UNIVERSIDAD POLITÉCNICA DE VALENCIA

DEPARTAMENTO DE INGENIERÍA HIDRÁULICA Y MEDIO AMBIENTE

"UTILIZACIÓN DE TÉCNICAS AVANZADAS EN EL TRATAMIENTO Y MANEJO DE DATOS. APLICACIÓN A LA GESTIÓN DE SISTEMAS DE ABASTECIMIENTO DE AGUA."

TESIS DOCTORAL

Presentada por:

José Luis Díaz Arévalo

Dirigida por:

D. Rafael Pérez García

D. Joaquín Izquierdo Sebastián

Valencia, 2010

Al recuerdo de José Domingo Díaz
Para María del Carmen Arévalo
Para Javier, Margarita, y Claudia

Para Ángela María y Daniel;
la compañía, entrega, energía
y vitalidad del día a día

AGRADECIMIENTOS

Esta tesis ha sido posible gracias a:

Los profesores D. Rafael Pérez García y D. Joaquín Izquierdo Sebastián directores de la misma.

El vicerrectorado de Investigación, Desarrollo en Innovación de la Universidad politécnica de Valencia por financiar parcialmente esta investigación a través de la Beca FPI concedida.

El ministerio de Educación y Ciencia por financiar parcialmente el desarrollo de esta tesis a través del proyecto *"Aplicación de técnicas de minería de datos en la planificación, operación, mantenimiento y gestión de sistemas de abastecimiento de agua (SIBA)"* (Ref. DPI 2004-04330)

A la Empresa Multipropósito de Calarcá S.A. ESP y especialmente a la ingeniera Luz Marina Arbeláez Arbeláez (Líder Investigación y Desarrollo), por facilitar los datos utilizados en esta investigación.

A los profesores y compañeros del Grupo Multidisciplinar de Modelación de Fluidos.

A Vicent y Eugenia, Miguel y Petra Amparo, Gonzalo y Amparo.

A mi esposa Ángela María y mi hijo Daniel.

A mi familia María del Carmen, Javier, Claudia Bibiana, María Paula, Santi, Gabriela, Andrés, Margarita, Felipe, Manuela, Claudia, Luis, Marthica, Martha Lu, Paula, Elisa, Toño, Fabi, María Victoria, Arturo, Gloria, Diego Luis y Ángela.

A Arturo y Olga, Mario y Estela, Mario y Nancy, Gabriel y Olga, Delva, Alba, Angel, Claudia Yaneth, y Eider.

RESUMEN

Según el historiador griego Heródoto, hacia el año 3050[1] antes de Cristo los antiguos faraones Egipcios recopilaron datos referentes a la población y riqueza del país con el objeto de preparar la construcción de las pirámides, lo que corresponde quizá al inicio histórico de la estadística, y al reconocimiento de la necesidad humana de recopilar y almacenar gran cantidad y diversidad de información. Por otra parte, la utilización de técnicas avanzadas en el tratamiento y manejo de datos conocida con el término *minería de datos,* se acuña a principios de los años 90 del siglo pasado, basado en los campos de la ciencia y el conocimiento de: las bases de datos; la recuperación de información; la estadística clásica; el aprendizaje automático; los sistemas para la toma de decisión; la visualización de datos; la computación paralela y distribuida; y otros cómo, el lenguaje natural; el análisis de imágenes; el procesamiento de señales; los gráficos por computadora, etc.

En esta tesis, de forma general, se realiza un planteamiento sobre el tratamiento y manejo de datos aplicado a los sistemas de abastecimiento de agua; con éste fin, se hace uso del paradigma del *descubrimiento de conocimiento en bases de datos* (*KDD*, de sus siglas en inglés) y se aplican específicamente algunos métodos de *minería de datos* (Data Mining). Debido a la diversidad de problemáticas que pueden y deben ser resueltas, y tomando como base la información (datos) con la que se cuenta en las diferentes etapas del diseño, operación y gestión de un sistema de abastecimiento de agua, consideramos que éste es un campo apropiado para la aplicación y desarrollo de estas metodologías. Muchas otras ramas de la ciencia y del conocimiento han sido exploradas por medio de estos métodos. El agua es un recurso natural vital que aunque se renueva naturalmente, presenta dotaciones limitadas en cantidad y calidad para cada lugar y momento concretos. Actualmente, hay un interés creciente más no

[1] Ruiz, M., D. (2004), *"Manual de estadística"*, editado por eumed.net ISBN: 84-688-6153-7

suficiente, hacia la promulgación y el avance en la utilización de técnicas novedosas en el manejo de la información por parte de la comunidad científica de ingenieros del agua, así como de los entes encargados del manejo y gestión de los sistemas de abastecimiento.

Por tanto, como aporte a la divulgación se ha desarrollado esta tesis, cuidando con rigor que los planteamientos y las discusiones aquí expuestas, permitan vislumbrar la contribución de la utilización de estas técnicas avanzadas hacia la gestión de los sistemas de abastecimiento de agua. Concretamente, a partir de la información disponible y las herramientas seleccionadas, se generó un modelo de gestión basado en reglas de decisión, para tratar los daños ocasionados y reportados en una red de abastecimiento de agua.

La metodología seguida consistió en realizar un amplio estudio del marco teórico del descubrimiento de conocimiento en bases de datos y como aporte presentarlo en idioma español, teniendo en cuenta que la mayor parte de las investigaciones y desarrollos del tema se han hecho en inglés, por lo cual es escasa la información en lengua Castellana. A continuación, se realiza un estudio exhaustivo del estado del arte acerca de las investigaciones y trabajos realizados sobre la utilización, aplicación, y desarrollo de los temas expuestos en el marco teórico de los sistemas de abastecimiento de agua, detallando con profundidad algunos de ellos como base.

Posteriormente se presenta una aplicación práctica real, consistente en encontrar las posibles causas de los daños ocurridos durante el año 2006 en la red del sistema de abastecimiento de agua potable del municipio de Calarcá, que se encuentra ubicado en la región cafetera de Colombia. Estos daños fueron reportados por la empresa que gestiona el abastecimiento. Para el desarrollo de esta aplicación se buscaron las posibles relaciones entre las diferentes variables encontradas con base en la información disponible en tales reportes, en el modelo hidráulico del abastecimiento, y en planos temáticos de factores de riesgo por causas naturales importantes en esta región del país. Parte de la información utilizada para esta aplicación práctica nos fue suministrada por la empresa de

capital mixto *Multipropósito de Calarcá S.A. ESP*, quien gestiona el abastecimiento de agua en el municipio.

Después de realizar los pasos pertinentes para el desarrollo de la metodología de *KDD* y de escoger los modelos apropiados, se hace uso de las siguientes herramientas de minería de datos: árboles de regresión y clasificación, redes neuronales y redes de Kohonen, con apoyo del programa Clementine 9.0 de SPSS para encontrar las relaciones entre las variables. Los mejores resultados se obtuvieron con los algoritmos de clasificación y regresión y, aunque con éstos no se llega a tener deducciones que concluyan relaciones fuertes de dependencia que permitan extraer causalidades entre los diferentes daños reportados y las diferentes variables tenidas en cuenta, sí es cierto que, con estos resultados y los desarrollos futuros propuestos, se puede contar con una herramienta que permita ayudar en el diseño, operación y manejo de los sistemas de abastecimiento de agua, basándose en la información que va generando el propio sistema.

Los resultados obtenidos, aparte de su gran potencial de aplicabilidad en sistemas de abastecimiento de agua potable, pueden ser mejorados si se cuenta con una información básica tomada para este fin, tal como se plasma en las recomendaciones finales.

La metodología seguida y el modelo práctico estudiado presentan las ventajas de poder ser realimentados continuamente, así como tomar decisiones en tiempo real por parte del gestor de la red de abastecimiento, lo cual indudablemente proporciona una útil y poderosa herramienta de gestión; y lo más interesante, a partir de la propia información real de la red de abastecimiento, lo cual salva la problemática de las incertidumbres presentes al momento de plantearse el modelado de la red de abastecimiento de agua potable, así como la valoración subjetiva de parámetros incluidos en estas formulaciones.

Para finalizar, se proponen algunas líneas futuras de actuación sobre la mejora de la información disponible y en general, la investigación a seguir para mejorar los resultados obtenidos hasta el momento.

ABSTRACT

According to Greek historian Herodotus, around 3050 BC, the ancient Egyptian pharaohs collected data about country's wealth and population in order to prepare for construction of the pyramids. This effort to compile great and diverse amounts of information just might correspond to the historical beginning of statistics. Additionally, the use of advanced techniques in data management known as *data mining* was coined in the early 1990s, based on fields of science and knowledge such: databases; information retrieval; classical statistics; machine learning; decision making process; data visualization; parallel and distributed computing; and others such as, natural language; image analysis; signal processing; computer graphics, etc.

The general objective of this PhD thesis is the treatment and handling of data applied to Water Supply Systems (WSS) from the paradigm *Knowledge Discovery in Databases* (KDD), specifically using data mining methods. Due to the variety of problems that could be solved from the information obtained during design, operation and management of WSS, we consider that this is an appropriate field for a KDD application model. Many other branches of science and knowledge have been explored by the KDD model. Although water is a natural renewable resource, it has limitations with respect to quantity and quality for each location and specific time. Engineering's scientific community and WSS managers have shown a growing interest in the development of innovative data management techniques, but much remains to be learned. The KDD and data mining techniques can help us to support the resolution of questions to be developed, managed and corrected in WSS.

We developed this PhD thesis to contribute to this effort, and we have taken rigorous care to ensure that proposals and discussions presented here enable us to understand the use of KDD in the management of WSS. Specifically, we used the available information and the selected tools to generate a management model based on decision rules to deal with reported damages in the water network.

The first step of the methodology was to conduct a comprehensive theoretical framework study of KDD. Since most investigations and developments of this subjects have been published in English we present it in Spanish. Next, we developed an exhaustive state-of-the-art study for the investigations involving the use, application and development of KDD methodology in WSS, and we describe it in detail.

Next, we present a practical application to find the damage that occurred during 2006 in the water supply network of Calarcá, a town located in the Colombian coffee region. (The water supply company reported these damages.) To develop the application, we searched relationships between variables found. Data used included information reports, the hydraulic model and risk factor level maps that measure natural disasters affecting the Colombian coffee region. Some of the information used for this application was given to us by the public-private partnership company *Multipropósito of Calarcá S.A. ESP*, which manages the water supply in the town.

After KDD's steps were done and appropriate models were chosen, we decided to use the following data mining tools: classification and regression trees, neural networks and Kohonen networks. The data mining software used was SPSS Clementine 9.0. The best results were obtained with the classification and regression algorithms, though these results do not indicate a strong dependency between variables. Nevertheless, with the outcomes of modeling and the future developments proposed, we can obtain a framework based on the information generated by the system itself. This tool would assist us in the management and solution of problems regarding the design, operation and management of a WSS.

The results obtained, besides their great potential for WSS applicability, can be improved with basic information taken for this purpose, as reflected in the final recommendations.

The methodology and the studied model have the advantage of continuous updating as well as real-time decision-making capability which undoubtedly provide a powerful and useful management tool. Most interestingly, the tool

comes from the actual information of the supply network itself, which eliminates the problem of uncertainties that arise when modeling the network, as well as subjective assessment of parameters included in these formulations.

Finally, we propose some future courses of action to improve the available information and the research to achieve better results than those obtained up to this point.

RESUM

Segons l'historiador grec Heròdot, cap a l'any 3050 abans de Crist els antics faraons Egipcis van recopilar dades referents a la població i riquesa del país amb l'objecte de preparar la construcció de les piràmides, la qual cosa correspon potser a l'inici històric de l'estadística. A més del fet de la necessitat humana de recopilar i emmagatzemar gran i diversa quantitat d'informació. A principis dels anys 90 del segle passat d'altra banda, comença la utilització de tècniques avançades quant al tractament i maneig de dades conegudes amb el terme mineria de dades. Aquelles tècniques es basen en els camps de la ciència i el coneixement, com ho són: les bases de dades; la recuperació d'informació; l'estadística clàssica; l'aprenentatge automàtic; els sistemes per a la presa de decisió; la visualització de dades; la computació paral·lela i distribuïda; i altres com, el llenguatge natural; l'anàlisi d'imatges; el processament de senyals; els gràfics per ordinador, etc.

En aquesta tesi, de forma general, es realitza un plantejament cap al tractament i maneig de dades aplicades als sistemes d'abastiment d'aigua. Amb aquest fi, es fa ús del paradigma del descobriment de coneixement en bases de dades (KDD, de les seues sigles en anglés) i, s'apliquen específicament alguns mètodes de mineria de dades (Data Mining). Considerem que aquest és un camp apropiat per a l'aplicació i desenvolupament d'aquestes metodologies, a causa de la diversitat de problemàtiques que poden i han de ser resoltes, i prenent com a base la informació (dades) amb la que es compte en les diferents etapes del disseny, operació i gestió d'un sistema d'abastiment d'aigua. Moltes altres branques de la ciència i del coneixement han sigut explorades per mitjà d'aquests mètodes. L'aigua és un recurs natural vital que encara que es renova naturalment, presenta dotacions limitades en quantitat i qualitat per a cada lloc i moment concrets. Recentment existeix un interès creixent dirigit a la promulgació i l'avanç en la utilització de tècniques noves en el maneig de la informació per part de la comunitat científica d'enginyers de l'aigua, així com, dels ens encarregats del maneig i gestió dels sistemes d'abastiment. Malgrat açò aquest es un camp en cosa poc explorat.

Per tant, com a aportació dins d'aquest camp s'ha desenvolupat esta tesi, cuidant amb rigor que els plantejaments i les discussions exposades, permeten albirar el suport de la utilització d'aquestes tècniques avançades en a la gestió dels sistemes d'abastiment d'aigua. Concretament, a partir de la informació disponible i les ferramentes seleccionades, es va generar un model de gestió basat en regles de decisió per a tractar els danys ocasionats i reportats en xarxes d'abastiment d'aigua.

La metodologia seguida va consistir a realitzar un ampli estudi del marc teòric del descobriment de coneixement en bases de dades i com a aportació presentar-ho en idioma espanyol, ja que la major part de les investigacions i desenvolupaments del tema s'han fet en anglès, contant-se amb escassa informació en la llengua Castellana. A continuació es realitza un estudi exhaustiu de l'estat de l'art sobre les investigacions i treballs realitzats quant a la utilització, aplicació i desenvolupament dels temes exposats en el marc teòric en els sistemes d'abastiment d'aigua, detallant amb profunditat alguns d'ells com a base.

Posteriorment es presenta una aplicació pràctica real, consistent a trobar les possibles causes dels danys ocorreguts durant l'any 2006 en la xarxa del sistema d'abastiment d'aigua potable del municipi de Calarcá, que es troba ubicat en la regió cafetera de Colòmbia. Aquests danys van ser reportats per l'empresa que gestiona l'abastiment. Per al desenvolupament d'aquesta aplicació, es van buscar les possibles relacions entre les diferents variables trobades amb base en la informació que es troba en aquests reports, en el model hidràulic de l'abastiment i, en plans temàtics de factors de risc per causes naturals importants en esta regió del país. Part de la informació utilitzada per a aquesta aplicació pràctica ens va ser subministrada per l'empresa de capital mixt *Multipropósito de Calarcá SA ESP*, qui gestiona l'abastament d'aigua al municipi.

Després de realitzar els passos pertinents per al desenvolupament de la metodologia de KDD, i d'escollir els models apropiats, es fa ús de les següents ferramentes de mineria de dades: arbres de regressió i classificació, xarxes neuronals i xarxes de Kohonen, amb suport del programa Clementine 9.0 de SPSS,

per a trobar les relacions entre les variables. Els millors resultats es van a produir amb els algoritmes de classificació i regressió i, encara que amb aquests no s'arriba a tindre deduccions que conclouen relacions fortes de dependència que permeten extraure causalitats dels diferents danys reportats i les diferents variables tingudes en compte, sí que és cert que amb els resultats obtinguts i els desenvolupaments futurs proposats es pot comptar amb una ferramenta que permeta ajudar en el disseny, operació i maneig dels sistemes d'abastiment d'aigua, basant-se en la informació que va generant el propi sistema.

Dels resultats obtinguts, estem convençuts que a banda del seu gran potencial d'aplicabilitat en sistemes d'abastiment d'aigua potable, poden ser millorats si es compta amb una informació bàsica presa per a este fi, tal com es plasma en les recomanacions finals.

La metodologia seguida i el model pràctic estudiat, presenten l'avantatge de poder ser realimentats contínuament, així com la presa de decisions en temps real per part del gestor de la xarxa d'abastiment, la qual cosa indubtablement proporciona una poderosa i útil ferramenta de gestió, i el més interessant, a partir de la pròpia informació real de la xarxa d'abastiment, la qual cosa salva la problemàtica de les incerteses presentes al moment de plantejar-se el modelatge de la xarxa d'abastiment d'aigua potable, així com la valoració subjectiva de paràmetres inclosos en aquestes formulacions.

Per a finalitzar, es proposen algunes línies futures d'actuació quant a la millora de la informació disponible i, en general, a la investigació que s'ha de seguir per a millorar els resultats obtinguts fins al moment.

TABLA DE CONTENIDO

INDICE DE FIGURAS

INDICE DE TABLAS

Capítulo I

Introducción y objetivos

I.1. Introducción

La aleatoriedad introducida en un modelo no es la única vía de incertidumbre. Los dos principales cursos de incertidumbre son la aleatoriedad y la carencia de conocimiento. La carencia de conocimiento proviene de nuestra inhabilidad para conceptuar los procesos del mundo real en forma matemática, especialmente para sistemas complejos. Como resultado, la incertidumbre corresponde a un hecho objetivo del fenómeno bajo consideración, o a una impresión subjetiva de la percepción humana (El-Baroudy y Simonovic, 2006). En el caso de los sistemas de abastecimiento de agua, para los cuales el análisis de su modelado en general se puede llevar a cabo ya sea simulando el estado de la red en un único instante (estático), o representando el comportamiento de la red a lo largo del tiempo (dinámico), no deja de presentarse cierto grado de incertidumbre con respecto al resultado obtenido, debido a la falta de conocimiento de algunas de las variables que intervienen en el prototipo (por ejemplo, la casuística estocástica de las demandas o, la rugosidad de los materiales) durante la vida útil del sistema; también se debe considerar que algunas formulaciones están basadas en la experiencia y conocimiento del experto (diseñador u operador). Por otra parte, además de la incertidumbre asociada a variables físicas del sistema, se debe tener claro que cualquier tipo de modelo sólo presenta una aproximación a la realidad, más que una representación exacta de la misma. Por tanto, el manejo e interpretación de la información producida durante la gestión de un abastecimiento de agua puede ser de ayuda para disminuir ese grado de incertidumbre que se plantea en el momento del modelado. En el último par de décadas se ha trabajado en el desarrollo de diferentes técnicas para el manejo y tratamiento de datos que pueden revelar patrones, tendencias, y comportamientos no conocidos de estos datos con anterioridad y que, tal como se plantea durante el desarrollo de esta tesis, permiten disminuir la incertidumbre generada durante la etapa de modelado de la red de abastecimiento.

En su introducción a la minería de datos y el descubrimiento de conocimiento, Fayyad *et al.* (1996) enumeran una serie de directrices para

seleccionar una posible aplicación de minería de datos. La primera de ellas es el impacto potencial significativo que pueda tener la aplicación; en referencia a esto, como ejemplo, *Langley y Simon (1995)* presentan el caso de una fundición sueca que hizo uso de reglas de inducción para desarrollar un sistema de análisis térmico para controlar su producción de aleaciones de hierro; como resultado de la aplicación de esta metodología se logró un ahorro de 50 dólares por tonelada al año, haciendo uso de prácticas de producción más eficientes. Si intentamos llevar este ejemplo al caso de los sistemas de abastecimiento de agua podríamos plantearnos el ahorro en costos de reparaciones de redes, si, por ejemplo, se consigue un esquema de rehabilitación de redes de acuerdo con diferentes factores que afectan la red de distribución de agua y que puedan ser implementados en modelos de minería de datos (*Díaz et al.*, 2003b); así como la detección de fraudes y fugas estudiando patrones de comportamiento de consumos, con lo cual se podrían obtener tanto ahorros en costos de producción como en conservación del recurso natural; o la optimización energética de los sistemas de abastecimiento, con el consiguiente ahorro en los costos derivados de la necesidad de energía requerida por la red para su correcto funcionamiento. Uno de los criterios básicos para la implementación de una herramienta de minería de datos es el conocimiento previo existente del problema a analizar; como en cualquier método computacional es importante el concurso de personas expertas en el tema para su desarrollo. Otros criterios tienen en cuenta la calidad de los datos en sí mismos: que se disponga de una cantidad suficiente de datos para analizar, que los atributos sean relevantes para las preguntas que se espera obtener respuesta y que los datos tengan bajo nivel de ruido.

Cuando se está en la tesitura de modelar un sistema desconocido o pobremente comprendido, un punto de partida inicial consiste en organizar campañas de medición o recolección de datos. Generalmente, se inicia por una parte, con la medición de funciones de influencia (o fuerza) desde el exterior del sistema (variables externas) midiendo a la vez la respuesta del mismo en vista del cambio de estado del sistema (variables de estado o internas), y por otra parte con el cambio de la salida del sistema (funciones resultantes). Por tanto, sólo

después de que se han conseguido gran cantidad de datos con la suficiente calidad, se puede identificar un sistema. Entonces se pueden presentar tres escenarios posibles Kompare, 1995:

1. No se puede concluir nada útil de las observaciones. Esto puede suceder si la campaña de medición fue pobremente diseñada o llevada a cabo sin el suficiente periodo de tiempo; o si simplemente no existen relaciones entre las variables. En caso de ser un problema en el diseño de las campañas de medición, podría plantearse el mejorar la solución con más mediciones o rediseñando la toma de observaciones.

2. Se pude finalizar con un modelo estadístico de caja negra. Con esta categoría de modelos seremos capaces de predecir el comportamiento del sistema, aunque no podremos caracterizar su estructura intrínseca y comportamiento. En otras palabras, seremos capaces de decir lo que el modelo hace, pero no cómo. Además, no seremos capaces de garantizar el comportamiento del modelo en regiones no cubiertas por los datos con los que se construye el modelo. Esto se debe al hecho de que el modelo sólo cubre las relaciones encontradas dentro de los datos dados.

3. En algunos casos podremos ser capaces de reconocer patrones dentro de los datos y, a partir de estos patrones, crear inferencias acerca de procesos básicos del sistema observado; entonces, después de mediciones repetitivas, seremos capaces de desarrollar un modelo conceptual. Tal modelo, también llamado "caja blanca" o modelo transparente, será capaz de expresar *"qué"* hace el modelo y *"cómo"* lo realiza. Debido al fondo conceptual del modelo, se tiene más certeza de que represente la realidad, al igual que puede ser de ayuda cuando el modelo es utilizado fuera del rango de los datos con los cuales fue construido.

Por otra parte, Berthold y Hand (2003) en la introducción de su libro hacen notar la ambigüedad de que si el análisis de datos puede ser inteligente puede ser a la vez poco inteligente. Datos distorsionados, elección incorrecta de cuestionamientos, aplicación incorrecta de las herramientas analíticas de datos,

sobreajuste, alto grado de idealización de un modelo, un modelo que va más allá de las varias fuentes de incertidumbre y de ambigüedad en los datos, entre otros, llevan a que el análisis pueda ser poco inteligente. Una de las razones por la cual es interesante el análisis inteligente de datos, es porque no consiste simplemente en aplicar una variedad de herramientas a un problema dado, sino que se debe realizar una valoración crítica, exploración, prueba y evaluación. Es un dominio que requiere de inteligencia y cuidado, así como, la aplicación de conocimiento y experiencia acerca de los datos.

En referencia a lo anterior, las herramientas de *software* utilizadas para aplicar las técnicas de minería de datos, generalmente producen listas con gran cantidad de resultados que pueden ser útiles en una aplicación particular. Sin embargo, en situaciones reales, no toda esta información resulta de utilidad. Por consiguiente, esto es importante para (Gibert *et al.,* 2008):

- Identificar la información relevante en los resultados obtenidos, basándose en los objetivos de cada análisis en particular.

- Encontrar la mejor forma de presentar los resultados seleccionados al usuario final.

En general, un sistema de abastecimiento de agua corresponde a una estructura compleja en la cual se llevan a cabo los procesos de producción, transporte y distribución de agua. Dentro de esta estructura, las redes de tuberías representan una de las infraestructuras de mayor valor de la sociedad industrial. Por tanto, el diseño de planes de mantenimiento tanto para las condiciones actuales como futuras de demandas y presiones, así como la reducción de costos futuros de mantenimiento, representa una parte integral dentro de la estrategia de la gestión de la red. El deterioro de una red de agua puede ser debido, por una parte, al incremento de las demandas y, por otra, al deterioro físico de los componentes del sistema que transportan el agua.

Dos de los principales criterios de funcionamiento de un sistema de distribución de agua son: que sea económico de construir, operar y mantener, y,

que opere de manera fiable (Dandy y Engelhardt, 2006); por tanto los requerimientos de funcionamiento del sistema pueden ser divididos en económicos, de fiabilidad, y de calidad del agua. El criterio económico tiene que ver con la operación y mantenimiento del sistema de distribución. Los principales costos asociados al sistema de distribución son los costos de capital de infraestructura, especialmente las tuberías; y así como que los mayores costos en rehabilitación del sistema corresponden al reemplazo de estos activos. El criterio de fiabilidad intenta cuantificar la capacidad del sistema para entregar las demandas de agua con las presiones requeridas en todos los puntos. El tercer criterio tiene que ver con la calidad del agua suministrada respecto a parámetros microbiológicos, químicos y estéticos.

La dificultad cuando se incluye el criterio de fiabilidad ya sea en el diseño u operación de un sistema de abastecimiento de agua, radica en el hecho de que no existe una definición universalmente aceptada, al igual que métodos unificados para cuantificarla. El conflicto se origina en identificar cuáles son los factores que se deben cuantificar cuando se mide la fiabilidad de un sistema de abastecimiento y modelar las variaciones existentes tanto temporales como espaciales. Entre los factores que se han identificado en la literatura (Dandy y Engelhardt, 2006) que contribuyen en la fiabilidad del sistema se tienen: la fiabilidad mecánica de los componentes y, la capacidad para suministrar la demanda requerida durante los eventos de falla.

Adicionalmente, la gestión del sistema de abastecimiento de agua se presenta como una tarea compleja *(León et al.,* 2000). La responsabilidad del manejo recae en primera instancia sobre los operadores. Aunque algunas tareas se realizan de forma automática, los operadores toman muchas de sus decisiones basándose en su propia experiencia e intuición para decidir la acción más adecuada en cada momento. La aplicación de modelos matemáticos de optimización es una alternativa viable; sin embargo estos modelos deben ser lo suficientemente flexibles para considerar cualquier cambio topológico en la red. Además, el principal problema con el que se cuenta para el desarrollo de estos

modelos es la consecución de la información, ya que en parte está basada en la experiencia de los operadores.

Igualmente, la elección de una estrategia de mantenimiento para un sistema de distribución de agua puede ser un problema con cierto grado de dificultad debido al gran número de elementos que componen el sistema (tubos, bombas, válvula, medidores, etc.), la evolución dinámica de la falla de una tubería, la existencia de cierto grado de asociación entre los componentes del sistema, el límite de recursos disponibles para las actividades de mantenimiento, y la dificultad asociada en cuantificar muchos de los beneficios y costos.

Muchas compañías de distribución de agua carecen de una visión general del estado de la producción y el sistema de distribución; es difícil, costoso y toma tiempo reconocer fallos dentro del sistema para identificar las operaciones no optimizadas y los recursos desperdiciados, y tomar las acciones de recuperación y mantenimiento necesarios. Una red de distribución de agua presenta un amplio conjunto de problemas específicos y requerimientos, entre otros *Afsarmanesh et al.* (1997), destacan:

- En general, la red está ubicada de manera dispersa a lo largo de un área geográfica extensa.

- Debido a la distribución geográfica del área cubierta, a los recursos disponibles y a los requerimientos de consumo, cada sistema es diferente y, en consecuencia, emplea mecanismos específicos de control y manejo.

- Para garantizar el suministro continuo de agua, las estaciones de bombeo y embalses se deben controlar y manejar de forma integrada.

- La calidad del agua se debe garantizar desde los puntos de toma hasta los consumidores.

- La red permanece en constante expansión debido al incremento de grandes consumidores: fábricas, naves industriales, etc.

- Se deben considerar incertidumbres en cuanto a los cortes de suministros totales o parciales por daños en las tuberías o derivaciones clandestinas, entre otros.

- Los criterios de mantenimiento regulares deben ser puestos en práctica, al objeto de optimizar la explotación del servicio.

Al efectuar el análisis hidráulico de una red de distribución de agua se pueden identificar tres tipos de variables (Revelli y Ridolfi, 2002). Se tienen variables tales como la longitud de la tubería, la cual tiene un valor preciso o exacto. Por otro lado, se tienen cantidades para las cuales se cuenta con la suficiente información para definir sus correspondientes distribuciones estadísticas con un grado razonable de fiabilidad; como es el caso de contar con dispositivos de medición de caudal instalados con lo que se podría determinar un patrón de consumos en cualquier nudo de la red. Finalmente, se tienen valores imprecisos o inexactos donde la dispersión de la información impide clasificarlos dentro de cualquiera de las dos categorías anteriores; uno de los casos más importantes es el valor del coeficiente de rugosidad, especialmente después de un cierto periodo de uso de la red. En general, se recurre para este último caso a alternativas como el valor medio sugerido por manuales técnicos, aunque siendo un procedimiento conveniente y rápido no deja de ser a menudo simplista, y que debe ser compensado con factores de seguridad apropiados. Para este tipo de variables se cuenta con técnicas y herramientas dentro de la minería de datos que permiten su manejo e interpretabilidad para un mejor conocimiento del funcionamiento del sistema.

El manejo sostenible de las redes de distribución de agua debe incluir, no únicamente metodologías para monitorear y reparar o reemplazar viejas infraestructuras, sino también, métodos para el modelado de condiciones en infraestructuras deterioradas, la valoración de datos históricos de incidentes y, el riesgo inherente de falla; concibiendo estrategias de reemplazo o reparación, enumerando costos del ciclo de vida, y visualizando las condiciones presentes modeladas por medio de bases de datos geoespaciales y sistemas de información

geográfica, para una mejor exposición de los modelos matemáticos subyacentes y las métricas estimadas hacia todos los entes o empresas interesadas (*Christodoulou et al., 2009*).

Para que un usuario de un abastecimiento tenga la posibilidad de abrir su grifo y poder satisfacer sus necesidades de requerimiento de agua, han tenido que intervenir previamente procesos de forma aislada y/o conjunta que, aunque no correspondan a un alto y complejo prototipo del quehacer científico, no es ni mucho menos cuestionable que corresponden a uno de los mayores progresos de la civilización para evitar epidemias, enfermedades y muertes humanas. Es por esto, principalmente, que el desarrollo de trabajos científicos que actúen directamente en la mejora de este tipo de infraestructuras del suministro de agua, debe ser constante y divulgados periódicamente para su conocimiento. En este sentido, el modelo aquí propuesto del descubrimiento de conocimiento en bases de datos presenta la posibilidad de integrar toda la operación y gestión del abastecimiento y, haciendo uso de la información correcta y adecuada, abre un abanico de factibles aplicaciones que pueden ir desde la fase de estudios y diseños técnicos de la obra, hasta la fase de su operación y manejo diario.

Entre esta gama de posibilidades, se puede pensar en:

- mejorar diseños a partir del conocimiento previo de otros abastecimientos de condiciones similares,
- optimizar estos diseños tanto hidráulica como económicamente,
- contribuir en la gestión del abastecimiento desde su punto de toma hasta la entrega a los usuarios finales,
- optimizar las tareas de tratamiento del agua cruda,
- colaborar en la gestión de pérdidas o fugas en la red a través de una monitorización de la misma, y un análisis de esta información por medio de las herramientas propuestas,

- detección de fraudes a partir del análisis de la información obtenida de consumos,

- mejoramiento del control de la calidad del agua en la red, lo cual implica la prevención de ataques a la misma,

- la posibilidad de un mejor acercamiento de cara al usuario, si se es capaz de prevenir el riesgo de fallos y roturas en la red,

- y el apoyo en la gestión económica del sistema.

De acuerdo con lo expuesto con anterioridad, se vislumbra la posibilidad de la utilización de técnicas avanzadas para el manejo de datos en los sistemas de abastecimiento de agua, permitiendo la integración de la gestión de los diferentes componentes del sistema. Por ser los sistemas de abastecimiento de agua grandes productores de información (datos), tener un estado que se podría establecer como virgen o "poco estudiado" en cuanto al manejo de esta información (para el gran tamaño de la información que se obtiene, se deja de aprovechar una gran cantidad), y por lo mencionado en los párrafos anteriores, tenemos la convicción de que el tema seleccionado de investigación plantea un reto interesante dentro del inicio de la actividad investigadora; a este respecto cabe mencionar que ya se han realizado algunos aportes en publicaciones y congresos al respecto (Díaz *et al.*, 2004a, 2004b, 2007, 2008, Izquierdo *et al.*, 2008a).

Este documento ha sido dividido en capítulos; en el primero, además de esta introducción, se presentan los objetivos generales y los objetivos específicos que se proponen para desarrollar. El segundo capítulo corresponde a una descripción de aspectos relacionados con los sistemas de abastecimiento de agua sobre los cuales se puede aplicar la metodología propuesta; así como con un cierto nivel de detalle a aspectos del fundamento teórico sobre el cual está basado el *descubrimiento de conocimiento en bases de datos.* La extensión del mismo responde al hecho de que en la actualidad es escasa la información de la que se dispone en idioma español; por tanto consideramos que es un aporte en este sentido. Adicionalmente, se profundiza en las técnicas de mayor divulgación en

cuanto a la minería de datos. El tercer capítulo versa sobre los antecedentes y el estado del arte del tema. Algunos artículos, considerados importantes como aporte en la aplicabilidad de la teoría desarrollada en el capitulo anterior, son descritos con algo más de profundidad. En este capítulo se presenta una aplicación a partir de algunas mejoras efectuadas en el algoritmo de *particle swarm optimization (PSO)* basándonos en parte de la información utilizada en esta tesis. En el cuarto capítulo se presenta una descripción de la zona sobre la cual fue desarrollada la aplicación práctica, así como, una descripción detallada de la información de la cual se dispuso. El quinto capítulo corresponde a los resultados obtenidos y la discusión planteada de acuerdo con estos resultados. En el capítulo sexto se presentan las conclusiones y líneas futuras de investigación, para finalizar con la bibliografía y los anexos del documento.

I.2. Objetivo general

El objetivo general que pretende este trabajo es, a partir de un ejemplo de aplicación práctica real, presentar la utilidad de las técnicas avanzadas en el tratamiento y manejo de datos como ayuda en la operación y gestión de los sistemas de abastecimiento de agua. Estas técnicas están fundamentadas en los pasos del descubrimiento de conocimiento en bases de datos KDD, y en especial la minería de datos. Para llevar a cabo este objetivo, se realizó una revisión exhaustiva del estado del arte del tema, así como del componente teórico del posible abanico de herramientas a utilizar.

Con este fin se investiga, estudia, y transmite el conocimiento acerca de la posibilidad de la utilización de técnicas de manejo de datos aplicables en los sistemas de distribución de agua; esto anterior como herramienta novedosa y práctica para su implementación como apoyo en la gestión del sistema.

Para el caso planteado en esta tesis; el objetivo es encontrar con base en la información disponible y las herramientas seleccionadas (tratamiento y manejo), un modelo para tratar los reportes de daños (datos) en la red de abastecimiento,

y así poder aplicar este modelo en el manejo (gestión) del sistema de abastecimiento.

I.3. Objetivos específicos

- Realizar una investigación lo más completa posible acerca del llamado *"descubrimiento de conocimiento en bases de datos"*, y la minería de datos.

- Difundir este conocimiento ya que la mayor parte de la información no se encuentra escrita en lenguaje castellano.

- Investigar en profundidad el estado del arte de la aplicabilidad de las técnicas de *minería de datos* en el diseño, planificación y gestión de los sistemas de abastecimiento de agua.

- Identificar posibles aplicaciones hasta ahora no desarrolladas en las cuales las herramientas de manejo de datos puedan ser útiles para el objetivo general de este trabajo.

- Investigar posibles desarrollos prácticos en diferentes actividades de la ciencia que permitan encontrar paralelismos a la problemática de los sistemas de abastecimiento de agua.

- Identificar entre las posibles herramientas y técnicas de minería de datos la apropiada para desarrollar una aplicación práctica, ajustándose a la información recopilada.

- Desarrollar un ejercicio práctico que permita presentar los beneficios y bondades de las técnicas de manejo de datos en los sistemas de abastecimiento de agua, a partir de los reportes de peticiones, quejas y reclamos del municipio de Calarcá en el departamento de Quindío en la región cafetera de Colombia.

- A partir de estos reportes e información como el modelo hidráulico de la red de abastecimiento, datos climatológicos, e información acerca de la caracterización de riesgos y vulnerabilidades por amenazas geológicas, generar

modelos para establecer y clasificar los daños en la red por medio de la metodología reseñada para el manejo de datos.

- Detallar unos resultados y conclusiones acerca de esta aplicación práctica, con el fin de aportar novedad con la investigación.

- Establecer pautas para que los entes encargados de la gestión y operación de abastecimientos hagan uso de estas herramientas.

- Sugerir futuros avances y líneas de investigación que permitan la mejora de lo que hasta el momento se ha desarrollado y aportado.

Capítulo II

Marco teórico

II.1. Introducción

El planteamiento en el que se basa el desarrollo de esta tesis doctoral, está fundamentado en la creencia (asentimiento) y convicción de la utilidad del llamado *descubrimiento de conocimiento en bases de datos* (KDD, de sus siglas en inglés) en general, y más específicamente de las técnicas de *minería de datos* en la búsqueda de patrones, tendencias, y conocimiento dentro de un sistema de abastecimiento de agua, que nos permitan aportar las bases necesarias para el establecimiento de una herramienta como apoyo en tareas de diseño, manejo, y gestión de redes de abastecimiento de agua.

Este capítulo se enfoca en una exposición sucinta de temas relacionados con los aspectos técnicos de los sistemas de abastecimiento de agua, en los cuales conceptuamos se podría hacer uso de las herramientas de minería de datos, así como en la explicación, con algo más de detalle y profundidad, de algunas de las técnicas más empleadas de éstas mismas.

Como se ha mencionado en el capítulo anterior, los sistemas de abastecimiento de agua generan gran cantidad de información, desde su etapa de diseño hasta su operación y gestión. Parte de esta información es utilizada para realizar tareas rutinarias o programadas, otra depende de la heurística empleada por los funcionarios encargados de ciertas tareas, los cuales, con el tiempo y conocimiento empírico han desarrollado sus propios prototipos de gestión pero que no se encuentra documentada; y gran parte no es aprovechada por falta de un correcto almacenamiento, o por no contar con las herramientas apropiadas para su análisis.

A continuación se realiza una descripción de generalidades de los sistemas de distribución de agua; los temas descritos son susceptibles de la utilización de la metodología expuesta en este documento del descubrimiento de conocimiento en bases de datos, especialmente la problemática de las fugas. El apartado de análisis de fiabilidad del sistema corresponde a un interés que se ha venido incrementando en los últimos años, debido al riesgo de que se puedan atacar

infraestructuras de servicios públicos, y ante el cual también cabe el uso de la metodología expuesta. Posteriormente, se presenta una breve introducción al proceso de descubrimiento de conocimiento haciendo énfasis en la metodología CRISP-DM para estandarizar los procesos del descubrimiento de conocimiento y la minería de datos, así como la metodología para realizar las tareas de pre y pos procesamiento de la información; algunas de las técnicas más utilizadas de minería de datos y de mayor divulgación, también son expuestas en este capítulo.

II.2. Sistemas de distribución de agua

Básicamente, un sistema de distribución de agua está compuesto por: una fuente de donde tomar el recurso natural; una estructura o medio de captación de esta agua; un sistema de aducción para llevar el líquido hasta la planta de tratamiento donde, deberá ser tratado para poderlo suministrar con garantía de potabilidad a los usuarios del sistema; y una red de distribución que lleva el agua hasta cada usuario. Si se trata de redes de riego, directamente de la aducción se toma el agua para satisfacer las necesidades hídricas de los cultivos. En regiones de climas áridos, con valores bajos de precipitaciones de lluvia o acumulados en períodos de tiempo muy cortos, como ocurre en ciertas zonas de España, se hace necesaria la construcción de presas para poder regular el caudal circulante por las fuentes y, así, poder garantizar un abastecimiento durante cada período hidrológico. También, es muy frecuente la utilización de aguas subterráneas, perforando el suelo para su extracción.

A continuación, se detallan algunos aspectos que pueden proporcionar mayor información como posibles aplicaciones para la minería de datos, debido a su importancia y grado de impacto dentro de la gestión de los abastecimientos de agua.

II.2.1. Red de distribución de agua

Las redes de distribución de agua pueden ser tan simples como un tubo

que lleva agua de un embalse a otro, o tan elaboradas, como un conjunto interconectado de redes de distribución para un área metropolitana.

Una red está constituida por dos elementos esenciales: nodos y tuberías, acompañada por elementos accesorios de protección: válvulas, depósitos, tanques, bombas etc.

Los sistemas de agua están constituidos en general por cuatro partes:

1. Suministro de agua (aducción)
2. Tratamiento
3. Distribución
4. Remoción Sanitaria

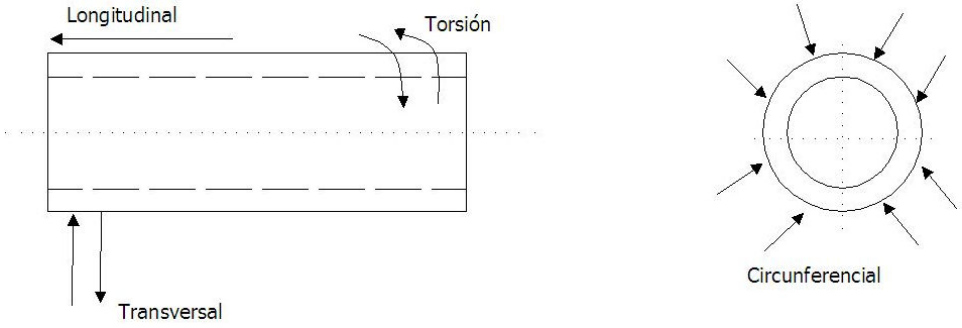
El **sistema de suministro** está compuesto por los diferentes sitios de captación del agua como ríos, quebradas, caños, embalses, presas, acuíferos subterráneos, y pozos, y los acueductos y tuberías de distribución, que entregan el agua a usuarios distantes. El **sistema de tratamiento** comprende básicamente filtración y otras plantas que eliminan impurezas y agentes perjudiciales, e instalaciones de saneamiento (típicamente cloración), que destruyen bacterias y otros contaminantes biológicos. El **sistema de distribución** está formado por embalses y torres de regulación de energía, mallas de tuberías, bombas, y otros componentes que entregan el agua desde los sistemas de tratamiento al usuario final. Finalmente, el **sistema de remoción sanitaria** y de desechos comprende alcantarillados y sistemas relacionados con la toma y entrega de aguas contaminadas, con residuos domésticos e industriales, a las instalaciones de tratamiento sanitario, y las instalaciones que retornan el agua tratada al medio ambiente.

A continuación, se presentan detalles sobre aspectos que comprometen el funcionamiento de las tuberías de conducción del agua, y que proporcionan los fundamentos para comprender el fenómeno del fallo en una tubería.

II.2.2. Tensiones Mecánicas sobre las tuberías

Una de las causas que provocan un mal funcionamiento en las redes de tuberías de agua es el incremento de la demanda; pero quizá la principal corresponde al deterioro físico de los componentes en el sistema de transporte del fluido; por tanto los componentes de mayor interés, en las estrategias de rehabilitación son las tuberías. Las roturas en tuberías, corresponden a fallas estructurales que requieren de acciones inmediatas por parte de la compañía de agua, para remediar la situación (*Babovic et al.*, 2002).

Al considerar la falla estructural de una tubería, es importante entender las cargas o tensiones que ésta soporta, ya que estas cargas juegan papel en el deterioro del tubo, al desmejorar la resistencia física de este pudiendo provocar su falla estructural. Los tipos de tensiones que pueden provocar rotura son:



Fuente: adaptado de Babovic *et al.*, 2002.

Figura II.1. Diferentes tipos de tensión soportados por una tubería

- Tensión circunferencial: es la tensión debida a la presión uniforme alrededor del tubo (interior y exterior). Se debe principalmente, al peso del suelo, o a presión interna.

- Tensión Longitudinal: es la tensión dirigida, a lo largo de la longitud del tubo, debido a expansión térmica, asentamiento diferencial del lecho, o presión interna cerca de los codos o válvulas.

- Tensión de corte transversal: es la tensión debida a la fuerza ortogonal a la longitud de la tubería, en una dirección específica (generalmente vertical). La causa principal en este caso, es una libre expansión de la tubería, es decir parte

de la tubería que no está soportada longitudinalmente por el lecho.

- Tensión cortante por torsión: es la tensión debida a la torsión de la tubería, a lo largo del eje longitudinal. Puede ser producida, por movimiento de otros tubos conectados transversalmente.

II.2.3. Fugas

Este es uno de los apartados de la gestión de un sistema de abastecimiento de agua, que genera retos tanto técnicos como estratégicos para su localización y control, en el cual, la aplicabilidad de herramientas de manejo de datos puede resultar de gran interés y apoyo. Indudablemente, las fugas representan uno de los mayores costos del sistema, ya sea por la rehabilitación, reparación o renovación necesaria de infraestructuras, o por el costo del agua no contabilizada, que se deja de facturar, independientemente de los daños ambientales que se puedan causar, así como el malestar por parte de los clientes.

Los factores que pueden provocar que una tubería falle son diversos, aunque principalmente estas fallas se deben a una combinación de factores. Las causas de falla de la tubería pueden dividirse en los siguientes grupos:

- Calidad del material y edad del tubo.
- Factores ambientales.
- Calificación del personal que realiza la colocación de la tubería.
- Condiciones de servicio bajo las cuales opera la tubería.

La calidad del material y la calificación del personal son variables difíciles de incluir en ecuaciones de predicción. Los factores ambientales son importantes en el historial de rotura de las tuberías, e incluyen tanto el suelo como la climatología. La interacción del suelo con el tubo puede causar fallas en diferentes formas, por ejemplo suelos expansivos debido a su alta humedad pueden generar cargas externas sobre la tubería, situación que se puede dar también si se

presentan asentamientos diferenciales. Otro factor ambiental que puede provocar fallas es la reacción electromecánica entre el suelo y el tubo, que puede provocar corrosión. La lluvia puede ser el principal factor en algunos tipos de fallas, ya que determina el contenido de humead del suelo, y por tanto su movimiento, y la corrosión que pueda ocurrir.

Las condiciones de servicio, que pueden provocar rotura en la tubería, corresponden a cargas internas y externas sobre ésta. Las presiones internas se deben al fluido transportado, que se pueden incrementar súbitamente debido a la ocurrencia de transitorios hidráulicos. Esta situación se puede presentar en situaciones normales de operación del sistema (encendido o apagado de bombas), o puede ocurrir por el cierre de una válvula, para reparar una rotura o fuga existente en la tubería. Las cargas externas sobre la tubería son debidas a elementos sobrepuestos a ellas. Estas cargas incluyen el suelo sobre la tubería, el tráfico rodado, y las construcciones o estructuras cercanas.

El agua no contabilizada se atribuye generalmente a fugas, errores de medición, y robo. Sin embargo, según informes de la "*Asociación internacional de suministro de agua (IWSA)*" (*Burn et al., 1999)*, las fugas son la causa principal del agua no contabilizada. Esta pérdida de agua tiene un costo, no solo en términos de gasto y desperdicio de un recurso natural, sino en términos económicos (*Díaz et al.*, 2003a). La primera pérdida económica producida por la fuga, corresponde al costo del agua bruta, su tratamiento, y transporte. Como resultado de la fuga se tienen daños en la red, erosión del material de lecho de la tubería, y el daño en cimientos y fundaciones de vías y edificaciones; esto podría considerarse como una segunda pérdida económica. La disminución de la seguridad de suministro, como resultado de la reducción en agua almacenada per cápita, también representa un costo si tal disminución requiere del aumento del suministro para mantener la seguridad. Sumado a las pérdidas económicas y ambientales generadas por las fugas, también se genera un riesgo de salud pública, teniendo en cuenta que cada fuga es un punto potencial de entrada de contaminantes, si ocurren caídas de presión en la red.

Detección de fugas

Los métodos que se utilizan en la actualidad, en mayor o menor medida, corresponden a los siguientes:

Monitoreo de caudales

Una herramienta útil en el programa de control efectivo de fugas es la monitorización de los caudales mínimos nocturnos. Ya que las fugas son continuas, el caudal nocturno menos cualquier flujo de *"uso legal"* da el volumen verdadero de agua perdida. Si esta diferencia se aproxima a cero es difícil determinar las fugas, si por el contrario, la diferencia es alta se garantiza una mayor detección de fugas. La monitorización de caudales nocturnos se puede tomar como un paso de inspecciones para la detección efectiva de fugas. En vez de monitorear la totalidad del sistema, este se puede dividir en zonas. Con la información resultante se puede planear una detección efectiva de fugas, en zonas donde se presentan caudales nocturnos altos. La técnica consiste en cerrar válvulas sucesivamente en la zona, para aislar secciones de tubería, y registrar la reducción en el caudal de agua. Una gran reducción de caudal indica la existencia de fuga entre en las dos secciones anteriores aisladas. Como resultado del procedimiento, la localización de la fuga puede ser llevada a una zona de menor tamaño, y utilizar métodos acústicos para su búsqueda.

Las técnicas de monitoreo de caudales nocturnos también son utilizadas para detectar fugas en curso. Realizando un monitoreo continuo de los caudales nocturnos, pueden detectarse los cambios inusuales en los volúmenes de agua. Basándose en la experiencia del operador de la red, se puede determinar si el incremento en caudal y volumen es producido por una fuga.

Detección de ruidos de fugas

Las fugas en un sistema de distribución de agua pueden ser comprobadas

sistemáticamente haciendo uso de equipamiento acústico, que detecta los sonidos o vibraciones que se generan a lo largo de la tubería, debidos al agua que se escapa de esta a presión. El agua pasa a través del orificio o fuga originando ruido en la tubería, generalmente en el rango de 500-800 Hz. Para que se pueda detectar por medio de métodos acústicos esta fuga, la presión debe ser como mínimo de 10 m.c.a. La atenuación de la vibración que viaja a lo largo del tubo depende del material y diámetro del mismo. El agua que fuga impacta sobre el suelo en el área de fuga, causando un sonido diferente, normalmente en el rango de 20-300 Hz. Un tercer ruido, normalmente también en el rango de 20–300Hz, es causado por el agua fugada circulando dentro del agua que permanece en la cavidad del suelo adyacente a la fuga (*Burn et al.*, 1999). Según el tipo de material, el primer ruido puede ser trasmitido a lo largo de la tubería a largas distancias, los otros dos sonidos, están generalmente limitados a las proximidades de la fuga.

Los dispositivos (mecánicos o electrónicos) para escuchar ruidos utilizan mecanismos o materiales sensitivos, tales como elementos piezoeléctricos, para detectar los ruidos y vibraciones de una fuga.

Los dispositivos electrónicos pueden incluir amplificadores de señal y filtros de ruido, lo cual es muy útil en condiciones ambientales adversas. El uso de dispositivos auditivos es generalmente correcto, pero su efectividad depende de la experiencia del usuario.

Métodos alternativos

Las fugas pueden ser detectadas haciendo uso de métodos alternativos, o equipamiento desarrollado para otras industrias. Aquellas técnicas no acústicas, cómo los métodos de trazadores, los infrarrojos o termografía, y el radar, entre otros, han sido utilizadas pero su efectividad no está establecida aún.

Métodos de gases trazadores: consiste en la introducción de un gas no tóxico, insoluble en el agua, y de color, en el sistema de distribución de agua. Los

gases generalmente utilizados son el helio y el hidrógeno.

Termografía: su principio radica en que el agua fugada de la tubería enterrada cambia las características térmicas del suelo adyacente, notándose la diferencia entre el suelo húmedo y seco alrededor de la fuga.

Radar (GPR): Este método consiste en la transmisión de pulsos electromagnéticos de onda corta en el suelo, utilizando una antena de radar. Cuando el pulso de radar encuentra una interfase entre dos materiales, se refleja parcialmente a la superficie, donde es detectado por una antena receptora. La reflexión parcial de las ondas electromagnéticas, en la interfase entre dos materiales o cualquier anomalía, es debida al contraste en las propiedades dieléctricas. El intervalo entre las ondas electromagnéticas trasmitidas y reflejadas es utilizado para determinar la profundidad de la superficie reflejada. Por la transmisión de los pulsos de radar, en posiciones regulares de superficie (haciendo un *barrido [Scan]),* pueden ser determinados el tamaño y la profundidad de objetos enterrados. El GPR puede ser utilizado para detectar fugas de agua de dos formas: (a) identificando cavidades de suelo, creadas por el flujo turbulento del agua fugada, (b) identificando segmentos de tubería, que parecen más profundos de lo esperado, por el incremento de las constantes dieléctricas del suelo adyacente saturadas por el agua fugada.

II.2.4. Análisis de fiabilidad del sistema

En este apartado queremos presentar una perspectiva de un tema que ha tomado cierta relevancia, en especial a raíz de las amenazas terroristas que se han presentado en los últimos años en diferentes partes del planeta, y del cual no se puede excluir una parte tan sensible dentro del funcionamiento normal de cualquier población, como es su abastecimiento de agua; y la utilización de técnicas de manejo de datos puede colaborar en su detección y prevención (Huang y McBean, 2009).

Grandes y complejos sistemas de la ingeniería están sujetos a un amplio

rango de posibles condiciones futuras (El-Baroudy y Simonovic, 2006). Muchas de estas condiciones no pueden ser controladas o estimadas con un grado aceptable de exactitud. La incertidumbre, asociada con la cuantificación de aquellas condiciones potenciales, impone un gran reto para el diseño, planificación y gestión de los sistemas. Los sistemas de abastecimiento de agua incluyen en general diferentes tipos de instalaciones interconectadas, sirviendo amplias regiones geográficas. Por consiguiente, estos sistemas de abastecimiento de agua tienen el riesgo de fallar debido a riesgos naturales, o a causas antropogénicas, sin intención (errores operacionales o equivocaciones), o intencionales (actos terroristas).

Los actos terroristas pueden corresponder a amenazas físicas, químicas y biológicas, y a amenazas de tipo cibernético (Haimes y Longstaff, 2002). Las amenazas físicas se refieren por ejemplo al uso de explosivos dentro de los componentes físicos o de control del sistema de abastecimiento. Los agentes químicos y biológicos se utilizan directamente en el agua para contaminar el suministro y perjudicar la salud de los consumidores. Las amenazas cibernéticas, dentro del contexto de la guerra de la información, se refieren a la introducción de virus informáticos e información errónea a través de la red informática, para comprometer la regulación y control de la entrega de agua.

La viabilidad de una amenaza física, depende de un número de factores, entre los cuales están:

1. La configuración general de los componentes físicos del sistema.

2. La habilidad de los terroristas para acceder a las instalaciones, donde los explosivos pueden ser colocados para producir un efecto máximo.

3. La habilidad de los terroristas para trasportar e implantar explosivos en el sistema en cantidades suficientes para impactar la operación del sistema.

4. La habilidad de los terroristas para obtener o manufacturar explosivos, sin llamar la atención de los órganos de seguridad del estado.

5. El grado al cual los terroristas son capaces de arriesgar su vida para llevar a cabo una amenaza física contra un sistema particular.

6. En general, los sistemas que dependen de bombeos son más vulnerables a amenazas físicas que los sistemas que operan a gravedad. Los químicos utilizados para el tratamiento del agua, tales como el carbón activado o algunos compuestos del cloro o el flúor, que tienen altos potenciales reactivos, pueden contribuir a la severidad de un fuego o explosión.

Teniendo un cierto grado de control de la información generada por los sistemas de abastecimiento, se puede hacer uso de herramientas que permitan el manejo de esta información, con el fin de tomar las correspondientes medidas luego de realizar las predicciones acerca de un posible atentado. En general, cualquier cambio en el comportamiento del funcionamiento diario de la red, puede ser un síntoma de alarma ante algún factor que está afectando este funcionamiento. Así, si por ejemplo, se está monitoreando en tiempo real la calidad biológica y química del agua, cualquier cambio anormal puede ser detectado para tomar las medidas oportunas (*Huang y McBean, 2009*).

La viabilidad de la introducción deliberada de amenazas químicas o biológicas depende, entre otros factores, de los siguientes:

1. Las propiedades hidráulicas y geométricas del sistema.

2. La habilidad de los terroristas para acceder al lugar donde los agentes químicos y biológicos puedan ser inyectados.

3. La habilidad de los terroristas para trasportar e inyectar dentro del sistema cantidades suficientes de agentes necesarios para dañar la operación del mismo.

4. La habilidad de los terroristas para obtener o manufacturar los agentes en cantidades suficientes, sin llamar la atención de los órganos de seguridad del estado.

5. Grado de arriesgar la vida de los terroristas.

Las amenazas cibernéticas se pueden presentar en los sistemas del tipo SCADA, ya que se puede proporcionar información falsa a los operadores, interrumpir la operación del sistema, dañar sus componentes, y/o perjudicar la población o el ambiente.

Se ha afirmado, que los sistemas de abastecimiento de agua urbanos y rurales están entre las grandes realizaciones de la ingeniería del siglo XX (Baecher, 2006) y, que el suministro de agua potable dado por estos sistemas ha hecho más por proteger la salud humana que toda la ciencia médica moderna. Las necesidades de investigación para la protección de los suministros de agua y los sistemas de saneamiento incluyen la valoración de la seguridad física, la detección, monitoreo y tratamiento de contaminantes y, en menor medida, la seguridad cibernética.

Otras áreas, que se pueden identificar como investigaciones para esta protección, corresponden a:

1. La valoración de los riesgos de amenaza o vulnerabilidad.

2. La identificación y caracterización de los agentes químicos y biológicos.

3. El establecimiento de un centro de excelencia para el apoyo a las comunidades en la conducción de la valoración de la vulnerabilidad y el riesgo.

4. La aplicación de técnicas de seguridad informatizadas, utilizadas por las empresas del agua.

En cuanto a la garantía de la seguridad física de los sistemas de abastecimiento de agua, el análisis de fiabilidad de un sistema utiliza los conceptos de *carga* y *resistencia* como nociones fundamentales para definir el riesgo de fallo. Estos dos conceptos han sido utilizados en ingeniería estructural para reflejar el comportamiento característico del sistema bajo, condiciones de cargas externas. En los sistemas de abastecimiento de agua, los conceptos de

carga y resistencia pueden ser reemplazados por los de *demanda* y *suministro* respectivamente, para reflejar las variables específicas de dominio del sistema de abastecimiento *(El-Baroudy y Simonovic, 2006)*. La demanda puede ser definida como la variable que refleja los diferentes requerimientos de agua impuestos al sistema durante su vida útil, y el suministro como la variable característica del sistema que describe la capacidad de este para satisfacer la demanda.

II.3. Minería de datos (Data Mining)

Aunque en algunos casos en la bibliografía suele hablarse de minería de datos y el proceso de KDD como la misma cosa, en nuestro caso la minería de datos corresponde a unos de los pasos del descubrimiento de conocimiento en grandes bases de datos.

La minería de datos es definida como: *"La extracción no trivial de patrones válidos, novedosos, potencialmente útiles y finalmente comprensibles a partir de los datos"* (Fayyad *et al.*, 1996). Es decir, la minería de datos corresponde a un proceso que es generalmente iterativo (tanto por el algoritmo del que se haga uso como por la afectación sobre el proceso), generalizable para el futuro, desconocido con anterioridad, aplicable (útil) para el objetivo propuesto, que lleva a la comprensión del fenómeno bajo estudio.

A su vez, el propósito del descubrimiento de conocimiento en bases de datos *(Knowledge Discovery in Databases-KDD),* es la exploración de una cantidad considerable de datos *brutos,* para extraer patrones *potencialmente* útiles, *válidos* y *comprensibles*, combinando herramientas que han sido desarrolladas en diferentes comunidades científicas, tales como las *bases de datos*, la *inteligencia artificial*, la *ingeniería del conocimiento*, la *estadística*, y el *aprendizaje automático*, entre otros.

Por lo tanto, los métodos de minería de datos, y el descubrimiento de conocimiento en bases de datos, convierten los datos en información, la información en conocimiento, y el conocimiento en práctica.

El descubrimiento de conocimiento, como se menciona anteriormente, es un proceso que combina, entre otras, técnicas de aprendizaje automático, estadística, reconocimiento de patrones, conjuntos borrosos (difusos) y aproximados, etc., para extraer conocimiento o información de grandes cantidades de datos. Es utilizado para dar soporte a las decisiones tomadas, o para la explicación de fenómenos observados. El descubrimiento de conocimiento es un proceso que ayuda a dar sentido a los datos de forma legible y aplicable.

La naturaleza del *manejo del descubrimiento* en la minería de datos, hace uso de algoritmos de aprendizaje que pueden buscar y encontrar *clusters, patrones, asociaciones y regularidades interesantes* en bases de datos. Su habilidad para representar *información y conocimiento* extraído en forma comprensible, tal como con árboles de decisión, reglas, modelos de datos y mapas de concepto y conocimiento, le dan unas capacidades descriptivas y predictivas útiles.

II.3.1. El Proceso de KDD

El proceso de *KDD* es interactivo e iterativo, ya que en el trascurso de las actividades se deben tomar decisiones que pueden generar que se realicen consideraciones sobre actividades previas. Este proceso involucra numerosos pasos con variedad de decisiones tomadas por el usuario.

Los pasos generales del *KDD* comprenden, (Figura II.2):

- Comprensión del tema, conocimiento de antecedentes y planteamiento de metas a obtener.
- Seleccionar los datos, de acuerdo con los objetivos establecidos.
- Limpieza de datos y preprocesamiento: consiste en la remoción de ruidos y valores extremos, así como el manejo de campos de datos perdidos.

- Reducción de datos y proyección: consiste en encontrar rasgos útiles para representar los datos dependiendo de la meta de la tarea. Se plantea el uso de la reducción adimensional, o métodos de trasformación para reducir el número de variables en consideración, o encontrar representaciones invariantes de los datos.

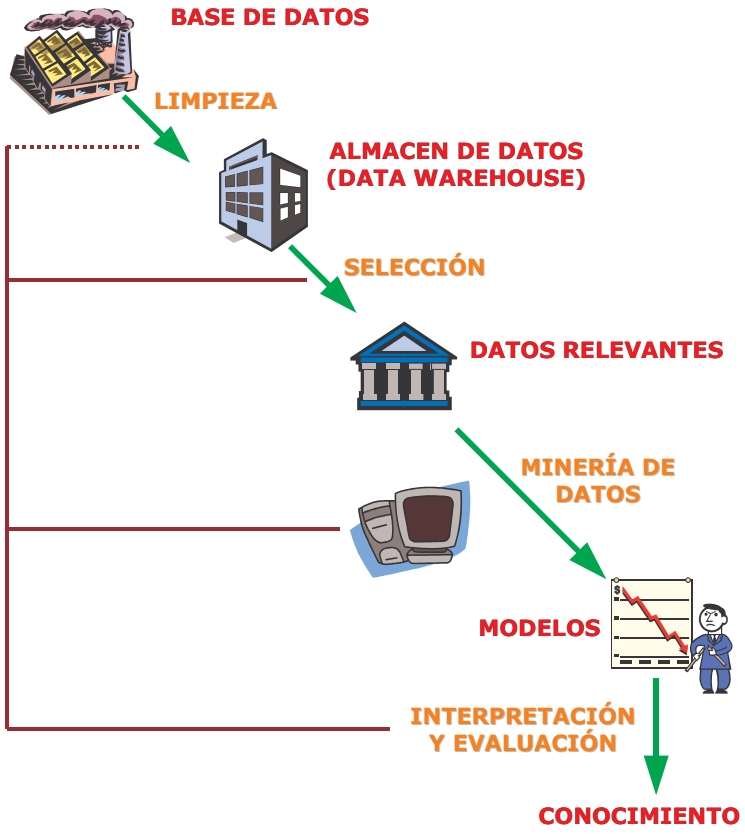


Figura II.2. Pasos del KDD

- Búsqueda de las tareas de minería de datos: decidir si la meta del proceso de *KDD* es clasificación, regresión, agrupamiento, etc.

- Búsqueda del algoritmo, o los algoritmos para la minería de datos: consiste en la selección del método o métodos a ser utilizados para la búsqueda de patrones en los datos.

- Minería de datos: corresponde a la propia búsqueda de los patrones de interés.

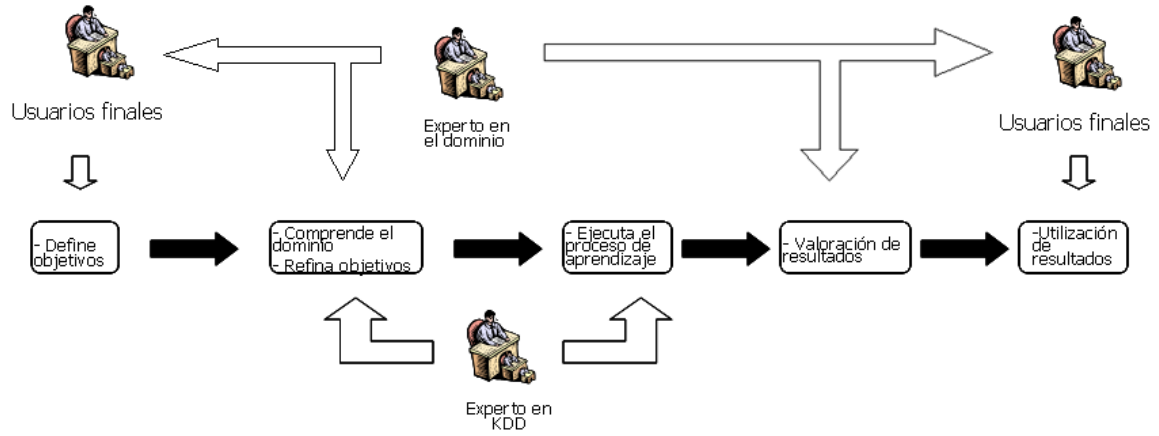
- Interpretación de los patrones minados, desde donde existe el posible

retorno a cualquiera de los pasos anteriores para futuras iteraciones.

- Consolidación del conocimiento descubierto: corresponde a la incorporación del conocimiento en la interpretación del sistema o, simplemente, a documentar y reportar las partes de interés.

La Figura II.3 visualiza los principales componentes y puntos de interacción humana, en un escenario típico para la aplicación de técnicas de aprendizaje automático, en un problema real.

Una situación típica puede consistir es un usuario final, con un objetivo preliminar, que presenta alguna posibilidad para la aplicación de técnicas de aprendizaje automático. El objetivo puede ser, por ejemplo, la necesidad de predecir el consumo de agua en una red de distribución.



Fuente: Adaptado de Camarihna y Martinelli, 1999.

Figura II.3. Escenario típico que representa la intervención humana en el proceso de aprendizaje

A veces, este objetivo puede estar bien definido y soportado por la información de todos los antecedentes para ejecutar la tarea. Sin embargo, es normal redefinir los objetivos. Con esta formulación preliminar se inicia el proceso. El siguiente paso corresponde a la interacción entre el experto en el dominio o tema de interés y el experto en aprendizaje automático. Se encuentran situaciones en que el experto en el tema es, a la vez, el usuario final. En esta fase de interacción, el experto en aprendizaje intenta conseguir el máximo posible del conocimiento del experto en el tema. En esta etapa el experto en aprendizaje,

junto con el experto en el tema, deben refinar el objetivo preliminar, definiéndolo en forma más fácil y útil, y transformándolo en una tarea formal de aprendizaje.

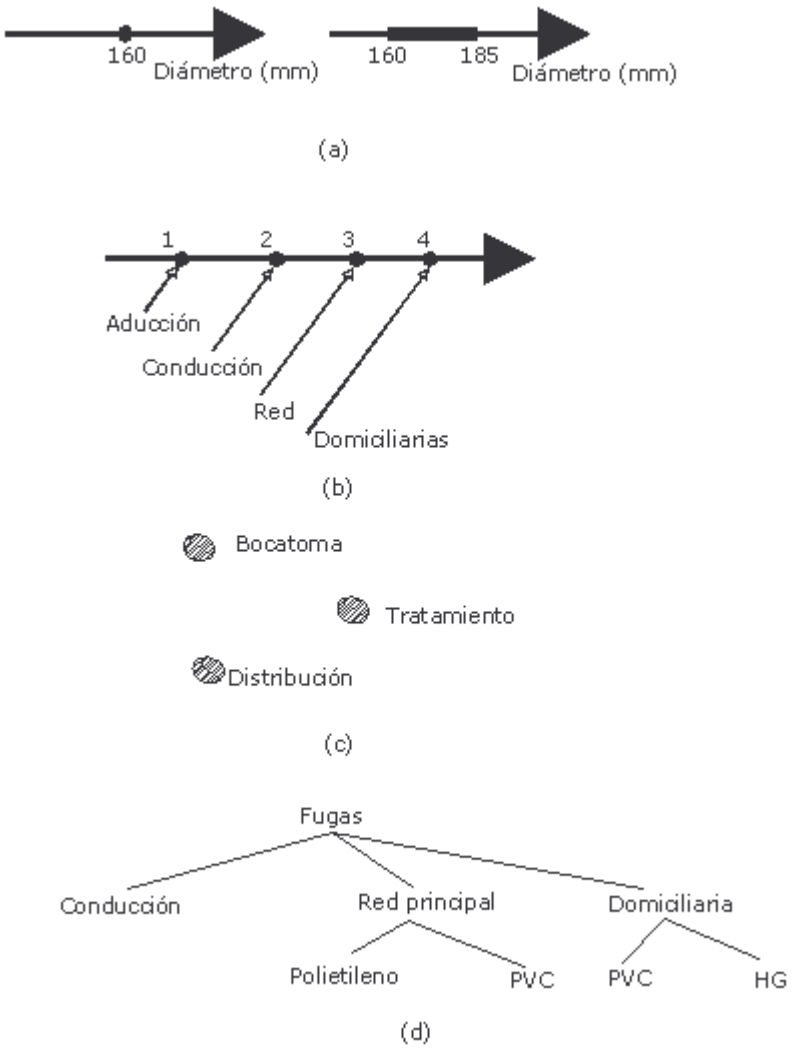
En una etapa posterior, el experto en aprendizaje automático utiliza su propio conocimiento acerca del proceso de aprendizaje y, tomando en cuenta el conocimiento del tema que ha adquirido del experto, intenta identificar resultados relevantes ilustrativos respecto a las tareas definidas de aprendizaje. Entonces, el experto en el tema valora los resultados obtenidos. Si los resultados son buenos y significativos, el conocimiento generado puede ser explorado por el ambiente de aplicación. Si los resultados no son lo suficientemente buenos, el proceso se reinicia desde la etapa preliminar para refinar los resultados.

II.3.1.1. Representación del conocimiento

La representación del conocimiento predetermina la ejecución de cualquier sistema de *descubrimiento de conocimiento;* se deben considerar diferentes tipos de datos, e identificar el papel de la granulación de la información en el sistema.

Los registros en las bases de datos pueden asumir diversidad de formas. En general, se puede diferenciar entre entradas numéricas y simbólicas. En la primera categoría se pueden encuadrar números, vectores, matrices bidimensionales, o multidimensionales. Las entradas simbólicas son utilizadas para describir alguna variable cualitativa (tales como oscuridad, claridad, etc).

El tema principal desde el punto de vista del descubrimiento de conocimiento es cómo manipular estos tipos de datos. En particular, se debe prestar atención al problema de comparación de dos segmentos de datos (calcular la distancia entre ellos). Cuando se trata con dos vectores numéricos, la tarea es sencilla: se puede utilizar cualquier definición de la distancia tal como la Euclídea, Minkowski, Tschebyschev, etc. El problema (ver Figura II.4) se presenta cuando se trata de la jerarquía de dos objetos (entidades).

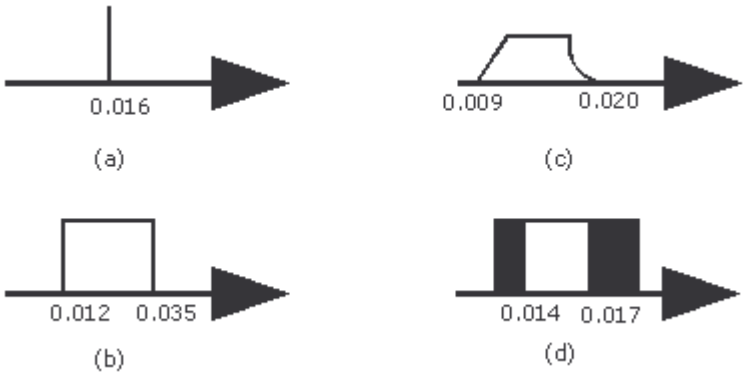


Fuente: Adaptado de Cios *et al.*, (1998).

Figura II.4. Ejemplos de tipos de datos (a) Continuos (b) Ordinales (c) nominales (d) estructura de árbol

La comprensión de los números por parte de los seres humanos es muy limitada. Por esto, en vez de hacer uso de lotes de números, se pueden reproducir agregados. Por ejemplo, cuando se busca en un mapa de isoyetas, no se presta atención a lluvias específicas sino a algo más abstracto, tal como zonas con altas o bajas precipitaciones en alguna región del país. Es decir, la información se hace granulada y representada en un nivel más alto de abstracción (agregación). Por tanto, todas las relaciones entre variables de interés, son cuantificadas de forma similar. Cuando se tienen datos con la intención de buscar algún patrón, se busca alguna distancia conceptual, formada por segmentos de conocimiento de una cierta granulación. La granulación de la información, es una forma de encapsular datos numéricos en una entidad única conceptual. Por ejemplo, cuando se está

interesado en un patrón que describa la rugosidad de las tuberías del abastecimiento, se podrían inspeccionar los datos desde este punto de vista, sin prestar en principio demasiada atención al orden de magnitud. Figura II.5



Fuente: Adaptado de Cios *et al.*, (1998).

Figura II.5. Diferentes modelos de granulación de la información (a) numérica (b) intervalo (c) borrosa (d) aproximada

Los gránulos de información son formados para reducir la complejidad de la descripción de sistemas en el mundo real (Bargiela y Pedrycz, 2001). La generalidad alcanzada con la granulación de la información, se produce por el sacrificio de la precisión numérica de los datos. En la figura anterior, la rugosidad de la tubería puede ser tomada como un valor de 0.016, pero a su vez ante la falta de conocimiento acerca de este valor en el tiempo, podría observarse el intervalo entre 0.012 y 0.035 para capturar mejor un valor medio. Igualmente, se podría introducir cambiando el carácter si-no de los límites, un conjunto difuso (Zadeh, 1965), o un conjunto aproximado (Pawlak, 1994), como dos paradigmas interesantes de la ventana de descubrimiento. Las ventanas de descubrimiento, pueden exhibir niveles muy diferentes de granularidad. El descriptor numérico (valor único), presenta un alto nivel de granularidad. El intervalo muestra un bajo nivel de granularidad, el cual es seguido por las representaciones del conjunto difuso, y del conjunto aproximado. El más bajo nivel de granularidad, podría ocurrir para un conjunto que cubra la totalidad del espacio. Esto expande la ventana de descubrimiento, e implica que el conocimiento descubierto se concentra en la totalidad de la base de datos.

Los patrones revelados de una base de datos, pueden ser representados

entre otros (comúnmente utilizados en la comunidad del aprendizaje automático) por: producción de reglas y árboles de decisión.

a. Reglas

Las reglas, son estructuras comunes de representación del conocimiento, y son especialmente populares en los sistemas basados en éste. Sus principales ventajas son su alta legibilidad, y representación en módulos. Las reglas, pueden ser añadidas y borradas fácilmente. El conocimiento representado de esta forma es altamente modificable. La desventaja de la generación de reglas es la dificultad de capturar y revelar una descripción general del problema, como en el caso de utilizar una representación gráfica. Un tipo de regla podría ser:

Si día = lunes, y clima = cálido, y humedad > 80%, entonces, demanda de agua sube 10%

Esta regla establece que si el día de la semana es lunes, y se tiene un clima cálido, y una humedad relativa superior al 80%, entonces la demanda de agua en un abastecimiento, se incrementa en un 10 por ciento.

Un aspecto importante y relevante de las reglas es su habilidad para manejar varios niveles de granularidad de la información, Figura II.5. Cambiando la granularidad de las condiciones y conclusiones, la regla puede ser más específica, o más general. Por ejemplo, si la granularidad de la conclusión se incrementa, la regla es más general. Por consiguiente, la regla:

Si rugosidad es 0.016, entonces, demanda es 10 l/s,

es más específica, que una regla cuya conclusión es un intervalo:

Si rugosidad es 0.016, entonces, demanda es [10,20] l/s.

Por otra parte, si la condición se hace menos específica, entonces la regla consigue mayor generalidad. Así, quitando la condición "día" de la regla, se gana en generalización:

Si clima = cálido, y humedad > 80%, entonces, demanda de agua sube 10%.

Otro ejemplo podría ser:

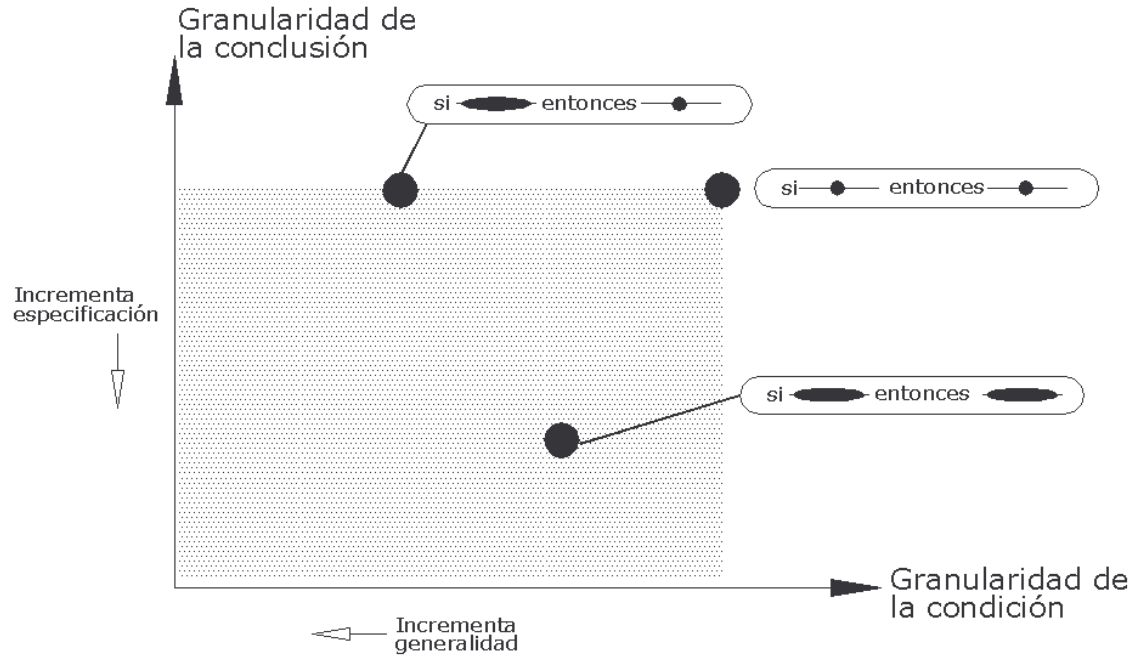
Si rugosidad es [0.014,0.017], entonces, demanda es 10 l/s,

donde la condición es reemplazada por un intervalo de la rugosidad del material.

Es evidente que incrementando la granularidad de la condición, se hace la regla más general, hasta un punto donde las reglas se vuelven incondicionales; por ejemplo:

Si rugosidad es CUALQUIERA, entonces, demanda es 10 l/s.

En la siguiente figura se representa el efecto de la granulación de la información mostrada con las reglas anteriores.



Fuente: adaptado de Cios *et al.*, 1998

Figura II.6. Efecto de la granulación en un sistema basado en reglas

b. Árboles de Decisión

Un método alternativo de adquisición de conocimiento es el aprendizaje

automático de los datos observados, que consiste en diseñar un algoritmo que pueda adquirir y afinar reglas de decisión de un conjunto de muestras o datos observados. Este método es conocido como aprendizaje inductivo o adquisición de conocimiento mediante ejemplos.

Los árboles de decisión corresponden a una construcción popular en el aprendizaje automático, puesto que son fácilmente visualizados y comprendidos. Se pueden tener árboles de clasificación, o árboles de asociación.

Aunque con los árboles de decisión se obtienen reglas fáciles de deducir e interpretar, estas pueden simplemente realizar divisiones paralelas a los ejes del espacio de estado, lo que conduce a la limitación de divisiones efectivas en el espacio cuando se encaran mapeos complejos no lineales entre los atributos de los datos de operación y las decisiones relacionadas, con lo cual se puede perder eficiencia en los resultados obtenidos.

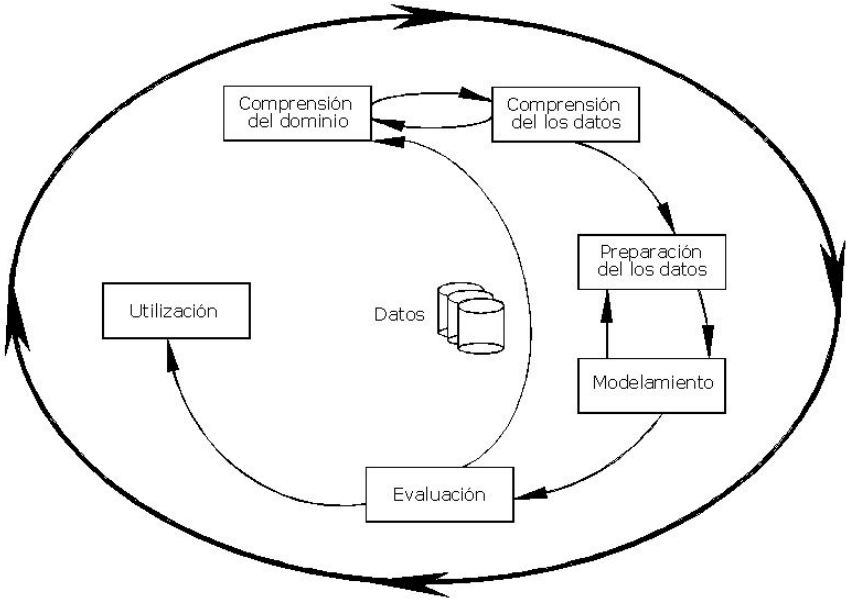
II.3.1.2. El modelo de procesos CRISP-DM

Esta iniciativa del CRISP-DM (*Cross-Industry Standard Process for Data Mining)* se realiza con la idea de estandarizar los procesos del *descubrimiento de conocimiento,* y de la *minería de datos.* Este proceso está compuesto por seis fases relacionadas entre sí (Figura II.7). Cada fase contiene un número de tareas las cuales producen resultados específicos.

Fase de comprensión del negocio (dominio)

Se dirige a en entender los objetivos y requerimientos del oficio, y convertirlos dentro de un problema de minería de datos, para diseñar un plan preliminar para alcanzar estos objetivos. Descuidar este paso puede llevar a un incremento del esfuerzo para producir respuestas correctas a cuestionamientos erróneos. El primer resultado de esta tarea son los antecedentes, los cuales detallan la información conocida acerca de la situación del dominio en cuestión y el inicio del proceso. Por tanto, en esta fase, además de determinar los objetivos

del dominio, se debe realizar una valoración de la situación, con lo cual se tiene un inventario de los recursos disponibles teniendo en cuenta restricciones, suposiciones, y cualquier otro factor considerado para determinar las metas de la minería de datos, puesto que el resultado de esta fase consiste en transformar los objetivos del dominio en metas de minería de datos (clasificación, descubrimiento de reglas, descubrimiento de subgrupos, ajuste de ecuaciones, agrupamiento, redes probabilísticas y causales, análisis de series temporales, análisis espaciales, etc.) Para finalizar, en esta fase se describe el plan para garantizar las metas de la minería de datos y, por consiguiente, garantizar los objetivos del negocio.



Fuente: Adaptado de Kosgen y Zytkow, 2002.

Figura II.7. Modelo de procesos CRISP-DM

Fase de comprensión de los datos

Corresponde a una recolección inicial de datos, y se continúa con actividades con el fin de familiarizarse con estos, para identificar problemas de calidad de la información, para tener una primera comprensión de la misma, o para detectar subconjuntos interesantes para futuras reexaminaciones. La primera tarea dentro de esta fase corresponde a listar los datos relevantes en el inventario de recursos. El resultado de esta tarea es un reporte inicial de los datos colectados con sus localizaciones, métodos utilizados para adquirirlos, y los problemas

encontrados.

Adicionalmente, en esta fase se realiza una descripción de los datos, examinando sus propiedades y generando un reporte en términos de formato, cantidad, identidades de campos, y cualquier otro rasgo superficial descubierto. Luego se debe verificar la calidad de los datos en función de que estén completos, correctos, y la cantidad de información perdida. Finalmente, se realiza una exploración de los datos haciendo uso de técnicas y herramientas de cuestionamientos *(querying)*, visualización y reportes, por ejemplo la distribución de atributos clave, relaciones entre pares de atributos, resultados de agregaciones simples, propiedades de subpoblaciones significantes, y análisis estadísticos simples.

Fase de preparación de los datos

Cubre todas las actividades con el fin de construir el conjunto a modelar. Las tareas de preparación de datos se realizan infinidad de veces y no necesariamente en un orden preestablecido. Estas incluyen tablas, registros y selección de atributos, así como transformación y limpieza. En general, esta fase genera dos salidas: el conjunto de datos para fases subsecuentes, y una descripción de los datos, que lista las características de la preparación de datos realizada.

En esta fase de preparación, una tarea importante corresponde a la selección de los datos, la cual cubre la selección automática o manual de tuplas, la selección de atributos o características, así como la reducción del número de valores, por ejemplo por medio de técnicas de discretización o de categorización. Los criterios de selección incluyen la relevancia para las metas de la minería de datos, así como las restricciones de calidad y técnicas, tales como los límites sobre el volumen o tipos de datos. Como salida se obtiene un criterio racional del por qué de la inclusión o exclusión de la información, y qué datos son retenidos.

Con la limpieza de datos se da a la información el nivel de calidad

requerido; tiene que ver con la selección de datos a limpiar, la introducción de omisiones adecuadas, o técnicas más ambiciosas tales como la estimación de datos perdidos por modelado; el reporte de esta tarea describe las decisiones y acciones para el manejo de los problemas de calidad de los datos y lista las transformaciones de los datos para su limpieza, y los posibles impactos sobre el análisis de resultados.

Luego se realizan las tareas para la construcción de datos, tales como la generación de variables derivadas, la creación de nuevos registros, o la transformación de valores por variables existentes. La siguiente tarea de integración, apunta hacia la combinación de la información de múltiples tablas o registros, para crear nuevos registros o valores; el resultado de esta tarea es la combinación de datos, uniendo dos o más tablas que contienen diferente información acerca de objetos iguales, o la agregación de datos, la cual tiene que ver con el cálculo de nuevos valores resumiendo la información de múltiples tablas o registros.

Otra tarea importante dentro de la preparación de los datos consiste en formatear la información, que principalmente cubre las modificaciones sintácticas que no cambian el significado, pero son requeridas por la herramienta de modelado. Los resultados típicos de esta tarea corresponden a registros reordenados, atributos reorganizados, y el reformateo dentro de valores.

Fase de modelado

En la fase de modelado se seleccionan y aplican las técnicas de modelado, y se calibran los parámetros para obtener valores óptimos. Puesto que muchas técnicas presentan requerimientos específicos sobre la forma de los datos, se requiere una interacción cercana con la preparación de los datos. En la primera fase de modelado se selecciona la técnica de modelado entre las técnicas y herramientas preseleccionadas. Si se aplican múltiples técnicas, esta tarea se desarrolla por separado para cada una de ellas.

Antes de construir los modelos, usualmente se genera un procedimiento para probar la calidad de los mismos. Por ejemplo, en tareas de minería de datos supervisada, tal como la clasificación, es común utilizar tasas de error como medida de calidad para los modelos generados. Por consiguiente, típicamente, se separan los datos en conjuntos de entrenamiento y de prueba, construyendo modelos sobre los primeros, y estimando su calidad sobre conjuntos de prueba.

La siguiente tarea dentro de esta fase, consiste en aplicar la técnica de modelado sobre los datos preparados para crear uno o más modelos. Puesto que las técnicas de modelado permiten la calibración de varios parámetros, se considera la elección de estos dentro de valores racionales. Los resultados de esta tarea corresponden a los modelos con su exactitud esperada, robustez y posibles defectos.

En la valoración del modelo se interpretan los resultados del modelado de acuerdo con los criterios establecidos y el diseño de prueba deseado. En general, se hace uso de herramientas estadísticas y otras herramientas disponibles para esta tarea. La valoración del modelo resultante resume los resultados de esta tarea, lista las cualidades de los modelos generados dando una clasificación con respecto a otros. Adicionalmente, esta valoración posibilita la revisión de los parámetros establecidos, con lo cual se puede hacer un afinamiento del modelo en pasos iterativos.

Fase de Evaluación

En este punto se cuenta con modelos que tienen una calidad alta desde la perspectiva de la minería de datos. En la fase de validación estos modelos son admitidos de acuerdo con las perspectivas del dominio, y se realiza una revisión de los pasos ejecutados para la construcción del modelo. Mientras que la valoración del modelo trata con factores como la exactitud y generalidad del mismo, esta tarea evalúa el modelo de acuerdo a los objetivos originales del dominio y los criterios de éxito. El resultado de esta tarea conduce al

establecimiento de si el proyecto alcanza los objetivos iniciales propuestos.

Adicionalmente a la evaluación del modelo, se considera la totalidad del proceso para determinar si hay algún factor importante o tarea que ha sido descuidada, o para identificar procedimientos genéricos para la generación de modelos similares en el futuro.

Finalmente, la última tarea de evaluación determina los pasos siguientes en el proyecto. De acuerdo con la valoración de los resultados y la revisión de procesos, se decide como proceder. Aquí se concluye si finalizar el proyecto e ir a su utilización si es apropiado, si iniciar iteraciones futuras, o si configurar un nuevo proyecto de minería de datos.

Fase de Utilización

La creación de modelos no es generalmente el fin del descubrimiento de conocimiento y la minería de datos. Dependiendo de los objetivos, la fase de utilización puede ser tan simple como la generación de un reporte, o tan compleja como implementar un proceso repetible de minería de datos. Para utilizar los resultados de la minería de datos dentro del negocio, en esta tarea se desarrolla la estrategia para su uso.

El monitoreo y mantenimiento son temas importantes si los resultados de la minería de datos forman parte del día a día del negocio y su ambiente. La preparación cuidadosa de la estrategia de mantenimiento, ayuda a prevenir de períodos largos de innecesaria utilización incorrecta de los resultados de la minería de datos.

Al finalizar el proyecto, se escribe un reporte final. Dependiendo de su plan de uso, este reporte puede ser un resumen del proyecto y sus experiencias, o una presentación de los resultados de la minería de datos.

Finalmente, se revisa la totalidad del proyecto y se valora su veracidad, así como donde fue bien y que necesita mejoras. La documentación de la experiencia

adquirida, consolida información importante para futuras necesidades. Por ejemplo, dificultades, aproximaciones erróneas, o pistas para seleccionar las técnicas apropiadas de minería de datos en situaciones similares forman parte de esta documentación.

II.3.1.3. Reducción de Datos

Las tareas de pre y pos procesamiento de datos corresponden a una de las partes críticas y más importantes de todo el proceso KDD (Gibert *et al.*, 2008). Las herramientas informáticas utilizadas para aplicar las técnicas de minería de datos, generalmente, producen listas llenas de resultados que pueden ser útiles en una aplicación particular; sin embargo, dado un caso real particular, no toda esta información es útil, por tanto, es importante identificar la información relevante de los resultados informáticos, basándose en los objetivos de cada análisis particular, y encontrar la mejor forma de presentar los resultados seleccionados al usuario.

Muestreo

La teoría del muestreo es uno de los mayores éxitos de la estadística moderna. Con la utilización de métodos de rutinas de muestreo, las características de las poblaciones pueden ser estimadas de forma fiable, imparcial y eficiente (Klösgen y Zytkow, 2002).

El muestreo aleatorio simple corresponde al tipo de muestreo más sencillo, en el cual cada grupo de objetos de un tamaño requerido tiene igual posibilidad de ser la muestra seleccionada. Mientras sea posible conseguir una muestra atípica de esta forma, las leyes de la probabilidad dictan que cuanto más grande es la muestra es más probable que represente la población de donde viene. En general, el método conduce a muestras en las cuales las características de los individuos son distribuidas similarmente a las de la población.

El muestreo aleatorio estratificado divide primero la población en estratos o grupos y, entonces, se hace uso de un muestreo aleatorio simple dentro de cada

grupo. Estos estratos se basan en una variable o variables que pueden estar relacionadas con la(s) variable(s) de interés, pero no de interés principal. Dividir la población en grupos que son similares antes del muestreo, producen dos ventajas: primero, el analista puede controlar el número de observaciones dentro de cada estrato; esto es particularmente útil, si un estrato tiene pocos miembros con respecto a otro, pero el analista desea conocer los efectos dentro del estrato. La segunda ventaja de la estratificación tiene que ver con la cantidad de error o varianza de los datos. Si los miembros dentro de un estrato son más similares entre ellos que los miembros de diferentes estratos, el estrato específico estimado será más preciso que uno de la muestra entera.

El muestreo por conglomerados es útil cuando los miembros de la población forman *clusters* naturales, tales como pacientes dentro de hospitales o empleados dentro de una compañía. Un enfoque puede ser muestrear dentro de cada clúster (muestreo estratificado), el otro, a menudo más fácil de aplicar, consiste en muestrear *clusters* aleatoriamente para luego incluir todos los miembros de cada clúster muestreado. El muestreo en dos etapas (generalmente etapa múltiple) combina las dos ideas de la elección aleatoria de *clusters* y, entonces, se muestrean miembros dentro de cada clúster. En el contexto de grandes bases de datos, una aplicación de muestreo por *clusters* consiste en seleccionar aleatoriamente bloques de datos y utilizar todos los datos sobre estos bloques. La motivación de esta aplicación es la recuperación de un registro de la base de datos de un bloque particular; el bloque entero debe ser leído en la memoria.

Cuando los individuos en una población son numerados de alguna forma, *el muestreo sistemático* es una opción. También conocido como muestreo *k*-ésimo, elige un miembro al azar de aquellos numerados entre 1 y *k*, para incluirlo dentro de la muestra. Ya que la selección no es aleatoria, este tipo de muestreo puede no ser representativo de la población, por tanto debe ser utilizado con cuidado.

El *muestreo en dos etapas* puede ser útil cuando se desea organizar la muestra basándose en los valores de una o más variables, pero sin conocerle rango o distribución de aquellas variables en la población.

En el contexto del modelado predictivo, *el muestreo incierto* puede proporcionar mejoras sobre el muestro al azar. Los métodos de muestreo por incertidumbre, solicitan de forma iterada etiquetas de clase para los ejemplos de entrenamiento, cuyas clases son inciertas a pesar de los ejemplos previos etiquetados.

Selección de Atributos

La selección de atributos corresponde a un problema de búsqueda que consiste en generar, evaluar, y seleccionar subconjuntos de atributos. El propósito de esta selección es triple: reducir el número de atributos, mejorar la exactitud de la clasificación, y simplificar la representación del aprendizaje. Entre los métodos representativos de selección de atributos se tienen las aproximaciones exhaustivas, las aproximaciones heurísticas, las aproximaciones no-determinísticas, y las aproximaciones basadas en instancias (Klösgen y Zytkow, 2002). En las tareas de minería de datos, la gran cantidad de la información puede hacer necesario considerar la estabilidad del algoritmo de selección de atributos. Cuando las tareas no son de clasificación (etiquetas de clase no disponibles), los algoritmos anteriores no son operativos, y se deben considerar otros como la selección de atributos escalable, y las aproximaciones supervisadas y no supervisadas. Para la reducción de la dimensionalidad, se hace uso de medidas de entropía para ordenar los atributos.

Agregación de Atributos

La agregación de atributos es un proceso a través del cual se crean nuevos atributos. El propósito es mejorar el funcionamiento de las tareas, tal como la estimación de la exactitud, la visualización y la comprensibilidad del conocimiento aprendido. Los dos principales grupos de agregación de atributos corresponden a la construcción y extracción de características. La construcción de atributos es un proceso que descubre información perdida acerca de las relaciones entre atributos, y aumenta el espacio de rasgos por inferencia, o creando nuevos

atributos. La extracción es un proceso que selecciona un conjunto de nuevos atributos de las características originales, a través de alguna función de mapeo.

Discretización de atributos numéricos

El tipo de datos puede modificarse para facilitar la utilización de técnicas que requieren tipos de datos específicos; es el caso de numerizar atributos con lo cual se reduce el espacio y se permite la utilización de técnicas numéricas. Por el contrario, la discretización es el proceso de convertir atributos numéricos en simbólicos, por la partición del dominio de atributos. Muchos algoritmos de minería de datos requieren atributos simbólicos (categóricos); por otra parte muchas bases de datos reales contienen atributos numéricos (también conocidos como continuos), con números enteros o reales como valores. Por tanto, es necesario preprocesar la información para el descubrimiento de conocimiento: los atributos numéricos deben ser convertidos a atributos simbólicos. Esta conversión se consigue partiendo el dominio de los atributos numéricos en intervalos; este proceso se conoce como discretización, y contribuye a un uso más eficiente del descubrimiento de conocimiento, y a la maximización de la exactitud de los resultados del conjunto de reglas inducido de los datos discretizados.

II.3.1.4. Modelos Básicos de Minería de Datos

La minería de datos está basada en una cantidad de modelos que capturan el carácter de los datos en diferentes formas.

Clustering (Agrupamiento)

El objetivo es encontrar grupos naturales (*clusters*) en datos con alta dimensionalidad. Los análisis clúster, desempeñan un papel central en gran variedad de campos, y a menudo son utilizados como una herramienta para el análisis de datos preliminar y descriptivo, y para clasificaciones sin supervisión. Su principal propósito consiste en identificar grupos homogéneos, identificando

similitudes entre objetos respecto a sus atributos característicos.

Modelos de Regresión

Se originan a partir del análisis de regresión estándar, y su parte aplicada conocida como identificación de sistema. La idea subyacente, es construir una función lineal o no lineal que explique los datos:

$$y = f(x, a),$$

Los parámetros de esta función, (a), son determinados mediante métodos conocidos de optimización, especialmente aquellos basados en el cómputo de matrices pseudo-inversas, o utilizando técnicas basadas en gradientes. El error medio cuadrático es el criterio utilizado con mayor frecuencia.

Clasificación

Corresponde al aprendizaje que clasifica datos dentro de categorías predeterminadas. El término se origina del reconocimiento de patrones, en el cual un gran número de clasificadores han sido desarrollados. Los clasificadores pueden ser considerados como un caso especial de modelos de regresión, en este caso la variable de salida (y) asume un número finito de valores correspondientes a las clases de interés.

Sumarización

Es una aproximación hacia la caracterización de datos por medio de un pequeño número de rasgos o atributos. La media y la desviación estándar corresponden a dos descriptores compactos de los datos. Esta técnica es, a menudo, aplicada a un análisis exploratorio de los datos, y a la generación automática de reportes.

Análisis de Enlace

Hace referencia a la determinación de relaciones (dependencias) entre campos en una base de datos. En un caso particular, se puede estar interesado en la determinación de correlaciones entre variables.

Análisis de secuencias

Este tipo de análisis está enfocado hacia problemas de modelado de datos secuenciales. Los modelos abarcan análisis de series temporales, modelos de series temporales y redes neuronales temporales.

II.3.1.5. Evaluación e Interpretación

La tarea de evaluar e interpretar los modelos obtenidos o patrones descubiertos por una técnica de minería de datos no es algo trivial, ya que estos patrones deben ser precisos, *comprensibles, útiles y novedosos*, por tanto, los criterios de evaluación son diversos y, en algunos casos, con bastante carga de subjetividad.

Para entrenar y probar un modelo se debe realizar la partición de los datos en dos conjuntos, uno de entrenamiento *(training set)* y otro de prueba *(test set)*, con el fin de garantizar que la validación del modelo sea independiente; en caso de no hacerlo se obtendrán resultados demasiado optimistas, y la precisión del modelo será sobreestimada.

El método de evaluación más básico, la *validación simple*, reserva un porcentaje del conjunto de datos como prueba y no lo utiliza para construir el modelo. Este porcentaje puede variar entre el 5% y el 50% de los datos. Esta división se realiza de forma aleatoria. En modelos predictivos, resulta fácil la interpretación de esta división, ya que en una tarea de clasificación, por ejemplo, después de generar el modelo se prueba sobre los datos de test y, de acuerdo a los resultados, se puede obtener la precisión, dividiendo el valor del número de

clasificaciones bien efectuadas por el total de datos clasificados.

En el caso de no tener una cantidad suficiente de datos para poder realizar la división necesaria para la evaluación, se puede utilizar la técnica conocida como *validación cruzada* (*cross validation*). Los datos son divididos en dos conjuntos equitativos de forma aleatoria; luego se construye un modelo con el primer conjunto que es utilizado para predecir los resultados en el segundo conjunto y obtener así una tasa de error o precisión; luego con el segundo conjunto se construye un modelo para predecir los resultados del primer conjunto y obtener así una segunda tasa de error. Finalmente, se construye un modelo con todos los datos y se calcula un promedio de las tasas de error que es utilizada para estimar mejor su precisión.

El método que se utiliza generalmente es el de la *validación cruzada con n pliegues* (*n-fold cross validation*). Aquí los datos se dividen de forma aleatoria en *n* grupos; uno de ellos se reserva para la prueba y con los *n-1* restantes, juntando todos sus datos, se construye un modelo que es utilizado para predecir el resultado del conjunto de prueba. Este proceso se repite *n* veces dejando cada vez un conjunto distinto para la prueba, obteniendo *n* tasas de error independientes. Finalmente, se construye un modelo con todos los datos y se obtienen sus tasas de error y precisión, promediando las *n* tasas de error disponibles.

Otra técnica utilizada para cuando se disponen de pocos datos es la conocida como *bootstrapping*, que consiste en construir primero un modelo con todos los datos iniciales; luego se crean numerosos conjuntos de datos llamados *bootstrap samples*, realizando un muestreo de los datos originales con reemplazo, es decir, se van seleccionando ejemplos del conjunto inicial, pudiendo seleccionar el mismo ejemplo varias veces. Luego se construye un modelo con cada conjunto y se calcula su tasa de error sobre el conjunto de prueba, que corresponde a los datos sobrantes de cada muestreo. El error final estimado para el modelo con todos los datos se calcula promediando los errores obtenidos para cada muestra.

Medidas de evaluación de los modelos

Dependiendo de la tarea de minería de datos realizada se tienen diferentes medidas de evaluación:

- Si se trata de una tarea de clasificación, la medida es su *precisión predictiva*, que se obtiene dividiendo el número de ejemplos clasificados correctamente, por el número total de ejemplos.

- En el caso de que la tarea corresponda a reglas de asociación, se evalúa de forma separada cada regla, teniendo en cuenta tanto su *cobertura* (*soporte*), como su *confianza*. La cobertura corresponde al número de ejemplos a los que la regla se aplica y predice correctamente. La confianza es la proporción de ejemplos que la regla predice correctamente, es decir, la cobertura dividida por el número de ejemplos a los que se puede aplicar la regla.

- Si la tarea es regresión, la forma más habitual es por el error cuadrático medio del valor predicho respecto al valor que se utiliza como validación.

- Para tareas de agrupamiento, estos modelos son más difíciles de evaluar ya que no existe una clase o valor numérico donde medir las veces que el modelo aprendido predice correctamente. En general depende del método utilizado, aunque suelen ser función de la cohesión de cada grupo, y de la separación entre grupos. Una primera aproximación podría ser utilizar la verosimilitud *(likelihood),* lo cual significa quedarse con la hipótesis con la cual los datos sean más verosímiles. Otra opción para formalizar la cohesión y separación entre grupos se puede dar, por ejemplo, utilizando la distancia media al centro del grupo de los miembros de un grupo, y la distancia media entre grupos, respectivamente. Otra alternativa para saber si es adecuado un modelo de agrupamiento podría ser por la comparación de utilizar diversas técnicas de aprendizaje para un mismo conjunto de datos. Maulik y Bandyopadhyay (2002) utilizan el índice de *Davies-Bouldin*, el índice de *Dunn´s*, el índice de *Calinski Harabasz*, el índice I, y el índice de *Xie y Beni,* para evaluar el desempeño de los

métodos de agrupamiento de *K-means, single linkage,* y *simulated annealing*, encontrando que el índice I es el más fiable de los utilizados.

II.3.1.6. Construcción de multiclasificadores

Bagging (Saqueo)

Se utiliza para generar clasificadores a partir de la evidencia; deriva del *bootstrap aggregation* visto anteriormente, en el que se generan conjuntos de entrenamiento por la selección aleatoria con reemplazo de una muestra de *m* ejemplos de entrenamiento del conjunto original. Aunque se esperaría que los conjuntos entrenados y aprendidos fueran similares, no siempre ocurre, por ejemplo, en los árboles de decisión; cambios poco notorios en el conjunto de entrenamiento causan que el árbol construido por el algoritmo sea diferente; este tipo de algoritmos son denominados *"weak learner"*, debido a esta inestabilidad o debilidad.

La estimación de nuevos ejemplos es realizada por medio de votación mayoritaria, eligiendo la clase que tenga mayor número de votos entre todos los clasificadores. En el caso de modelos de regresión, se calcula la media de los valores estimados para cada uno de los modelos del conjunto, la cual reduce el error cuadrático con respecto a un modelo único se regresión.

Boosting (Estimular)

Aquí se asigna un peso a cada ejemplo del conjunto de entrenamiento; por tanto, en cada iteración se aprende un modelo que minimiza la suma de los pesos de los ejemplos clasificados erróneamente. Los errores en cada iteración, son utilizados para actualizar los pesos en el conjunto de entrenamiento, dándoles más importancia a los mal clasificados. Con esta estrategia, se construyen los modelos corrigiendo los errores anteriores y, a diferencia de *bagging*, se da relevancia a los modelos con mejor comportamiento, en vez de nivelarlos a todos.

II.3.2. Inteligencia artificial

"La Inteligencia Artificial es la ciencia de construir máquinas para que hagan cosas que, si las hicieran los humanos, requerirían inteligencia" (Marvin Misnky, en Escolano *et al.*, 2003).

La inteligencia Artificial es un campo multidisciplinar cuya meta es automatizar actividades que requieren inteligencia humana (Williams, 1983). Las áreas que se manejan en la inteligencia artificial pueden ser resumidas como *Percepción*, *Manipulación, Razonamiento, Comunicación y Aprendizaje.* La percepción es la construcción de modelos del universo físico a través de los sentidos (visual, auditivo, etc.). La manipulación es la articulación de aditamentos (por ejemplo brazos mecánicos o dispositivos de locomoción) para incidir en un estado del universo físico. El razonamiento atañe con funciones de alto nivel cognitivo tales como la planeación, la esquematización de conclusiones por inferencia de un modelo del universo, el diagnóstico, el diseño, etc. La comunicación trata el problema de la comprensión y el conocimiento de la información a través del uso del lenguaje. Finalmente, el aprendizaje trata el problema de mejorar automáticamente el desempeño del sistema en el tiempo basado en la experiencia del sistema.

El objetivo científico último de la inteligencia artificial es construir un modelo formal de la *Mente* que explique completamente la actividad del "pensamiento" y su funcionamiento en los humanos.

II.3.3. Técnicas de minería de datos

A continuación, se detallan algunas de las técnicas de minería de datos de mayor impacto en la bibliografía, y se presenta una mayor profundización en aquellas que tienen una aplicación específica en el tema de los abastecimientos de agua, de acuerdo con el estado del arte del mismo. Cabe anotar que no todas estas técnicas son utilizadas en el desarrollo de esta tesis, pero se aportan por el interés que suscitan en el momento de plantearse posibles desarrollos y

continuaciones al trabajo aquí expuesto, así como para una mejor comprensión tanto de la temática expuesta como de la revisión bibliográfica de las aplicaciones en los sistemas de abastecimiento.

II.3.3.1. Técnicas estadísticas

El modelado estadístico se fundamenta en explicar el comportamiento de una variable a partir del conocimiento de otras. Se parte de la hipótesis de que una variable tiene cierta variabilidad, y que esta variabilidad está relacionada con el comportamiento de otras variables.

Este modelado corresponde a la técnica estadística de mayor uso; puede ser empleada tanto si se plantea la predicción de una cierta variable de respuesta, como si se busca encontrar un modelo de causalidad, en cuyo caso las variables explicativas son causa de la variabilidad de la respuesta.

El modelado estadístico puede ser representado como:

$$y_1 = f(x_{i1},...,x_{ip}) + \varepsilon_i \text{ ,}$$

donde f es la función que relaciona los valores de la variable de respuesta y_i con las variables explicativas x_{ij} y ε_i corresponde a la parte específica del individuo i que no está explicada por ninguna variable. La función f representa la parte determinista, que permite representar el comportamiento de la variable de respuesta y realizar predicciones sobre ella, mientras que la parte específica representa la parte impredecible y aleatoria que se denomina *el error*. Este término se determina a partir de una distribución de probabilidad generadora de los distintos valores para cada individuo, que se puede suponer centrado, es decir, el valor esperado es 0:

$$E[\varepsilon_i] = E[y_i - f(x_{i1},...,x_{ip})] = 0$$

Dentro de las modelaciones estadísticas se incluyen las regresiones lineales, los modelos aditivos generalizados (*GAM*, de sus siglas en inglés) y las regresiones

adaptativas multivariadas por *splines (MARS)*.

Se habla de modelo de regresión cuando la variable de repuesta y las variables explicativas son cuantitativas. Si se tiene una variable explicativa se habla de *regresión simple*, mientras que si se dispone de varias variables explicativas se tiene un problema de *regresión múltiple*.

Una forma simple de relacionar la respuesta con los predictores es a través del modelo (modelo de regresión múltiple lineal), expresado como:

$$E(Y) = Xb = b_0 + b_1 X_1 + ... + b_p X_p$$

Los modelos aditivos generalizados pueden acomodar los efectos de la no-linealidad de los estimadores originales:

$$G(E(Y)) = G(\mu) = b_0 + f_1(X_1) + ... + f_p X_p,$$

donde los términos *f,* son funciones arbitrariamente definidas generalmente estimadas a partir de alguna clase de suavización tal como *splines* cúbicos. Este modelo está compuesto de funciones de una sola variable, una para cada predictor, que son estimaciones no paramétricas de tendencias, manteniendo los otros predictores en el modelo constantes.

La regresión adaptativa usando *splines* corresponde a una especie de modelo aditivo, que puede ser visto como una generalización de los árboles de regresión, introducido para vencer algunas de sus limitaciones. *Friedman* (1990) describe estas limitaciones detalladamente y muestra cómo la regresión adaptativa por esplines puede vencerlos. El modelo resultante es implementado en el sistema *MARS* (*Multivariate Adaptive Regression Splines*), el cual construye modelos de regresión flexibles ajustando esplines separados o funciones base a diferentes intervalos de la variable predictora.

II.3.3.2. Algoritmos Genéticos

Los algoritmos genéticos, fueron formalmente introducidos por Holland en

1975 en la Universidad de Michigan y, son utilizados para encontrar la solución óptima de problemas realizando una búsqueda aleatoria, paralela y global, especialmente en problemas de optimización basados en la mecánica de la selección natural y la genética.

Los algoritmos genéticos trabajan realizando una selección aleatoria de combinaciones de genes *"cromosomas"* para formar la primera generación. Cada cromosoma individual en cada generación corresponde a la solución en el dominio del problema. Se utiliza una función de *aptitud* para evaluar la calidad de cada cromosoma, tal que cromosomas con calidad alta sobrevivirán y generarán la siguiente generación. Haciendo uso de los tres operadores básicos (reproducción, cruce y mutación), se recombina una nueva generación para encontrar la mejor solución. El proceso se repite continuamente hasta obtener la solución óptima o hasta que se realice un número constante de iteraciones.

El algoritmo genético es una técnica de búsqueda para encontrar máximos o mínimos globales en la solución de problemas. Si bien no se garantiza encontrar la solución óptima global, se tiene evidencia empírica de que el algoritmo genético encuentra soluciones con un nivel aceptable, y en un tiempo competitivo con respecto a otros algoritmos de optimización combinatoria. En general, la posibilidad de obtener el máximo o mínimo global utilizando un algoritmo genético está relacionada con la complejidad del problema; entre más complejo sea este, se presenta mayor dificultad en encontrar el óptimo global.

El problema de la complejidad puede ser definido por la dimensionalidad del espacio de solución del problema. Es decir, en cada parámetro por determinar, la solución será considerada como una de las dimensiones en el dominio de ésta. Por consiguiente, si se necesitan considerar bastantes parámetros para obtener la solución, la complejidad del problema será bastante alta. Por otra parte, si solo un limitado número de parámetros son evaluados para un problema, la complejidad de éste es baja. La idea se puede utilizar haciendo uso de la terminología de los algoritmos genéticos; cada gen de un cromosoma es uno de los parámetros para obtener la solución; por tanto la complejidad del problema es proporcional al

tamaño del cromosoma.

Los algoritmos genéticos corresponden a técnicas de optimización que pueden ser utilizadas para mejorar otros algoritmos de minería de datos, obteniendo como resultado el mejor modelo para una serie de datos. El modelo resultante es aplicado a los datos para descubrir patrones escondidos o para realizar predicciones.

La computación evolutiva hace referencia a la evolución de la población como tipo de optimización, en contraste con los métodos estándares de optimización no lineal que dependen de un punto único de búsqueda (que es la migración de un elemento único a través del espacio de búsqueda); la computación evolutiva explota una población entera de posibles soluciones, y las desarrolla de acuerdo a algunos principios genéticamente manejados.

Los algoritmos genéticos fueron concebidos por *Holland* como un algoritmo basado en un tipo darwiniano de estrategia de supervivencia adecuada con reproducción sexual, donde los individuos fuertes de la población tienen una probabilidad alta de crear descendencia. Un algoritmo genético es implementado cómo una búsqueda computerizada y un procedimiento de optimización que utiliza principios de genética y selección natural.

La computación de evolución es el nombre dado a un grupo de algoritmos basados en la evolución de una muestra hacia la solución de un cierto problema. La muestra de posibles soluciones evoluciona de una generación a otra, llegando a una solución satisfactoria del problema. Estos algoritmos difieren en la forma en que se generan las muestras a partir de una presente, y en la forma en que los miembros son representados dentro del algoritmo. Se han logrado diferentes técnicas informáticas de evolución, entre las cuales se pueden destacar: los *algoritmos genéticos (GA's)*, *la programación genética (GP)* (Koza, 1992 y 1994) y *los algoritmos de evolución (EA's)* (Schwefel, 1995). Los *EA's* pueden ser divididos en *Estrategia de Evolución (ES)* y *Programación de Evolución (EP)* (Aesche y Simpson, 1994), aunque los más utilizados para el descubrimiento de reglas han sido los algoritmos genéticos y la programación genética, que difieren

principalmente en cuanto a la forma como es representado un individuo o miembro.

Todos estos algoritmos comparten iguales conceptos básicos, difieren en la forma en que son codificadas las soluciones, y en los operadores que utilizan para crear una nueva generación. En un algoritmo genético los individuos se representan por una cadena de reglas de condición, donde cada condición puede ser un par atributo-atributo o atributo-valor, y se pueden representar como una regla o como un conjunto de reglas, como:

Diámetro = "alto" caudales>100 ó

Caudales > 100 Rugosidad = "baja" Diámetro = "alto"

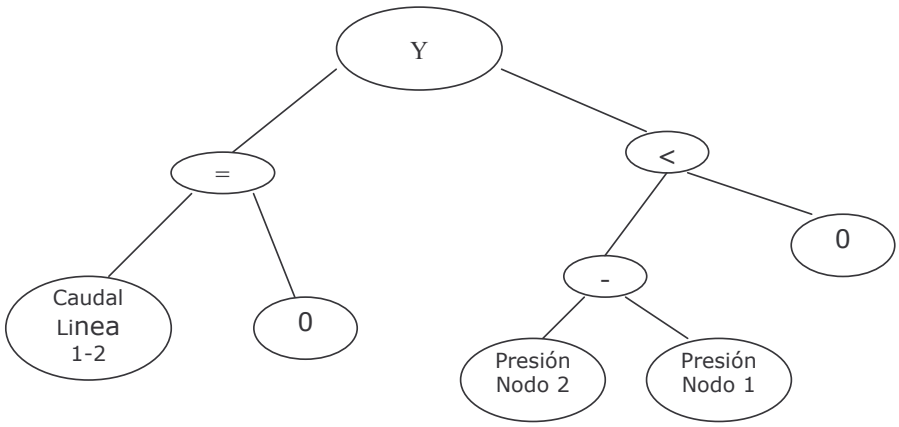


Figura II.8. Ejemplo de programación genética para descubrimiento de reglas

En la programación genética, los miembros son representados por un árbol, con reglas de condición y/o valores de atributos en las hojas de los nodos, y funciones (lógicas, relacionales u operadores matemáticos) en los nodos internos. Por ejemplo en la Figura II.8, se visualiza un árbol para representar una regla, SI (Caudal = 0) Y (diferencia de presiones en los nudos < 0), aplicándolo a un problema de fugas en tuberías, la regla consecuente podría ser (Fuga = SI).

La motivación de hacer uso de los algoritmos de evolución para el descubrimiento de reglas es que estos evalúan la regla como un todo, con el fin de adaptar la función más que evaluar el impacto de añadir o remover una condición *de* o *para* una regla.

Muchos métodos de evolución para selección de atributos están basados en un simple algoritmo genético; un individuo es una cadena binaria con *m* genes, donde *m* es el número de atributos. Cada gen puede tomar el valor 0 ó 1 indicando si el atributo correspondiente es o no seleccionado. Por ejemplo el individuo 01101000 representa una solución candidata donde el segundo, tercero y quinto atributo son seleccionados.

Los algoritmos genéticos han sido exitosamente utilizados para optimizar parámetros de algunos tipos de algoritmos de *KDD* (Santos *et al.*, 2000; Hruschka y Ebecken 2000), por ejemplo para optimizar los pesos de un clasificador de vecindad próxima (*K-nearest neighbour*), o como optimización de conjuntos borrosos utilizados para la construcción de árboles de decisión.

Más formalmente, un algoritmo genético está definido como un cuádruplo

$$GA = \{\Xi, \Pi, apt, \Omega\}$$

donde los componentes de este esquema corresponden a:

- Ξ es el mecanismo de descodificación
- Π corresponde a una población de tamaño finito (N)
- apt es una función de idoneidad, aptitud, o adaptación (*fitness)* asignando un cierto valor de utilidad a cada cadena. Su objetivo es distinguir soluciones buenas, de aquellas que no lo son.
- Ω es un conjunto finito de operaciones genéticas (reproducción, cruzamiento, mutación)

El poder real de esta técnica es la habilidad para tratar con problemas altamente no lineales, con parámetros no continuos, y objetivos complejos de optimización. Son nueve los pasos básicos para la implementación de un problema con algoritmos genéticos (Vitkovsky y Simpson 1997):

1. Codificación de una cadena

2. Generación de la población inicial
3. Cálculo de los costos de la cadena
4. Análisis de cada cadena
5. Asignación de penalizaciones de costos
6. Cálculo de costos totales
7. Cálculo de la habilidad de la cadena
8. Generación de nuevas poblaciones utilizando operadores genéticos
9. Producción de generaciones sucesivas

La codificación de una cadena o cromosoma define las características de soluciones particulares; cada bit o grupo de bits en una cadena corresponde directamente a un parámetro en la definición del problema. Entre otros, los tipos de codificación son: codificación binaria, codificación entera, codificación real.

La generación de poblaciones se realiza aplicando un número generador uniforme aleatorio que genere un conjunto inicial de cadenas. Un estimativo del tamaño inicial de la población puede ser obtenido por procesos de ensayo y error, donde el mejor tamaño de población es aquel para el cual cualquier incremento de tamaño no produce, en promedio, mejores soluciones.

El costo de una solución particular se puede calcular descodificando la cadena por medio de una tabla. Usando los parámetros de descodificación y sus costos asociados se puede encontrar el costo material de una cadena. Para cada cadena o posible solución se hace necesario probar si la solución viola cualquier restricción.

Hay tres vías diferentes de penalizar cualquier cadena que no se adapte a las restricciones del problema: eliminación, penalización alta, y penalización moderada. Un método común de determinar el costo de penalización es por el método de tanteos, donde son encontrados algunos límites para la penalización, lo que maximiza la diversidad produciendo un número aumentado de soluciones factibles. El costo total para una cadena particular en la población debe ser

tomado como la suma del costo material y el costo de penalización.

El cálculo de la capacidad de una cadena permite que una comparación numérica sea hecha para probar cómo aquella cadena es medida entre otras cadenas en la población. Una vez los ajustes de cada cadena en la población han sido determinados, se pueden utilizar operadores genéticos para generar nuevas poblaciones. Hay tres operadores genéticos principales cada uno con un papel importante para maximizar el flujo de material genético hacia una solución óptima. Estos son:

- Selección o reproducción
- Entrecruzamientos
- Mutación

Los pasos 3 a 8 representan la producción de generaciones sucesivas en la ejecución de un algoritmo genético; estos pasos son repetidos muchas veces (100 a 1000) para garantizar que los bajos costos o ajuste de soluciones se desarrollen. Generalmente, el tamaño del problema, complejidad, y número de variables de decisión determina el número total de generaciones necesarias. Para probar si un algoritmo converge, se revisa si el costo promedio de las mejores soluciones cambia dramáticamente con la próxima generación; si esto no sucede se puede decir que la solución converge.

Operadores Genéticos

Operadores de reproducción o selección

Al igual que en la naturaleza, una vez son calificados todos los individuos de una generación, el algoritmo debe seleccionar los individuos más capacitados, mejor adaptados al medio para tener mayores posibilidades de reproducción

Como su nombre lo indica, la estrategia de evolución procede del análisis de un gran número de generaciones, cada una con un tamaño de población

constante. La creación de nuevos individuos en nuevas generaciones sigue los operadores que envuelven la evolución biológica. Esencialmente los descendientes son creados como una combinación de genes de sus padres.

Crear una nueva población para la próxima generación da ocasión a la necesidad de algún tipo de esquema de selección. Se tienen cuatro tipos principales de esquema: selección proporcionada, selección por concurso, clasificación lineal, y selección Genitor (fija).

El esquema de selección proporcionada opera similar a una ruleta, donde se seleccionan soluciones al azar en razón de su capacidad con relación a la capacidad de todas las otras soluciones en la generación actual. Un inconveniente de este esquema es que, después de varias generaciones, la capacidad de diferentes soluciones es muy similar; por tanto, son virtualmente la misma y causa que el esquema se revierta, y posibles buenas soluciones podrían perderse debido a la presión de selección.

El esquema de selección por concurso empareja todas las cadenas al azar y la cadena con capacidad más alta, en el par, progresa a la siguiente generación. El proceso de clasificación lineal comienza ordenando la población de mejor a peor capacidad; a cada cadena se le asigna un número de copias de acuerdo con una función de asignación no creciente; luego se aplica un proceso de selección proporcionada usando cada asignación; lineal se refiere a la función de asignación.

El operador Genitor trabaja sobre individuos, más que a una escala de población. Un descendiente es buscado de acuerdo a una clasificación lineal, y el peor individuo es reemplazado con un descendiente.

Operadores de entrecruzamiento

Un operador de entrecruzamiento es la permutación parcial de bits entre dos cadenas padre (intercambio de genes entre cromosomas homólogos durante la meiosis).

El operador al azar empareja cadenas de la población; luego prueba si el cambio ocurrirá basado en la probabilidad de entrecruzamiento. Existen varios tipos de operadores de cruzamiento, algunos de los cuales son: Cruzamiento de un punto, cruzamiento de dos puntos, cruzamiento uniforme, y cruzamiento promedio.

Un cruce de un punto identifica dos cadenas de la población y aleatoriamente selecciona una posición en la cadena con la cual intercambia su contenido, tal como se visualiza en la Figura II.9.

La operación de entrecruzamiento incrementa la diversidad de la población. La intensidad del cruce se caracteriza en términos de la probabilidad en que se ven afectados los miembros de una población. Valores altos de probabilidad indican que más individuos son afectados por la operación.

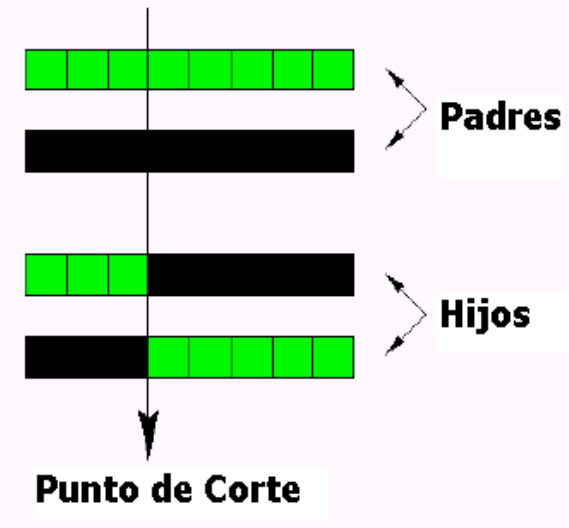


Figura II.9. Cruzamiento de un punto (one-point crossover)

Operadores de Mutación

El objetivo del operador de mutación es asegurarse de que ningún material genético importante es pasado por alto durante la ejecución del algoritmo genético. Cada cadena en la población es probada sobre una probabilidad base, para determinar si será mutada o no. Existen tres grupos distintos de operadores de mutación: creación de genomas, mutación aleatoria, y mutación por

deslizamiento.

Figura II.10. Mecanismos de mutación en cadenas binarias

La mutación añade diversidad extra de naturaleza estocástica. En cadenas binarias este mecanismo es implementado por la inversión de los valores de algunos *bits* seleccionados aleatoriamente (ver Figura II.10). La tasa de mutación está relacionada con la probabilidad en la cual el *bit* es afectado. Por ejemplo, una tasa de mutación del 5% aplicada a una población de 50 cadenas cada una con 20 *bits* significa que el 5% de los 1000 *bits* ha cambiado.

Para finalizar este apartado, podemos decir que el objetivo principal de una optimización operacional, es ofrecer un nivel aceptable de servicio a los clientes dentro de las restricciones del sistema y las regulaciones legales, minimizando los costos operacionales (Van Zyl et. *al,* 2004). Estas metas, entran en conflicto a largo plazo, ya que los intentos por minimizar los costos operacionales generan sitios en el sistema con mayor estado de vulnerabilidad, y menos capaces para manejar anormalidades, tales como las roturas en las tuberías, reduciendo así el nivel de servicio; por tanto cuando se aplica a los sistemas de distribución de agua, en una situación de emergencia es más importante encontrar una buena solución rápidamente que encontrar la solución óptima global.

En los últimos años esta técnica de optimización ha tenido un gran interés dentro de la comunidad científica, y en el tema de los abastecimientos de agua se han realizado desarrollos en diversas aplicaciones, Murphy y Simpson (1992), Aesche y Simpson (1994), Wu y Simpson (1996), Reis *et al.,* (1997), Montesinos *et al.*, (1999), Cisty (2000), Dandy y Engelhardt (2001), Van Zyl *et al.* (2004).

II.3.3.3. Particle Swarm Optimization (PSO)

Particle Swarm Optimization (PSO) es una técnica de optimización evolucionaria bien fundamentada y establecida. Aunque originalmente esta técnica fue diseñada para tratar con variables continuas, la variante considerada de PSO supera tres debilidades de la técnica de optimización; la utilización de variables discretas y continuas en el problema de optimización; la dificultad para mantener buenos niveles en la diversidad de la población y, por tanto, el balance entre búsquedas locales y globales; y finalmente, la elección de los valores correctos para los parámetros se lleva a cabo a través de un control de parámetros dinámico y auto-adaptativo (Izquierdo *et al.,* 2009a, Montalvo *et al.,* 2010).

PSO es una técnica de optimización de la población basada en estocástica desarrollada por Kenneddy y Eberhart en 1995, inspirada en el comportamiento social de bandadas de pájaros o bancos de peces; y ha mostrado gran potencial y buenas perspectivas en la solución de diversidad de problemas de optimización (Herrera *et al.,* 2009, Izquierdo *et al.,* 2008b y 2009b, Montalvo *et al.,* 2008b). Adicionalmente, la PSO puede ser fácilmente implementada y es computacionalmente poco costosa, puesto que, los requerimientos de memoria y velocidad de procesador son bajos.

Esta técnica de optimización comparte similitudes con las técnicas de computación evolutiva, tales como los algoritmos genéticos. El sistema se inicia con una población de soluciones al azar y busca el óptimo actualizando las generaciones. A diferencia de los algoritmos genéticos, en la PSO no se tienen operadores de evolución como los cruces y las mutaciones. En la PSO, las posibles soluciones, llamadas partículas, vuelan a través del espacio del problema siguiendo a la partícula actual óptima; por tanto una partícula con una idoneidad baja puede sobrevivir durante el proceso de optimización y, por consiguiente, el método puede potencialmente visitar cualquier punto del espacio de búsqueda. La PSO puede fácilmente manejar la no-linealidad y no-convexidad del dominio del problema; la búsqueda no depende de la población inicial, y supera el problema de óptimos locales que es común en algunas técnicas convencionales de

optimización (Afshar y Rajabpour, 2009).

La PSO es un sistema de optimización multiagente inspirado en la metáfora del comportamiento social. En la PSO, cada agente llamado partícula, vuela en un espacio *D*-dimensional de acuerdo con su experiencia histórica y aquella de sus compañeros. El movimiento de las partículas es estocástico. La PSO mantiene la huella de la mejor solución alcanzada hasta el momento para cada partícula, así como la mejor solución para todas las partículas alcanzada hasta el momento. Al finalizar una iteración de entrenamiento, la PSO cambia la velocidad de cada partícula hacia su mejor posición y la mejor posición lograda por el enjambre.

Suponiendo que el espacio de búsqueda es *D*- dimensional, entonces la *i-ésima* partícula del enjambre puede ser representada por un vector *D*-dimensional, $X_i = \left(x_{i1}, x_{i2}, ..., x_{iD}\right)^T$. La velocidad (cambio de posición) de esta partícula, se puede representar como otro vector *D*-dimensional, $V_i = \left(v_{i1}, v_{i2}, v_{iD}\right)^T$. Cada partícula recuerda su mejor posición entre aquellas que ha visitado; se denomina como $pbest_i$. Esta información corresponde a una analogía de experiencias personales de cada agente. Sin embargo, cada agente conoce el mejor valor alcanzado en el grupo, $gbest$ entre $pbest$. Esta información corresponde a una analogía del conocimiento de cómo otros agentes alrededor suyo se han desempeñado. Cada agente intenta modificar su posición, haciendo uso de la siguiente información:

- la posición actual X_i;
- la velocidad actual V_i;
- la distancia entre la posición actual X_i y $pbest$;
- la distancia entre la posición actual X_i y $gbest$.

Esta modificación, puede ser representada por el concepto de velocidad. La velocidad de cada agente es, por consiguiente, modificada por la siguiente ecuación:

$$V_i^{n+1} = wV_i^n + c_1 r_1 \times \left(pbest_i - X_i^n \right) + c_2 r_2 \times \left(gbest - x_i^n \right) \quad (1)$$

La posición actual (punto buscado en el espacio de solución), es modificada por la siguiente ecuación:

$$X_i^{n+1} = X_i^n + V_i^{n+1} \quad (2)$$

donde V_i^n = vector velocidad del agente i en la iteración n; w = peso de inercia; c_1, c_2 = dos constantes positivas, llamados parámetros cognitivo y social, respectivamente; r_1, r_2 = números aleatorios independientes uniformemente distribuidos entre 0 y 1; X_i^n la posición actual del agente i en la iteración n; $pbest_i^n$ = la mejor posición alcanzada por un agente i en la iteración n; $gbest^n$ = mejor posición global del grupo en la iteración n; i = 1,2,...,*N*, y *N* el tamaño del enjambre. Obsérvese, que la posición de la partícula X_i, la velocidad de la partícula V_i, la mejor posición de la partícula $pbest_i$ y la mejor experiencia de enjambre $gbest$ son todos vectores de dimensión *D.*

Las dos ecuaciones anteriores definen la versión inicial del algoritmo PSO. Desde el desarrollo básico propuesto por Kennedy y Eberhart (1995), muchos avances se han realizado para introducir nuevas ideas para mejorar la eficiencia y efectividad de la PSO para la solución de problemas de optimización en ingeniería. Estos esfuerzos, se han concentrado principalmente en la selección de parámetros y la introducción de nuevos mecanismos para el balance propio de exploraciones y explotaciones inherentes del algoritmo (Afshar y Rajabpour, 2009). Uno de los más exitosos de estos esfuerzos fue el desarrollo de PSO con vecinos próximos locales. De acuerdo con la variante global original definida anteriormente, cada partícula se mueve hacía su mejor posición previa y hacia la mejor partícula de todo el enjambre. De acuerdo con la variante local, sin embargo, cada partícula se mueve hacia su mejor posición previa y hacia la mejor partícula en su restringida vecindad. En la versión local del PSO, la velocidad de una partícula se define como (Eberhart *et al.*, 1995):

$$V_i^{n+1} = wV_i^n + c_1 r_1 \times \left(pbest_i - X_i^n \right) + c_2 r_2 \times \left(lbest_i - X_i^n \right) \quad (3)$$

En las primeras experiencias con PSO, un valor máximo $v_{máx}$, fue utilizado para controlar la velocidad de la partícula. Si la velocidad excedía este umbral, se establecía como igual a $v_{máx}$. Esto fue necesario porque valores grandes podían resultar moviéndose en soluciones buenas pasadas, mientras que valores pequeños podrían resultar en exploración insuficiente del espacio de búsqueda. El papel del peso de inercia, w de las primeras ecuaciones, se observó como crítico para el comportamiento de convergencia de la PSO. La función del peso de inercia es controlar el impacto de la historia previa de velocidades en la actualidad. Por consiguiente, el parámetro w regula la compensación entre las habilidades locales y globales del enjambre. Un gran peso de inercia, da como resultado una exploración global (búsqueda de nuevas áreas), mientras que uno pequeño tiende a facilitar la exploración local, es decir, una puesta a punto fina del área de búsqueda actual. Un valor apropiado del peso de inercia w, puede ser considerado como un balance entre la exploración global y local, y que consecuentemente resulte en la reducción del número de iteraciones requeridas para localizar la solución óptima. Inicialmente, este peso de inercia fue constante; sin embargo resultados experimentales indican que es mejor iniciar con un valor grande, para promover la exploración global en el espacio de búsqueda, y gradualmente decrecer para conseguir soluciones más refinadas. Así, Shi y Eberhart (1999), realizan un mejoramiento en el desempeño de la PSO con un peso de inercia variante sobre las generaciones, el cual varía linealmente de $w_{máx}$ en el inicio de la búsqueda, a $w_{mín}$ al final, de acuerdo con la siguiente formulación:

$$w = w_{máx} - \frac{\left(w_{máx} - w_{mín} \right) \times n}{iter_{máx}} \quad (4)$$

donde w_{max} = peso inicial; w_{min} = peso final; $iter_{max}$ = numero de iteración máximo; n = número de iteración actual. No obstante, hemos tenido mejores resultados haciendo uso de la ecuación 5.

Los parámetros c_1 y c_2 de las dos primeras ecuaciones, no son criticas para la convergencia de la PSO (Afshar y Rajabpour, 2009). Sin embargo, una buena puesta a punto puede resultar en convergencias rápidas y alivio de locales mínimos (Clerc, 1999). Eberhart y Shi (2000) concluyen que los resultados del algoritmo PSO pueden ser mejorados, seleccionando cuidadosamente el peso de inercia w, y los parámetros c_1 y c_2.

Haciendo uso de las tres primeras ecuaciones, las velocidades y las posiciones de las partículas son actualizadas repetidamente sobre iteraciones para obtener la solución óptima. Aunque la PSO es rápida en encontrar soluciones de calidad comparada con otras técnicas de evolución, se presentan dificultades para obtener soluciones de mejor calidad mientras se exploran funciones complejas. Se pueden encontrar convergencias prematuras, y sufrir de una pobre capacidad de puesta a punto para la solución final.

II.3.3.4. Lógica Fuzzy (Borrosa o Difusa)

Para formalizar proposiciones como, "*la caída de presión ha sido moderada*", o "*el nivel de cloro en la red es muy alto*", o "*se requiere de un suministro nunca menor a 10 L/s, ni más alto de 30 L/s, y a menudo entre 15 y 20 L/s*", se hace uso de la lógica *fuzzy o "difusa"*, que corresponde a una ampliación de la lógica clásica para definir conceptos imprecisos. Es decir, la lógica *difusa* es una técnica de múltiples valores que permite una gradación continua del valor real de una proposición, utilizando cualquier valor en el intervalo [0,1]. La teoría fue formulada con el propósito específico de describir matemáticamente información imprecisa, e interpretar conceptos definidos por expresiones lingüísticas, como las anteriores, dentro de un modelo matemático.

Actualmente, la lógica *difusa* es utilizada en el desarrollo de sistemas expertos difusos. El número de aplicaciones en la ingeniería y control industrial es elevado, ya que los sistemas difusos permiten, haciendo uso de reglas sencillas, un "control suave" de los procesos (gracias a la gradación de valores).

Generalidades

La lógica *difusa* maneja conceptos imprecisos (como pequeño, grande, joven, viejo, alto, bajo) y es más flexible que otras técnicas. Proporciona la noción de un conjunto borroso más que una clara demarcación de límites; por ejemplo, en vez de sólo 0 ó 1 se tiene también 0.9, 0.85, 0.93, 0.21, 0.05 etc.

La palabra "fuzzy" proviene del inglés "fuzz", que hace referencia al plumón de los pollos recién nacidos, cuyos contornos se muestran aún muy imprecisos. El significado de la palabra "fuzzy" se puede relacionar con algo difuso, ambiguo, vago (Hernández y López, 1997).

Ya en la Grecia antigua hubo discusiones entre partidarios de diferentes tipos de lógica. *Aristóteles, Parménides* y sus seguidores (unos 400 años antes de Cristo) aportaron su esfuerzo para desarrollar una teoría de la lógica que fuese precisa. Para ello desarrollaron una serie de leyes. Una de ellas establece que toda proposición solo puede ser verdadera o falsa. Sin embargo, había partidarios de lo contrario. Por ejemplo, *Heráclito* propuso que las cosas podían ser simultáneamente verdaderas y no verdaderas. *Platón* propone que existe un tercer campo (que él llama "más allá de lo verdadero y lo falso").

Finalmente, es la lógica propuesta por *Aristóteles* la que prevalece a lo largo de los siglos en el pensamiento occidental. A principios del siglo XX es *Jan Łukasiewicz* (1878-1956), quien describe una lógica basada en tres estados. Al tercer estado (llamado *"posible")* le asigna un valor entre el verdadero y el falso.

Definiciones

En 1965, el profesor *Lotfi A. Zadeh*, publicó su trabajo *"Fuzzy Sets"* (conjuntos *difusos*). En el mismo, Zadeh propone una lógica basada en un número infinito de estados entre los valores de verdadero y falso (1= verdadero, 0= falso). Hoy se considera a Lofti A. Zadeh, profesor de Ciencias de la Computación en la Universidad de California (Berkeley), el padre de la Lógica Fuzzy.

La lógica *difusa* se fundamenta en un razonamiento matemático que permite calcular de forma exacta las magnitudes correspondientes a conceptos vagos o situaciones poco previsibles, para poder tener control sobre ellos.

Un conjunto *difuso* es un conjunto de elementos que cumplen total o parcialmente con una característica. El grado de pertenencia de cada elemento al conjunto viene dado por el hecho de que cumplan en mayor o menor medida con la condición. La lógica *difusa* se basa en que cualquier elemento concreto puede pertenecer a varios conjuntos en diferente grado. Este grado de pertenencia varía entre 0 y 1. Una diferencia entre la lógica clásica y la lógica *difusa* es que en la primera el valor es 0 ó 1 necesariamente (verdadero o falso), mientras en la segunda el valor está entre 0 y 1.

Por tanto, el conjunto *difuso* (*Zadeh*, 1965) está caracterizado por una función que asocia los elementos del dominio, espacio, o universo de discurso $\boldsymbol{X}$ al intervalo de unidad [0,1], esto es:

$$A : \boldsymbol{X} \rightarrow [0,1]$$

Así, un conjunto *difuso A* definido en $\boldsymbol{X}$, puede estar representado por un conjunto de pares ordenados $x \in \boldsymbol{X}$ y su grado de asociación, denotándolo por $A = \{A(x) \mid x \in \boldsymbol{X}\}$. Un conjunto *difuso* es una generalización del concepto de un conjunto cuya función de asociación toma la forma *si – no* de cuantificación. El valor de $A(x)$ es el grado de asociación de un elemento x en A. Por ejemplo, al considerar el concepto de alta temperatura, en el contexto de un ambiente con temperaturas entre $\boldsymbol{X} = [0,50]$ definidas en ºC, el grado de asociación de 0ºC, A(0ºC), a la clase de temperaturas altas es *cero*; por otra parte, para 50ºC A(50ºC) está dentro de la clase altas temperaturas, y se le puede asignar un grado de compatibilidad total con el concepto.

En la lógica *difusa* cada variable se representa por múltiples opciones y grados. Relacionando la variable con sus grados de pertenencia se desarrolla la función de pertenencia. Un conjunto expresado en estos términos se denomina conjunto *difuso*, y se llama etiqueta a un conjunto *difuso* determinado.

Por ejemplo, la temperatura puede ser una variable la cual se ha acotado por una serie de rangos según un criterio propio (alta, media, baja, muy baja) que son las etiquetas de esta variable. Habitualmente se utilizan por cada variable entre 3 y 7 etiquetas, según las necesidades de precisión y complejidad del sistema de control.

Los conjuntos *difusos* pueden estar definidos en universos finitos o infinitos, por tanto se pueden encontrar diferentes notaciones en la literatura. Si un universo ***X*** es discreto y finito, con cardinalidad[2] *n*, entonces un conjunto *difuso* esta dado en forma de un vector *n*-dimensional cuyas entradas indican el grado de asociación de los elementos correspondientes de ***X***.

Un controlador *difuso* es la aplicación más importante de la teoría *difusa*. Se trata de un conjunto de funciones de pertenencia y reglas que se utilizan para razonar con datos.

Los pasos que sigue un control *difuso* son los siguientes:

Establecer funciones de pertenencia: se trata de dar valores a las variables de entrada y salida. Por ejemplo, si se tienen valores de temperatura y presión como variables de entrada, y la apertura de válvula como variable de salida, estableciendo rangos convenientes al proceso para la temperatura (caliente, normal, fría, etc.), para la presión (alta, medio alta, correcta, medio baja, baja), y para la apertura (total, bastante, media, poco, casi cerrada, cerrada), a cada una de las etiquetas se le dan valores numéricos y se establecen así las funciones de pertenencia.

Establecer las reglas: depende de la experiencia y conocimiento que se tenga del funcionamiento del proceso.

---

[2] Dado un conjunto difuso *A* en un universo finito ***X***, su cardinalidad designada por card(*A*), está definida por la suma de todas sus funciones de asociación

$$card(A) = \sum_{x \in X} A(x)$$

Por ejemplo:

Si temperatura = normal, y presión = correcta, entonces, válvula = poco;

Si temperatura = caliente, y presión = medio alta, entonces, válvula = bastante

Se deben establecer las reglas que se consideren necesarias para un buen control. Lo más habitual es un control de 9 a 12 reglas para un resultado aceptable (en términos generales).

Fuzzificación: etapa de antecedentes. En este paso el controlador estudia las reglas cuyos antecedentes se ven implicados (según los valores de temperatura y presión en el ejemplo) y las pone en juego, descartando aquellas que no han sido pensadas para esa circunstancia concreta. Es decir, si un antecedente no se cumple, se cancela la acción de la regla.

Inferencia: en este paso se procede a la asignación de los valores de las funciones de pertenencia de cada regla, se valora la acción que se debe emprender.

Defuzzificación: en esta etapa se traducen los valores *difusos*, es decir los valores asignados a los grados de pertenencia de los diferentes elementos (temperatura, presión, etc.), a valores concretos válidos para la salida del sistema de control, ya que un motor o una válvula no trabajan con grados de pertenencia, sino con valores concretos de magnitudes físicas. Esta conversión y evaluación del tipo de salida se denomina *"defuzzificación"*. Esta defuzzificación incluye el último paso donde se decide la calidad de salida que queremos que se adopte. Para ello se tienen dos métodos más utilizados, el del máximo (utilizar como salida el mayor valor) y el del centro de gravedad, en el cual se obtiene una media ponderada de todos los valores posible de salida (Stout, 2001).

Tipos de funciones de pertenencia

La función de pertenencia es el componente crucial de un conjunto difuso; por consiguiente, todas las operaciones con conjuntos difusos son definidas a

través de sus funciones de pertenencia (El Boroudy y Simonovic, 2006). Una función de pertenencia no corresponde a una función de distribución de probabilidad; por tanto, no se le puede dar un significado estadístico (Revelli y Ridolfi, 2002). La función de pertenencia indica cuánto "pertenece" un valor a un conjunto difuso, o qué tan cercano está a su valor más probable. Un número difuso, por consiguiente, tiene menos información que una distribución estadística; sin embargo, la ventaja de esta aproximación no es inventar información cuando no está disponible, sino intentar trasladar la información cualitativa que esta disponible en lenguaje matemático.

Tal como se mencionó anteriormente, la definición básica de un conjunto difuso es:

$$A : \boldsymbol{X} \rightarrow [0,1]$$

donde:

A es el conjunto difuso en el universo de discurso ***X;*** y

X es el dominio, o el universo de discurso.

La función de la ecuación anterior describe la función de pertenencia asociada con un conjunto difuso *A.* Se dice que un *conjunto difuso* es normal si por lo menos uno de sus elementos tiene un valor de pertenencia de uno (1). Un *conjunto difuso convexo* ***Z*** es un conjunto en el cual para cada número real *a,b,c* con *a*<*b*<*c* se mantiene lo siguiente:

$$Z(b) \geq \min(z(a), z(c))$$

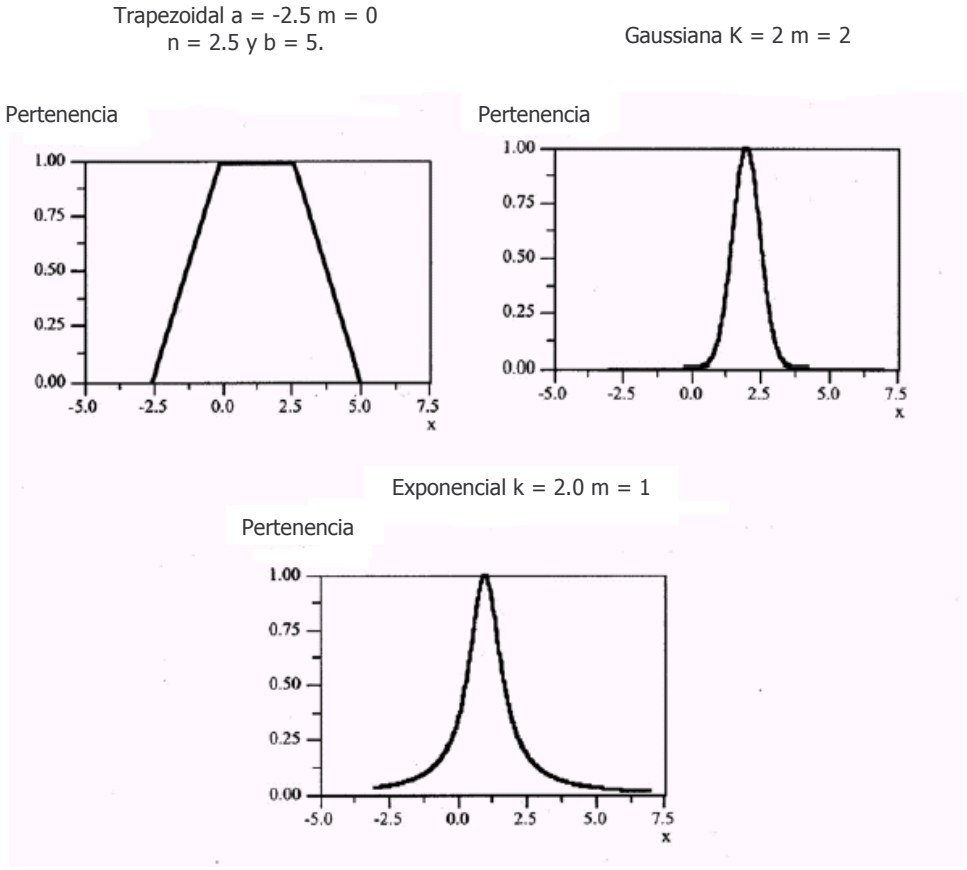
donde:

Z() es el valor de pertenencia; y

min( ) es la función mínimo.

Esta función puede tener diferentes formas y puede ser continua o discreta, dependiendo el contexto en el que se utilice. La Figura II.11 visualiza tres tipos

diferentes de funciones de pertenencia continuas.



Fuente: Cios *et al.*, 1998.

Figura II.11. Funciones de Pertenencia Trapezoidal, Gaussiana y Exponencial

Alguna de las funciones de pertenencia comúnmente utilizadas son las siguientes:

<u>Funciones Triangulares de pertenencia</u>

$$A(x) = \begin{cases} 0 & si\ x \le a \\ \dfrac{x-a}{m-a} & si\ x \in [a,m] \\ \dfrac{b-x}{b-m} & si\ x \in [m,b] \\ 0 & si\ \ x \ge b \end{cases}$$

Aquí *m* es un valor modal, mientras que los límites inferior y superior para los valores diferentes de cero de *A*(*x*), son nombrados por "*a*" y "*b*"

respectivamente.

Una notación equivalente, y algunas veces más conveniente resaltando explícitamente los parámetros de la función de pertenencia, es:

$$A(x;a,m,b) = \max\{\min[(x-a)/(m-a),(b-x)/(b-m)],0\}$$

Los tres parámetros de las funciones de asociación triangular asumen una interpretación sencilla. El límite inferior y superior identifican las regiones donde hay grados de asociaciones de funciones diferentes de cero, mientras que el valor modal identifica el elemento más probable del conjunto difuso.

Funciones S de pertenencia

$$A(x) = \begin{cases} 0 & si\ x \le a \\ 2\left(\dfrac{x-a}{m-a}\right)^2 & si\ x \in [a,m] \\ 1-2\left(\dfrac{b-x}{b-m}\right)^2 & si\ x \in [m,b] \\ 1 & si\ \ x \ge b \end{cases}$$

El punto $m = \dfrac{a+b}{2}$ es conocido como la diagonal de la función *S*.

Funciones Trapezoidales de pertenencia

$$A(x) = \begin{cases} 0 & si\ x < a \\ \dfrac{x-a}{m-a} & si\ x \in [a,m] \\ 1 & si\ x \in [m,n] \\ \dfrac{b-x}{b-n} & si\ x \in [n,b] \\ 1 & si\ \ x > b \end{cases}$$

Una notación equivalente es de la forma:

$$A(x;a,m,n,b) = \max\{\min[(x-a)/(m-a), 1, (b-x)/(b-n)], 0\}$$

Función Gaussiana de pertenencia

$$A(x) = e^{-k(x-m)^2}$$

donde $k > 0$.

Aplicaciones de Lógica Difusa

Las situaciones más comunes en las que conviene usar equipos o software que incluyen lógica difusa son las siguientes (Castro *et al.*, 2001, Gómez *et al.*, 2002, Hernández *et al.*, 1997, Kok *et al.*, 2001, Zhou *et al.*, 2001):

- Para procesos complejos, para los que no existe un modelo de solución sencillo.
- Para procesos no lineales.
- Para el caso en el que haya que introducir la experiencia de un operador experto (utilizando expresiones del tipo "demasiada presión" "poca rugosidad", etc.)
- Cuando ciertas partes del sistema a controlar son desconocidas, no medibles (procesos biológicos, químicos), con medidas no fiables (con posibles errores).
- Cuando el ajuste de una variable puede producir el desajuste de otras.

Entre las ventajas de un sistema de control difuso se tienen:

- La necesidad de menos reglas y variables, la utilización de expresiones lingüísticas y no descripciones numéricas, le relación entre las entradas y las salidas para cualquier tipo de aparato sin necesidad de conocer su funcionamiento interno.
- El uso de expresiones permite introducir en el sistema de reglas,

condiciones imprecisas de razonamiento resultando, así, modelos más intuitivos y de mejor comportamiento.

- Robustez del sistema, ya que atiende a muchas variables y a través de varias reglas. Al ser el resultado del conjunto de reglas el que incide sobre la salida del sistema, ese no se verá enormemente afectado si se produce una perturbación. El control difuso presenta una alta tolerancia al ruido (entendido como cambios).
- Al ser muy robusto, mantiene la estabilidad, aún en el caso de una caída del sistema. Esta se produciría lentamente, dando tiempo a tomar las medidas necesarias.
- Un sistema de este tipo no necesita un modelo matemático preciso del ente a controlar.
- Los controladores difusos permiten tanta o más precisión que los no difusos. En muchas aplicaciones alcanzan más rápido la estabilidad en etapas transitorias. Todo ello indica, en general, iguales o mejores prestaciones para un coste y complejidad menor.

Algunas de las aplicaciones de la lógica difusa que han aportado grandes mejoras en diversos campos son (Hernández y López 1997):

- Control automático de compuertas para plantas hidroeléctricas de producción de energía.
- Control simplificado de robots.
- Seguimiento de elementos por videocámara para la trasmisión de eventos deportivos.
- Previsión de fluctuaciones de temperatura en un sistema de aire acondicionado.
- Control estable y eficiente de vehículos.
- Control de tráfico.
- Mayor eficacia y optimización de funciones en aplicaciones de control industrial.
- Posicionamiento de elementos en la producción de semiconductores.
- Optimización de horario de servicio de autobuses.
- Sistemas de archivo para documentos.

- Sistemas de predicción de terremotos.
- Tecnología médica: diagnóstico del cáncer.
- Combinación de lógica difusa y redes neuronales.
- Reconocimiento de escritura manuscrita en ordenadores portátiles.
- Reconocimiento de elementos a enfocar en videocámaras.
- Control de luminosidad de fondo en videocámaras.
- Compensación de vibraciones en videocámaras.
- Control para máquinas lavadoras.
- Navegación con helicópteros.
- Simulación de procesos legales.
- Diseño de software de procesos industriales.
- Control de velocidad y temperatura en maquinaria de acerías.
- Control de transporte suburbano (metro) para mejorar la conducción, precisión en las paradas y el consumo de energía.
- Reducción del consumo de combustible en vehículos.
- Mejora de la eficacia y suavidad de ascensores.
- Mejora de la seguridad en reactores nucleares.
- Detección de errores.
- Control de maniobras de aproximación y acoplamiento entre vehículos espaciales.
- Proceso y reconocimiento de imágenes.
- Bases de datos difusos.
- Diseño asistido por ordenador.

II.3.3.5. Conjuntos aproximados (Rough Sets)

La base de la teoría de los conjuntos aproximados está en la suposición de que cada elemento del *universo de discurso* tenga rasgos característicos, los cuales son presentados por información (conocimiento, datos) acerca del elemento, (Pawlak y Slowinski 1994, y Pawlak 1998, 1999, y 2001). Los elementos que tienen las mismas características son indiscernibles. La teoría ofrece herramientas matemáticas para descubrir patrones escondidos en los datos, identifica dependencias parciales o totales, es decir relaciones causa – efecto en bases de datos, elimina redundancia en los datos, da aproximaciones a valores

nulos o inválidos, datos perdidos, datos dinámicos, etc.

La teoría de conjuntos aproximados es adecuada para problemas que pueden ser formulados como tareas de clasificación, y ha ganado un interés científico significativo como estructura de minería de datos y KDD (Ohrn, 1999). Es complementaria a los métodos estadísticos de inferencia, y ofrece el formalismo necesario para conducir el análisis de datos y el descubrimiento de conocimiento de datos imprecisos y ambiguos.

La teoría de los conjuntos aproximados tiene enlaces con métodos de razonamiento booleano, estadística, morfología matemática, y puede ser utilizada en combinación con otras técnicas como conjuntos difusos o borrosos, algoritmos genéticos, métodos estadísticos, redes neuronales, etc., (Adjei *et al.*, 2001, Düntsch y Gediga, 1997a, Gryzmala-Busse y Ziarko, 2000).

Algunas clases o categorías de elementos en un sistema de información no pueden ser distinguidas en términos de los atributos disponibles, solo pueden ser aproximadamente definidas. La idea de los conjuntos aproximados está basada en las relaciones de equivalencia que dividen un conjunto de datos en clases de equivalencia, y consiste en la aproximación de un conjunto por un par de conjuntos, llamados ***aproximaciones inferior y superior***. La *aproximación inferior* de un conjunto de elementos (concepto) contiene todos los elementos que, basados en el conocimiento de conjunto de atributos dado, pueden ser clasificados como pertenecientes al concepto. La *aproximación superior* de un conjunto contiene todos los elementos que no pueden ser clasificados categóricamente como pertenecientes al concepto; por tanto un conjunto aproximado es definido como la aproximación de un conjunto definido por un par de conjuntos, la aproximación superior e inferior del conjunto.

En la teoría de los conjuntos aproximados la información es representada en forma de *tabla*. Esta tabla está formada por *elementos o casos y* por *atributos*. Las entradas en la tabla son los valores categóricos de las características (atributos), y para algunos sistemas de información también clases asociadas

(categorías).

La teoría de los conjuntos aproximados puede ser utilizada en varias etapas del procesamiento de información. Por ejemplo, pueden ser utilizados para (Cios *et al.*, 2000):

- Organización de tablas de decisión, representando información que contiene datos inciertos o imprecisos.
- Análisis de conocimiento.
- Análisis de consistencia y conflictos en la serie o conjunto de datos.
- Cálculo de los conjuntos de aproximación inferior y superior.
- Identificación y evaluación de dependencias de un conjunto de atributos.
- Cálculo del poder discriminatorio del atributo.
- Calcular la exactitud y calidad de la aproximación.
- Calcular reducciones (*reducts*) como un conjunto de número mínimo de atributos describiendo conceptos.
- Reducir datos (con preservación de información), removiendo atributos innecesarios.
- Calcular el núcleo (*core*) y determinar un subconjunto de los atributos más significativos.
- Determinar un conjunto mínimo de atributos de reducción.
- Raciocinar con presencia de incertidumbre.
- Derivar algoritmos de decisión como un conjunto de reglas.

Los pasos seguidos en la estructura de los conjuntos aproximados son los siguientes:

Selección: el vehículo básico para la representación de datos en la estructura de la teoría de conjuntos aproximados es plano; tablas de datos en dos

dimensiones. Esto no implica que la tabla sea una simple tabla física; una tabla puede ser una vista lógica entre algunas tablas adyacentes. Las columnas de las tablas son llamadas atributos, las filas elementos, y las entradas en la tabla son los valores de los atributos.

Pre-procesamiento: si la tabla seleccionada contiene "huecos" en forma de valores perdidos o entradas de celdas vacías, la tabla puede ser preprocesada de varias formas para llenar o completar la tabla.

Transformación: los atributos numéricos pueden se pueden convertir en categorías, es decir, se utilizan intervalos o rangos en vez de los valores de los datos exactos.

Minería de datos: en la metodología de los conjuntos aproximados se producen conjunciones de proposiciones elementales o reglas si-entonces. Esto se realiza en un proceso de dos etapas, en las cuales subconjuntos de mínimos atributos son primero computados, antes de que los patrones o reglas sean generadas.

Interpretación y evaluación: Los patrones individuales o reglas pueden ser ordenados por alguna medida de "bondad" y manualmente inspeccionados. Conjuntos de reglas pueden ser empleados para clasificar nuevos casos y registrar el desempeño de clasificación.

Conceptos básicos de la teoría de los conjuntos aproximados

Sistemas de información

Un conjunto de datos puede ser representado por una tabla finita de elementos, donde cada fila representa un elemento, caso o evento. Cada columna representa un atributo, observación, o propiedad que puede ser medida de cada elemento, o suministrada por un usuario. Cada entrada en la columna q y fila x tiene el valor $f(x,q)$. Cada fila en la tabla describe la información acerca de algún

elemento en *S*.

Un sistema de información es una representación de datos recopilados de mediciones de algún fenómeno físico, como diálogos, textos, secuencias de imágenes, señales de procesos industriales, etc.

El sistema de información está compuesto por una *n*-tupla:

$S = \langle U, Q, V, f \rangle$ donde,

U - es el universo, conjunto no vacío finito de *N* elementos $\{x_1, x_2, ..., x_N\}$

Q – es un conjunto no vacío finito de *n* atributos $\{q_1, q_2, ..., q_n\}$

$V = \cup_{q \in Q} V_q$ donde V_q es un dominio (valor) del atributo *q*,

$f : U \times Q \rightarrow V$ – es la *función de decisión* llamada *función de información* tal que $f(x, q) \in V_q$ para cada $q \in Q$, $X \in U$. El termino universo se refiere a un universo cerrado (*Pawlak*, 1991).

Un *descriptor* del sistema de información *S,* es un par *(q,v)* para $q \in Q$, $v \in V_q$. Cualquier conjunto no vacío de elementos *X* es llamado un concepto en *S*. Un concepto puede tener un cierto significado; por ejemplo, en una serie de datos de tuberías con fallos y diagnósticos, se puede definir un concepto como un conjunto de elementos representando tuberías enfermas o con posibilidad de fallo.

Relación de no discernido (Indiscernibility)

Siendo, $S = \langle U, Q, V, f \rangle$ un sistema de información, $A \subseteq Q$ un subconjunto de atributos, y ($x,y \in U$) elementos, entonces los elementos "*x*" e "*y*" no son discernibles por el conjunto de atributos *A* en *S* (indicado por $x\tilde{A}y$) si:

$f(x,\alpha) = f(y,\alpha)$, para cada $\alpha \in$ A.

Cada subconjunto de atributos *A* determina una relación de equivalencia del universo *U,* la cual es referida a la relación de no discernido. Para cualquier subconjunto dado de atributos $A \subseteq Q$, *IND(A)* (mostrada por $\tilde{A}$) es una relación

de equivalencia en el universo *U,* y es expresada como una relación de no discernido. La relación de no discernido *IND(A)* (teniendo un conjunto de pares *(x, y)* de elementos de *U*) esta definida como:

$$IND(A) = \{(x, y) \in U \times U : \text{para todo } a \in A,\ f(x,a) = f(y,a)\}$$

Si el par de elementos *(x, y)* pertenecen a la relación *IND(A) ((x, y) ∈ IND(A)),* entonces los elementos *x* e *y* son nombrados como no discernibles con respecto a *A*. En otras palabras, no se pueden distinguir *x* de *y* en términos de atributos de únicamente el conjunto *A*.

La relación de no discernido *IND(A)*, como una relación binaria de equivalencia, divide el universo dado, *U,* en una familia de clases de equivalencia $\{X_1, X_2, ..., X_r\}$. La familia de todas las clases de equivalencia $\{X_1, X_2, ..., X_r\}$ definida por la relación *IND(A)* en *U*, genera una partición de *U* y es denotada por *A**. La familia de las clases de equivalencia *A** es también referida como una *clasificación* y se expresa como *U/IND(A)*.

Los elementos que pertenecen a las mismas clases de equivalencia X_i son no discernibles; en caso contrario, los elementos son discernibles con respecto al subconjunto de atributos *A*. Las clases de equivalencia X_i, $i = 1, 2, ..., r$ de la relación *IND(A),* son llamados *conjuntos A-elementales* en un sistema de información *S*.

$$[x]_A = \{y \in U : xIND(A)y \text{ ó } x \tilde{A} y\},$$

expresa un conjunto *A-elemental* (una clase de equivalencia) incluyendo un elemento *x*. Los conjuntos elementales X_i son bloques de la partición de *U* y representan los grupos más pequeños discernibles de elementos y se nombran como *conocimiento A-básico*.

De un sistema de información *S,* un subconjunto dado de atributos $A \subseteq Q$ genera la relación de no discernido *IND(A)* (relación de equivalencia). Un par ordenado *AS = (U,IND(A))* es referido como un *espacio de aproximación*. Cualquier unión finita de conjuntos elementales en AS, es nombrada como un

conjunto definible o *conjunto creado* en *AS*. Los conjuntos *Q-elementales* son llamados *átomos* de un sistema de información *S*. $Des_A(X)$ indica la descripción del conjunto *A*-elemental $X \in A^*$ (clase de equivalencia) y es definido como:

$$Des_A(X) = \{(\alpha, b) : f(x, a) = b, \forall x \in X, \alpha \in A\} \text{ ó,}$$

$$Des_A(X) = \{(\alpha = b) : f(x, a) = b, \forall x \in X, \alpha \in A\}$$

La relación de no discernido *IND(A)* impuesta por *A,* divide el universo en clases de equivalencia cuya familia (conjunto) tiene una clasificación *U/IND(A)* (partición *A** de *U*). Cualquier conjunto A-elemental X $\in$ X* (una clase de equivalencia), puede ser descrito por $Des_A(X)$.

Cada conjunto *A-elemental* X_i (conocimiento *A-básico*) representa una *célula* (el grupo de elementos más pequeño discernible), en el espacio de aproximación *AS = (U, IND(A))*. Cualquier unión finita de conjuntos *A-elementales* es un conjunto definible en *AS*.

Aproximación de conjuntos, regiones límite, espacios de aproximación

Siendo $S = \langle U, Q, V, f \rangle$, un sistema de información y tomando $A \subseteq Q$, y $X \subseteq U$ (*concepto de X*), se puede aproximar el *conjunto X* utilizando únicamente la información de *A* construyendo las llamadas aproximaciones ***A-inferior*** y ***A-superior*** de *X*, indicadas por $\underline{A}\,X$ y $\overline{A}\,X$ respectivamente, donde:

$$\underline{A}\,X = \{x : [x]_B \subseteq X\} \text{ y } \overline{A}\,X = \{x : [x]_B \cap X \neq 0\}.$$

La *aproximación inferior* corresponde a reglas ciertas, mientras que la *aproximación superior* a reglas posibles (con confianzas superiores a cero). La aproximación *A-inferior* de *X* es el conjunto de todos los elementos que pueden ser *verdaderamente* clasificados en *X* utilizando atributos de *A*. La aproximación inferior $\underline{A}\,X$ del conjunto *X* es la unión de todos aquellos conjuntos elementales, cada uno de los cuales esta contenido por *X*. Para cualquier $x \in \underline{A}\,X$ *es cierto que* $x \in X$, es decir, la aproximación inferior $\underline{A}\,X$ del conjunto *X* contiene todos los

elementos que basados en el conocimiento de los atributos *A*, pueden ser clasificados como ciertamente pertenecientes al concepto *X*.

La aproximación superior $\overline{A}X$ del conjunto *X* corresponde a la unión de aquellos conjuntos elementales, cada uno de los cuales tiene una intersección no vacía con *X*. Para cualquier $x \in \overline{A}X$ solo se puede decir que *x* posiblemente pertenece a *X*, es decir, la aproximación superior $\overline{A}X$ del conjunto *X* contiene todos aquellos elementos que basados en el conocimiento de los atributos de *A* pueden no ser clasificados como pertenecientes al concepto *X*.

El conjunto *U* - $\overline{A}X$ se conoce como la región *A-exterior* de *X,* y está formada por aquellos elementos que pueden ser clasificados como no pertenecientes a *X* utilizando atributos de *A*. El conjunto $BN_A(X) = \overline{A}X - \underline{A}X$ es conocido como la región *A-limite* de *X* (región incierta de *IND(A))*, y está formado por aquellos elementos que sobre la base de los elementos de *A* pueden ser claramente clasificados en *X*.

Matriz de discernido

Los conceptos de *matriz de discernido* y *función de discernido* sirven para construir algoritmos eficientes relacionados con, la generación de subconjuntos mínimos de suficientes atributos, para describir todos los conceptos en un sistema de información dado. Con estas dos nociones se pueden almacenar las diferencias entre atributos de cada par de elementos en una matriz llamada *matriz de discernido*.

Siendo $S = \langle U, Q, V, f \rangle$ un sistema de información, y asumiendo que *Q* = $\{\alpha_1, \alpha_2, ..., \alpha_n\}$, *U* = $\{x_1, ..., x_n\}$. Una matriz de discernido *M(Q),* para un sistema de información *S* con el conjunto de atributos *Q,* es una matriz de dimensiones *N* x *N*, con filas y columnas etiquetadas por los elementos x_i ($i = 1, 2, ..., N$). Cada entrada $m_{i,j}$ de la matriz de discernido (para una fila *i* y columna *j* dada, representando dos elementos x_i y x_j de *U*), es un subconjunto de atributos que

diferencia aquellos elementos.

La matriz de discernido puede ser definida como:

$$m_{ij} = \begin{cases} 0 & x_i\ , x_j \in iguales\ clases\ de\ equivalencia\ de\ IND(Q) \\ \{a \in Q : f(x_i, a) \neq f(x_j, a)\}, & x_i\ , x_j \in diferentes\ clases\ de\ equivalencia\ de\ IND(Q) \end{cases}$$

donde x_i , $x_j \in U$. La entrada $m_{i,j,}$ contiene todos aquellos atributos cuyos valores no son iguales para x_i y x_j, o sea que pertenecen a diferentes clases de la partición generada por *IND(Q)*. La matriz *M(Q)* es simétrica y $w_{ii} = 0$.

Cuando se ha creado la matriz de discernido se puede definir la *función de discernido*. Esto es, un signo concreto de como cada elemento en los datos se puede discernir de los otros. La función de discernido *fs,* de un sistema de información *S*, es una función booleana de *n* variables $\bar{a}_1, \bar{a}_2, ..., \bar{a}_n$, que corresponde a los atributos $a_1, a_2, ..., a_n$ respectivamente, y es definida como:

$$f_A(\bar{a}_1, \bar{a}_2, ..., \bar{a}_n) = \wedge\{\vee(m_{ij}) : 1 \leq j < i \leq m,\ m_{ij} \neq 0\}$$

donde $\vee(m_{ij})$ es la separación de todas las variables $\bar{a}$, tal que $a \in m_{ij}$.

Tablas de decisión

Un sistema de información puede ser presentado como *tablas de decisión* si el conjunto de atributos *Q* es divido en dos conjuntos; el conjunto de atributos de condición *C,* y el conjunto de atributos de decisión *D,* es decir, $C \cup D = Q$ y $C \cap D \neq 0$. Las *tablas de decisión* pueden ser expresadas como *(U, C ∪ D)*. Se dice que una tabla de decisión es *determinística,* si cada valor de los elementos de los atributos de decisión está únicamente especificado por un elemento particular de los atributos de condición, y es *no-determinística,* si un número de valores de los atributos de decisión (acciones) puede ser tomado de un atributo de condición dado.

Precisión de aproximación

La *precisión* de una aproximación del conjunto *X* por el conjunto de atributos *A (precisión* de *X),* está definida cómo:

$$\alpha_A(X) = \frac{card\ \underline{A}X}{card\ \overline{A}X}$$

Si el conjunto *X* es exactamente aproximado a *A* en el espacio de aproximación *AS* definido por $A \in Q$, entonces $\alpha_A(X) = 1$. Si el conjunto *X* es aproximado a *A* en *AS*, entonces $0<\alpha_A(X)<1$.

La *aproximación* como opuesto a *precisión*, representa el grado de aproximación inexacta de un conjunto *X* en el espacio de aproximación *AS* = (*U,IND(A)*), definido por $A \in Q$.

$$\rho_A(X) = 1 - \alpha_A(X)$$

Clasificación y reducción

Una de las tareas más frecuente de muchas aplicaciones de minería de datos es la *clasificación de elementos*; que es el proceso de determinar una única clase para un elemento dado, por ejemplo, para un conjunto de elementos dados caracterizados por atributos de condición y decisión, se puede establecer una clasificación en familias diferentes de acuerdo a los valores de los atributos de decisión. Comúnmente unos pocos atributos importantes son suficientes para clasificar unos elementos.

Algunos atributos en un sistema de información pueden ser redundantes, y pueden ser eliminados sin perder información esencial clasificatoria. El proceso de encontrar un conjunto de atributos de menor tamaño del original con el mismo poder clasificatorio del conjunto original, es conocido como *reducción de atributos*. Como resultado el gran sistema de información original puede ser reducido a un sistema pequeño conteniendo los atributos.

Los conjuntos aproximados *(rough sets)* permiten determinar para un sistema de información dado los atributos más importantes desde un punto de vista clasificatorio. La *reducción (reduct)* es la parte esencial de un sistema de información (relacionado a un subconjunto de atributos), el cual, puede discernir todos los elementos perceptibles por el sistema de información original. Un *núcleo (core)* es una parte común de toda reducción.

Algunos atributos pueden depender unos de otros; cambios en un atributo dado pueden causar cambios en otros de una forma no lineal. Los conjuntos aproximados determinan el grado de dependencia de los atributos y su significado. En la relación de no discernido, la dependencia de atributos es una de las características más importantes de un sistema de información.

Dado un sistema de información $S = \langle U, Q, V, f \rangle$, con atributos de condición y decisión $Q = C \cup D$, de un conjunto de atributos de condición dado $A \subset C$, se puede definir la *región positiva A, $POS_A(D)$,* en la relación *IND(D),* como:

$$POS_A(D) = \bigcup \{ \underline{A}X \mid X \in IND(D) \}$$

La región positiva $POS_A(D)$ contiene todos los elementos en *U* que pueden ser clasificados perfectamente sin error, dentro de diferentes clases definidas por *IND(D)*, basado solo en información de la relación *IND(A)*. La definición de la región positiva puede ser creada de dos subconjuntos cualesquiera de atributos *A*, $B \in Q$ en el sistema de información *S.* La región positiva *A* de *B* se define como:

$$POS_A(B) = \bigcup_{x \in B^*} \underline{A}X$$

La región positiva *A* de *B* contiene todos los elementos que utilizando atributos de *A,* pueden estar ciertamente clasificados en una de las diferentes clases de la clasificación *B** (partición de *B)*.

Los conjuntos aproximados definen el grado de dependencia de conjuntos de atributos. La cardinalidad (tamaño) de la región positiva *A* de *B* es utilizada para definir una medida ((grado) $\gamma_A(B)$) de dependencia del conjunto de atributos

B en *A*:

$$\gamma_A(B) = \frac{card(POS_A(B))}{card(U)}$$

Es decir el conjunto de atributos *B* depende del conjunto de atributos *A* en un grado $\gamma_A(B)$.

Reducción (reduct) y núcleo (core)

Los sistemas de información a menudo incluyen atributos de condición que no suministran información adicional acerca de los objetos en *U*, eliminando estos atributos se puede reducir la complejidad y el coste de los procesos de decisión. Para un sistema de información dado, algunos atributos pueden ser redundantes o innecesarios con respecto a una clasificación específica *A** generada por el conjunto de atributos $A \subseteq Q$. Esto significa, que un sistema de información puede ser sobrecargado por esta información redundante.

En virtud de la propiedad de dependencia de los atributos, se puede definir un conjunto *reducido* de atributos removiendo los atributos *innecesarios*, sin perder el poder clasificatorio del sistema de información reducido. Esto conduce a la reducción sustancial del sistema de información, encontrando el conjunto óptimo de suficientes atributos de clasificación, robusto, y con un alto grado de generalización (Chan, 1998).

Para un sistema de información *S* y un subconjunto de atributos $A \subseteq Q$, un atributo $\alpha \in A$ es conocido como *prescindible* (*descartable*) en el conjunto *A* si *IND(A)* = *IND(A* − $\{\alpha\}$), lo cual significa que las relaciones de no discernido generada por los conjuntos *A* y $A - \{\alpha\}$ son idénticas, de otra forma un parámetro α es *indispensable* en *A*.

El conjunto de todos los atributos indispensables en el conjunto $A \subseteq Q$, es llamado el *núcleo* de *A* en *S,* y es representado como *CORE(A)*. El núcleo contiene

todos los atributos que no pueden ser removidos del conjunto *A* sin cambiar (perder) la clasificación original *A**. El núcleo de *A* puede ser un conjunto vacío (Bautista *et al.*, 1999.)

Reglas de decisión

Una de las aplicaciones de los conjuntos aproximados, es la generación de reglas de decisión de un sistema de información dado, para clasificación de elementos conocidos o predicción de clases de nuevos elementos no vistos durante el diseño (*An et al., 1997, Nguyen et al., 2001, Zhong et al., 2001*). El proceso de reducción lleva a la inducción de mínimas reglas de decisión. Cualesquiera de tales reglas contiene un número mínimo de descriptores en su parte condicional, tal que su conjunción define el amplio subconjunto de una clase de decisión (Chan, 1991).

Las reglas de clasificación son argumentaciones que discriminan un concepto, es decir, un conjunto de elementos caracterizados por una cierta propiedad de otros conceptos (Zhan y Ziarko, 1995). Las reglas de clasificación generadas por un sistema de aprendizaje son generalmente evaluadas por dos criterios: su exactitud clasificatoria sobre los elementos no conocidos, y su complejidad.

Una tabla de decisión puede ser clasificada como *determinística* o consistente si el conjunto de atributos de condición *C,* discierne un conjunto de atributos de decisión *D*, *aproximadamente determinística* si *D* depende sobre *C* pero, *C* no discierne a *D,* y *completamente no determinística* si *C* no esta relacionado con *D.*

Las reglas de decisión son escritas, como ya se ha visto anteriormente, de la siguiente forma:

si (conjunto de condiciones), entonces, (conjunto de decisiones)

En el Capítulo III, se presenta una aplicación (An *et al.*, 1997) de los

conjuntos aproximados a la predicción de la demanda de agua potable.

II.3.3.6. Redes Neuronales Artificiales

El ser humano tiene la capacidad de resolver problemas a partir de la experiencia, sin poseer instrucciones o algoritmos predefinidos en su ser. Las redes neuronales artificiales corresponden a una forma de imitar la capacidad de los humanos de memorizar y asociar realizaciones, es decir, son modelos artificiales y simplificados del cerebro humano.

Las redes neuronales artificiales basan su funcionamiento en las redes neuronales reales, estando formadas por un conjunto de unidades de procesamiento conectadas entre sí. Por analogía con el cerebro humano se denomina "neurona" a cada una de estas unidades de procesamiento. Cada neurona recibe muchas señales de entrada, y envía una única señal de salida (como ocurre en las neuronas reales).

Las redes neuronales son conocidas en la estructura del aprendizaje automático cómo *"aproximadores universales",* con un gran carácter paralelo de cálculo, y buenas capacidades de generalización, pero también como cajas negras, debido a la dificultad para penetrar dentro de las relaciones aprendidas; por tanto es necesario un conocimiento experto acerca del dominio que se este desarrollando.

El aprendizaje de una red se puede realizar de tres formas diferentes, supervisadas, sin supervisión y auto supervisadas (Jain y Martín, 1999).

En las redes de ***aprendizaje con supervisión*** la red es provista de entradas y salidas esperadas (objetivos). La red de varias capas de retroceso o propagación hacia atrás (*multilayer back-propagation network* - *MLBPN*), estudiada en los 80 por *Rumelhart (1986) y Parker (1985)*, es el ejemplo más popular de una red supervisada. Es una poderosa técnica para construir funciones de trasferencia no lineales (Duch y Jankowski, 2000, y Nikov y Stoeva, 2001) entre algunos valores de entrada continuos y uno o más valores de salida

continuos. La red básicamente utiliza una arquitectura ***perceptron*** (una capa de entrada y una de salida Vaughn 1996, 1997, 1999 y 2001) y toma su nombre de la forma en la cual se procesan los errores durante el ejercicio.

En las redes con ***aprendizaje no supervisado*** no se conoce el valor objetivo; los pesos de la red son modificados de tal forma que se generen vectores de salida que posean propiedades estadísticas similares, dados vectores de entrada los cuales tienen también propiedades estadísticas similares; es decir, se presentan los patrones de entrada a la red y ésta los clasifica en categorías según sus rasgos más sobresalientes. En este tipo de redes el comportamiento colectivo es dependiente de cada componente (*Girolami et al.*, 1998).

La teoría de la *resonancia adaptativa* (*Adaptive Resonance Theory*) es un ejemplo de red sin supervisión u organizada por sí sola propuesta por *Carpenter y Grossberg (2003)*, la cual pretende resolver el dilema de la estabilidad – plasticidad, esto es, adaptarse (plasticidad) ante los datos relevantes, y mantenerse estable ante los datos irrelevantes.

Otro ejemplo es el algoritmo de *Kohonen* de *mapas auto-organizados (Self Organized Maps, ver anexo 1)*, cuya aplicación fundamental es la agrupación de datos; estas redes están compuestas por dos capas de unidades: una capa de entrada unidimensional, y una capa competitiva bidimensional organizada en una malla de dos direcciones, cuyas unidades pueden considerarse como ocultas y de salida simultáneamente. Cada unidad de esta capa tiene asociado un vector de referencia que, tras el entrenamiento, recuerda un determinado patrón de entrada. El algoritmo de aprendizaje agrupa los datos de entrada y, además, genera una aplicación espacial tal que los patrones de entrada similares, tienden a producir activaciones en unidades próximas de la malla. Para realizar el aprendizaje es necesario normalizar los vectores de entrada, así como los pesos de la capa competitiva.

En el ***aprendizaje auto supervisado*** la propia red corrige los errores en la interpretación empleando una realimentación.

Las redes neuronales artificiales supervisadas presentan buenos resultados en predicción, evaluación y generalización, mientras que las redes neuronales artificiales sin supervisión son más apropiadas en problemas de clasificación y reconocimiento.

El interés en las redes neuronales artificiales radica principalmente en que se asemejan a la forma como trabaja el cerebro humano, (Ali y Chen, 1999):

- Habilidad para aprender y generalizar;
- Adaptabilidad. La aptitud para adaptarse y continuar aprendiendo después de la fase inicial de entrenamiento. Las redes neuronales artificiales pueden modificar su comportamiento en respuesta a su ambiente;
- Contenido de direccionabilidad;
- Alto grado de tolerancia a fallos. La naturaleza de proceso distribuido de las redes neuronales artificiales tiene la ventaja de que, puesto que muchas neuronas están involucradas a la vez, sus entradas individuales no son mayormente significativas. Por consiguiente, de manera similar al proceso del cerebro, la pérdida o daño de una neurona o conexión es improbable que afecte significativamente el modelo de red neuronal en una forma adversa;
- Organización propia;
- Robustez;
- Simplicidad de cálculos básicos paralelos;
- Índices altos de cálculo dados por el paralelismo masivo;

Otros beneficios en la utilización de redes neuronales artificiales son:

- El programa de red neuronal artificial modifica sus parámetros internos como respuesta a su ambiente, y establece relaciones entre múltiples variables donde las reglas de decisión de otro modo no podrían ser

determinadas.

- La flexibilidad introducida a través de los pesos y funciones de trasferencia no lineales, significa que una red neuronal artificial de múltiples capas puede solucionar casi cualquier problema complejo no lineal, debido a que un número suficiente de procesamiento de elementos en capas escondidas es utilizado.
- La habilidad para ver a través de ruido y distorsión, de nuevo por su naturaleza de conocimiento distribuido o procesamiento.
- Alta compatibilidad con tecnología existente.

Algunas desventajas en el uso de las redes neuronales artificiales son las siguientes:

- Las redes neuronales no son muy útiles para problemas que requieran cálculos precisos.
- No hay actualmente ninguna metodología definida para el diseño de la estructura de red para una clase dada de problema que tiene que ser solucionado.
- No hay un buen sistema de auditoria que claramente delinee el camino de decisiones tomadas, como existe en los algoritmos de sistemas expertos.
- Las redes neuronales pueden requerir 2 o 3 órdenes de magnitud más que los datos y tiempo de cálculo en los métodos convencionales, para llegar a una configuración adecuada.

Las redes neuronales artificiales, son utilizadas para tratar problemas donde las características no numéricas necesitan ser tenidas en cuenta, pero excluyen el uso de funciones matemáticas; estas aplicaciones incluyen:

- Reconocimiento de patrones para la toma de decisiones inteligentes en grandes bases de datos comerciales tales como las manejadas por los bancos.

- Predicción en muchas áreas diferentes, tales como financiera y clima.
- Procesamiento de señales.
- Control de procesos industriales, en combinación con dispositivos apropiados de detección.
- Visión computacional y control robótico.
- Reconocimiento de caracteres de escritura.
- Reconocimiento de caracteres ópticos
- Conversión de textos en discursos altamente comprensibles.

Desarrollo histórico

En 1943, *Warren McCulloch* y *Walter Pitts* propusieron el clásico modelo de neurona en el que se basan las redes neuronales actuales. Seis años después, Donald Hebb en su libro *The Organization of Behavior*, presenta su conocida regla de aprendizaje.

En 1958, *Frank Rosenblatt* presentó el **Perceptron**, una red neuronal con aprendizaje supervisado cuya regla de aprendizaje era una modificación de la propuesta por *Hebb*. El **Perceptron** trabaja con patrones de entrada binarios, y su funcionamiento, por tratarse de una red supervisada, se realiza en dos fases: una primera en la que se presentan las entradas y la salidas deseadas; en esta fase la red aprende la salida que debe dar para cada entrada. La principal aportación del **Perceptron** es que la adaptación de los pesos se realiza teniendo en cuenta el error entre la salida que da la red y la salida que se desea. En la fase siguiente, de operación, la red "es capaz" de responder adecuadamente cuando se le vuelven a presentar los patrones de entrada.

En los años 60 se propusieron otros dos modelos, también supervisados, basados en el *Perceptron* de *Rosenblatt* denominados **Adaline** y **Madaline**. En estos, la adaptación de los pesos se realiza teniendo en cuenta el error, calculado como la diferencia entre la salida deseada y la dada por la red, al igual que en el

Perceptron. Sin embargo, la regla de aprendizaje empleada es distinta. Se define una función error para cada neurona que da cuenta del error cometido para cada valor posible de los pesos, cuando se presenta una entrada a la neurona. Así, la regla de aprendizaje hace que la variación de los pesos se produzca en la dirección y sentido contrario del vector gradiente del error. A esta regla de aprendizaje se la denomina *Delta*.

La era moderna de las redes neuronales artificiales surge con la técnica de aprendizaje de propagación hacia atrás o *Back Propagation*. La estructura de las redes citadas anteriormente (*Perceptron, Adaline y Madaline*) consta de dos capas: una capa primera formada por unidades que dejan pasar la entrada y que no tienen aprendizaje, y una segunda capa formada por una o varias neuronas en el caso del *Madaline*. Ya que las redes del tipo *Perceptron* no son capaces de aprender todas las posibles combinaciones entre entradas y salidas, la solución del problema consiste en añadir capas intermedias de neuronas, introduciendo de esta forma el problema de cómo enseñar a estas capas intermedias. Aquí es donde tiene importancia el algoritmo de propagación hacia atrás. En este se compara la salida real con la salida deseada. La diferencia entre ambas constituye un error que se propaga hacia atrás desde la capa de salida hasta la de entrada, permitiendo así la adaptación de los pesos de las neuronas intermedias mediante una regla de aprendizaje *Delta*. Sin embargo, también tiene sus limitaciones.

Posteriormente, se han desarrollado otros modelos que permiten un aprendizaje no supervisado como los mapas auto-organizados (*Self Organized Maps*) de *Kohonen*, los basados en la *Teoría de Resonancia Adaptativa* (*ART*) de *Grossberg y Carpenter*, o los modelos de control motor de *Bullock*, *Gaudiano* y *Grossberg*, entre otros (Catalina, recuperado 20 de julio de 2009).

Una *red neuronal artificial o red neuronal (NN)* en forma breve, puede ser definida por la especificación de los siguientes componentes:

- Modelo(s) de neuronas utilizado,
- Topología,
- Regla de aprendizaje utilizada para actualizar los pesos.

Modelos de Neuronas

Todos los modelos de neuronas biológicas, llamadas neuronas artificiales o brevemente neurona, intentan imitar sus equivalentes biológicos lo más parecido posible.

McCulloch y Pitts (1943) propusieron un modelo de neurona artificial como el visualizado en la Figura II.12. Realiza dos operaciones, una suma de los productos de W_i y X_i, seguida de una función de activación de salida de umbral que funciona sobre esta suma. Cómo en una neurona biológica, la neurona recibe inputs X_i de otras neuronas, y compara la suma con el umbral ϑ_i, para producir un resultado *y,* el cual es enviado por medio de la función de activación (*fact*) a otras neuronas solo si está por encima del umbral.

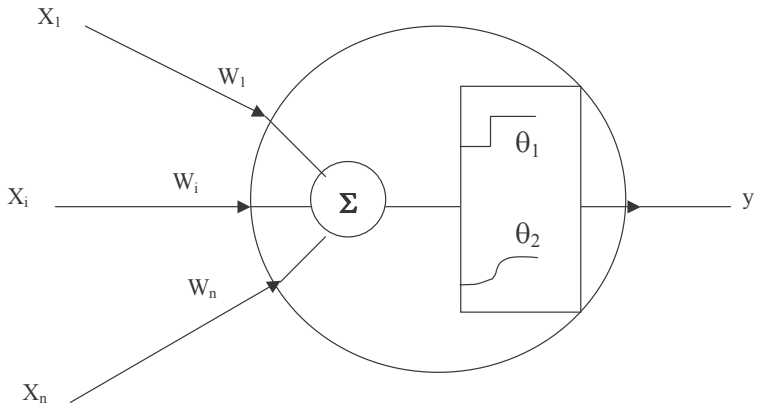


Figura II.12. Modelo de neurona artificial.

Este proceso se puede describir por la siguiente ecuación:

$$y = fact\left(\sum_{i=1}^{n} W_i\ X_i - \theta_1\right),$$

donde *n* es el número de *inputs* o la dimensión del vector X de entrada, y W_i los pesos de las conexiones que representan la sinapsis.

La función continua sigmoidea mostrada en la parte de abajo del rectángulo, es una de las funciones de activación no lineal más popular y está definida como:

$$fact \quad (X) = \left(1 + \exp\left(-\frac{X}{\theta_2}\right)\right)^{-1},$$

donde *x* es la entrada neta a la sigmoidea, y θ_2 representa la pendiente de la curva sigmoidea.

Topologías de las redes neuronales

En muchos casos, la topología de las redes neuronales necesita ser especificada por el usuario, con excepción de las redes ontogénicas (generan su propia topología), las cuales hacen uso de algunos criterios para guiar su propio diseño. La topología de las redes neuronales puede ser vista como un grafo dirigido donde las neuronas son los nodos y los pesos de las conexiones entre neuronas son los arcos. Todas las topologías de redes neuronales pueden ser divididas en dos tipos generales:

Sin ciclos *(feedforward)*, sin lazos ni conexiones dentro de la misma capa, también conocida como con propagación hacia adelante o progresiva (Ampazis *et al.*, 2001, Menéndez *et al.*, 1994).

Cíclica *(recurrent)*, con bucles de retroalimentación (Romero *et al.*, 1999) o con retropropagación o propagación regresiva.

Un ejemplo de una red neuronal sin ciclos es la *Función de Base Radial (Radial Basis Function)*, y una red recurrente es la red neuronal de *Hopfield*. En la Figura II.13 se visualizan las topologías de una red neuronal sin ciclos y una cíclica. En la red neuronal sin ciclos, los inputs solo distribuyen datos a las capas o unidades ocultas o escondidas (*hidden layer*) sin ejecutar cálculos.

Las limitaciones que presentan las redes neuronales convencionales con una única representación de los datos de entrada y procesos convencionales de aprendizaje, corresponden a la imposibilidad de representar cada faceta intrínseca de la información contenida en los datos de entrenamiento, y la tarea de aprendizaje se acomete como un todo, lo cual hace que la calidad y desempeño

sobre las características de generalización y robustez sean pobres y complejas. El arquetipo de red neuronal convencional emplea una representación de señal única para todo el proceso de aprendizaje.

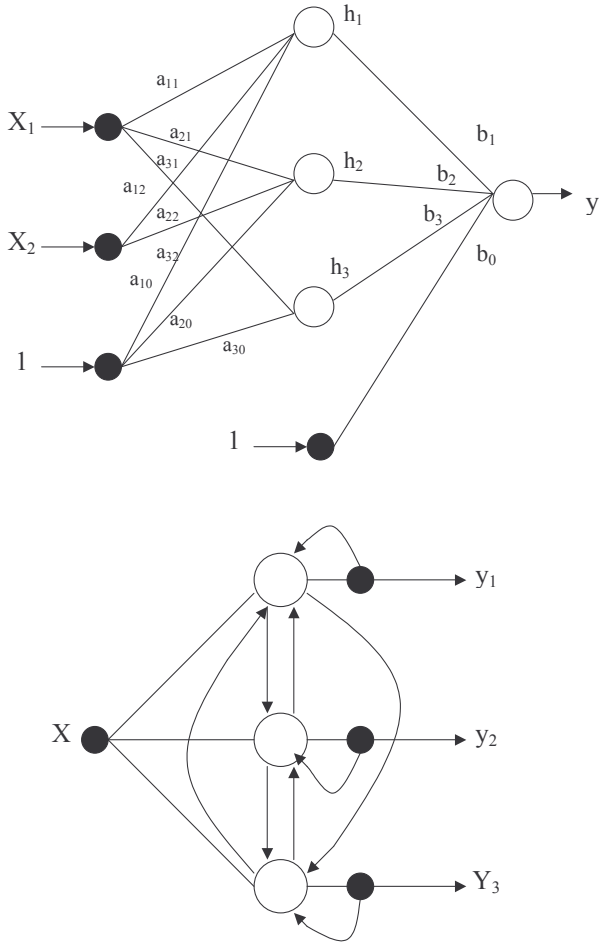


Figura II.13. Topología de redes neuronales sin ciclos y cíclica.

La generalización es un requerimiento clave para el trabajo con redes neuronales, y se refiere a la habilidad de una red para producir respuestas apropiadas a patrones que no fueron incluidos en la serie de datos de entrenamiento empleada. Para mejorar la habilidad de generalización de las redes para la predicción, *Liang y Liang* (2001) han empleado un arquetipo de aprendizaje llamado aprendizaje de resolución múltiple *(multiresolution learning, NNMLP)*, basado en el análisis de resolución múltiple de la teoría de pequeñas ondas (Sheikholeslami *et al.*, 2000), el cual es empleado para descomponer la señal original y aproximarla a diferentes niveles de detalle. El arquetipo de aprendizaje por resolución múltiple explota la secuencia de aproximación-

representación, por representación de la resolución gruesa a la fina durante el proceso de entrenamiento de la red; en consecuencia se pueden explotar las estructuras de correlación en los datos de entrenamiento para múltiples resoluciones, lo cual de otra forma podría obscurecer la resolución original de los datos de entrenamiento.

En las redes con varias capas, tanto el número de capas como el de neuronas son parte esencial del diseño. El número de capas está íntimamente relacionado con la velocidad de aprendizaje, por lo que comúnmente el número de capas ocultas se reduce a una o dos. Del número de neuronas en las capas ocultas depende el número de parámetros libres en el modelo. Si son pocas, es imposible lograr el ajuste; si son excesivas, se pierde generalidad.

Reglas de Aprendizaje

Algunas de las reglas utilizadas para actualizar los pesos en la generación de salidas *(outputs)* de las redes neuronales artificiales, son las siguientes:

Aprendizaje Hebbiano y anti-Hebbiano

Hebb (1949) estableció que si una neurona *i* antes de la sinapsis, repetidamente apunta hacia una neurona *j* post-sináptica, la fuerza de sinapsis entre las dos se incrementa. Esto se puede representar mediante la siguiente formulación:

$$w_{ij}(t+1) = w_{ij}(t) + \eta\, y_j(t) x_i(t),$$

donde $\eta \in [0,1]$ es el índice de aprendizaje, y t el tiempo de la iteración, w_{ij} indica el valor anterior de la conexión. Una variación a esta fórmula está basada en la correlación inversa entre las dos neuronas, a menudo conocida como la regla anti-Hebbiana de aprendizaje:

$$w_{ij}(t+1) = w_{ij}(t) - \eta\, y_j(t) x_i(t).$$

Aprendizaje Competitivo

En esta regla biológica de aprendizaje, las neuronas compiten entre ellas por la activación, y sólo la neurona ganadora, w_i, es capaz de encender y a la misma vez reforzar su peso:

$$w_i(t+1) = w_i(t) + \eta(t)\,(x - w_i(t))\,.$$

Aprendizaje de la corrección del error simple

Esta regla de aprendizaje fue utilizada en el *Perceptron* (*Rosenblatt*, 1958), donde se mide el error de la diferencia entre la salida *(y),* y la salida deseada conocida *(d)*, para actualizar los pesos.

$$w_i(t+1) = w_i(t) + \eta\,(d - y)\,x_i\,,$$

donde *(y),* y *(d),* $\in$ (0,1). Esta regla solo aplica a redes neuronales de una capa sin ciclos.

Aprendizaje de la corrección del error por retropropagación (Backpropagation)

Los pesos asociados a los elementos del proceso de las capas ocultas *(hidden layers)* son ajustados asumiendo que los errores en la salida son atribuidos a todos los elementos del proceso. La retropropagación entrena las capas escondidas, propagando el error en la capa de salida hacia atrás a través de la red, capa por capa, ajustando los pesos en cada capa.

Las redes neuronales corresponden a paradigmas de "cajas negras", las cuales encuentran una función de asociación, entre las entradas *(x),* y las salidas *(y),* donde la tarea consiste en encontrar los pesos adecuados *w* para aproximar esta función:

$$y = f(x, w)\,.$$

La regla de retropropagación (*Werbos*, 1974), es utilizada para determinar y actualizar los pesos:

$$w(t+1) = w(t) + \eta \nabla f\,(x(t+1), w(t))\,(d(t+1) - f(x(t+1), w(t)))\,,$$

donde ∇f es el gradiente de f con respecto a los pesos w. Los pesos son actualizados de una forma similar a la regla de corrección del error,

$$(d(t+1) - f\,(x(t+1),\, w(t))),$$

influenciado por el cambio de peso y el valor del índice de aprendizaje; ∇f, debe tender a cero para que la retroalimentación encuentre una solución.

Aprendizaje de Boltzmann

Este es un tipo de aprendizaje no determinístico de naturaleza estocástica y muestra la diferencia entre dos salidas de neuronas en los modos supervisado y no supervisado.

$$\Delta w_{ij} = \eta\,(\,pc_{ij} - pf_{ij}\,)\,,$$

donde pc_{ij} y pf_{ij} son las probabilidades de que en las dos neuronas i y j estén *"encendidos"* los modos supervisado y no supervisado respectivamente, cuando la red este en equilibrio. Esta regla de aprendizaje es generalmente conocida como máquina de Boltzmann (Hinton y Sejnowski, 1983).

Algoritmos de Aprendizaje

Cuando una regla de aprendizaje es utilizada dentro de un algoritmo, tal algoritmo se define como un algoritmo de aprendizaje. Los algoritmos de aprendizaje pueden ser utilizados en tres modos:

Aprendizaje supervisado

En el aprendizaje supervisado se suministran a la red entradas (*inputs*), y

salidas esperadas u objetivo (*outputs*). Las salidas de la red deberán ser añadidas a los valores objetivo; esto se realiza cambiando (actualizando) las conexiones de la red. Ejemplos de este tipo de aprendizaje son las redes neuronales sin ciclos, cuyo entrenamiento se realiza utilizando varias versiones del algoritmo de aprendizaje de retropropagación.

Aprendizaje sin supervisión

En este caso los valores de salida no están rotulados y, se espera que el algoritmo de aprendizaje presente la forma de darle sentido a los datos y agrupar patrones, colocando magnitudes similares en grupos iguales.

Combinación de las dos

Los algoritmos híbridos generalmente trabajan primero en un modo sin supervisión y luego intercambian al modo supervisado.

II.3.3.7. Árboles de decisión

Un árbol de decisión es una estructura en forma de árbol que visualmente describe una serie de reglas (condiciones), que conducen a la toma de una decisión.

Se tienen dos tipos de árboles de decisión (Berry y Gordon, 2000):

Árboles de Clasificación: los cuales rotulan registros y los asignan a la clase correspondiente. Los árboles de clasificación también pueden dar una confianza de que la clasificación sea correcta. En este caso, el árbol de clasificación da la probabilidad de clase, la cual es la probabilidad de que el registro este dentro de una clase dada.

Árboles de Regresión: estiman el valor de una variable objetivo que toma valores numéricos.

Todos los árboles tienen la misma estructura. Cuándo un modelo de árbol es aplicado a los datos, cada registro avanza a través del árbol por una vía determinada por una serie de controles tales cómo *"es el campo 3 mayor que el 27"* o *"es el campo 4 rojo, verde o azul"*, hasta que el registro alcance una hoja o nodo terminal del árbol.

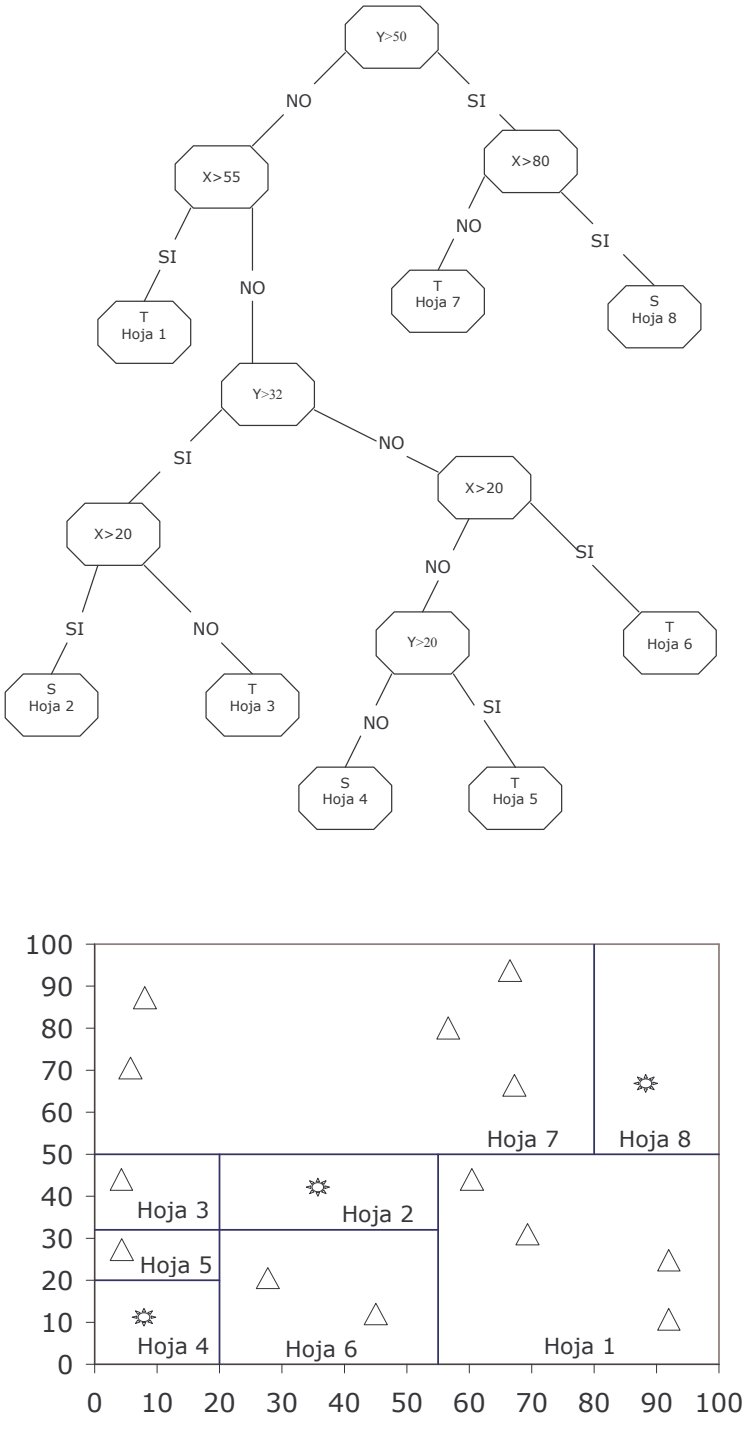


Figura II.14. Un árbol de decisión corta el espacio en cajas.

Cada rama de un árbol de decisión es un control de una variable única que corta el espacio en dos o más partes.

En la Figura II.14 se visualiza un ejemplo de árbol de decisión, el cual ha crecido hasta que cada caja es completamente pura en el sentido de que solo contienen una especie de figura (triángulo o sol).

Los árboles de decisión son construidos a partir de un proceso conocido como partición recursiva, el cual es un proceso iterativo que consiste en dividir los datos en partes.

Inicialmente todos los datos están juntos en una gran caja, y el algoritmo entonces parte los datos utilizando cada división binaria posible en cada campo. El algoritmo busca la división que fraccione los datos en dos partes que son más puras que la original. Esta división o partición es aplicada a cada una de las nuevas cajas. El proceso continúa hasta no encontrar más divisiones útiles. Así, el núcleo del algoritmo es la regla que determina la división inicial, (Chauchat y Rakotomalala 2001).

Para encontrar la división inicial se empieza con un conjunto de entrenamiento consistente en registros preclasificados, esto significa que el campo objetivo o variable dependiente tiene una clase conocida. La meta es construir un árbol que distinga entre las clases. El árbol puede ser utilizado para asignar una clase al campo objetivo de nuevos registros basado en los valores de otros campos o variables independientes.

La primera tarea es decidir cual de los campos independientes realiza la mejor división. La mejor división está definida como la que hace el mejor trabajo de separar los registros en grupos donde predomina una clase única. La medida utilizada para evaluar un divisor potencial es la reducción en *diversidad* (lo cual es otra forma de decir *"el incremento de pureza")*. El índice de diversidad corresponde a la probabilidad de que la segunda cosa palpada pertenezca a una clase diferente de la primera.

Para buscar el mejor divisor en un nodo, el algoritmo del árbol de decisión considera cada campo de entrada a su vez. En esencia cada campo es clasificado, entonces cada posible división es ensayada. La medición de diversidad es calculada para las dos particiones, y la mejor división es la que tiene mayor disminución en *diversidad.* Esto se repite para todos los campos y el ganador se escoge como el divisor del nodo.

El índice de diversidad ha sido desarrollado en diferentes campos, y por consiguiente con diferentes nombres. Para los biólogos estadísticos es el índice de diversidad de *Simpson*, para los criptógrafos es uno menos el índice anterior, para los econometristas es el índice *Gini*.

Independientemente de su nombre, su objetivo es medir la diversidad de una población. Puede ser interpretado como la probabilidad de que cualesquiera de dos elementos de la población escogidos al azar, pertenezcan a clases diferentes. Ya que la probabilidad de cualquier clase escogida dos veces en una muestra es simplemente P_i^2, el índice de diversidad es simplemente uno menos la suma de todos los P_i^2. Cuando hay sólo dos clases, las cosas se hacen aún más simples ya que la probabilidad de una clase es uno menos la probabilidad de la otra:

$$
\begin{aligned}
&1-(P_1^{\,2}+P_2^{\,2})\\
&1-(P_1^{\,2}+(1-P_1)^2)\\
&1-(P_1^{\,2}+(1-P_1)\,(1-P_1))\\
&1+-1\,(P_1^{\,2}+(1-P_1)-P_1+P_1^{\,2})\\
&1+-P_1^{\,2}+-1+P_1+P_1+-P_1^{\,2}\\
&2P_1-2P_1^{\,2}\\
&2P_1(1-P_1)
\end{aligned}
$$

La división inicial produce dos nodos, cada uno de los cuales es dividido de igual forma que el nodo raíz. Si todas las salidas en el nodo son iguales, entonces no tiene sentido dividirlo. En este caso el nodo se rotula como un nodo hoja.

Por otra parte, el algoritmo del árbol de decisión examina todos los campos

de entrada para encontrar divisores candidatos. Si el campo toma un único valor, es eliminado como candidato ya que no puede ser dividido. El mejor divisor para cada campo restante es determinado. Cuando no se encuentra divisores que disminuyan significativamente la diversidad de un nodo entonces se deja como un nodo hoja.

Eventualmente solo los nodos hoja permanecen y el árbol de decisión total ha ido creciendo; sin embargo el árbol completo no es generalmente el que realiza el mejor trabajo de clasificar un nuevo conjunto de registros.

Podar *(Pruning)* es el proceso de remover hojas y ramas para mejorar el desempeño del árbol de decisión. Un árbol podado es un subconjunto del árbol total de decisión. El árbol de decisión sigue creciendo mientras nuevas divisiones pueden ser encontradas, lo que mejora la capacidad del árbol de separar los registros de entrenamiento en clases. Si los datos de entrenamiento son utilizados para la evaluación, cualquier poda del árbol solo incrementa el error (Rastogi y Shim, 2000).

Los algoritmos de construcción de árboles realizan su primera división en el nodo raíz, donde hay una gran población de registros. Cada división subsiguiente tiene menor y más pequeña población representativa con la cual trabajar. Las idiosincrasias de los registros de entrenamiento en un nodo particular, muestran patrones que son peculiares solo para aquellos registros. Estos patrones carecen de significado y son absurdos para predicción. Por ejemplo, decir que un árbol de decisión intenta predecir la demanda de agua en Valencia y tiene como entrada en un nodo las demandas en México y algunas otras ciudades de centro América, se puede disminuir la diversidad en el nudo diciendo que "ciudades con características similares a México tienen esa demanda"; esta regla ayuda a clasificar los datos de entrenamiento, pero no son adecuados para determinar la demanda en Valencia.

Cuando se producen varias divisiones y se llega a niveles muy profundos dentro del árbol de decisión, se pueden presentar problemas de sobre valoración del conjunto de datos, es decir, en este nivel probablemente se hace una

distinción que resulta ser verdadera en el conjunto de entrenamiento, pero parece improbable para sostener en general. Se tienen algunas aproximaciones para tratar este problema, entre ellas las técnicas de podado o las técnicas de *bonsái*. Las técnicas de *bonsái* intentan detener el crecimiento del árbol antes de que esté demasiado profundo. El trabajo se realiza aplicando controles en cada nudo para verificar si una división es útil. Este control puede ser simple requiriendo de un número mínimo de registros que deben estar en el nodo, o más complicado aplicando pruebas estadísticas para verificar la importancia de la división propuesta. Las técnicas de poda dejan crecer el árbol para luego encontrar vías para podar las ramas que fallan al generalizar.

Los métodos de árboles de decisión son una buena elección cuando la tarea de minería de datos consiste en clasificar registros, o predecir resultados. Son utilizados cuando el objetivo es asignar a cada registro una de unas pocas categorías. Los árboles de decisión también son una buena alternativa cuando se trata de generar reglas que puedan ser fácilmente comprensibles, explicadas y trasladadas a un lenguaje natural (Apte y Weiss, 1997, Fukuda *et al.*, 1996).

Inducción de Árboles de Decisión

Existen una serie de algoritmos desarrollados desde principio de los años 60 del siglo anterior para la construcción de árboles de decisión, *CLS* (*Hunt et al.*, 1966), *CHAID* (Hartigan, 1975), *ID3* (Quinlan, 1979), *CART* (*Breiman et al.*, 1984), *ACLS* (Paterson y Niblett, 1982) *ASSITANT* (*Cestnik et al.*, 1987), *C4.5* (Quinlan, 1993) y *C5.0* (Quinlan, 1997 versión comercial), etc. A continuación de detallan dos de los algoritmos más conocidos.

Algoritmo ID3

Este algoritmo fue diseñado para el caso de la existencia de demasiados atributos y un conjunto de entrenamiento con bastantes elementos, donde se obtiene un árbol de decisión sin demasiado coste computacional. La estructura

básica de ID3 es iterativa; un subconjunto del conjunto de entrenamiento llamado *ventana,* se selecciona aleatoriamente y se forma un árbol de decisión a partir de este conjunto; este árbol clasifica correctamente todos los elementos en la *ventana*; el resto de elementos del conjunto de entrenamiento son entonces clasificados haciendo uso del árbol. Si el árbol da la respuesta correcta para todos los elementos, entonces es válido para todo el conjunto de entrenamiento y el proceso termina, si esto no ocurre se realiza una selección de elementos incorrectamente clasificados y se añaden a la *ventana* y el proceso continúa; en resumen siguiendo la premisa *"divide y vencerás"*, los pasos a seguir son los siguientes:

1. Se crea un nodo raíz con *S*: = todos los ejemplos
2. Si todos los elementos de *S* son de la misma clase, el subárbol se cierra. Solución encontrada.
3. Se elige una condición de partición siguiendo un criterio (split criterion).
4. El problema (y *S*) queda subdividido en dos árboles (los que cumplen la condición y los que no), y se vuelve al paso dos para cada uno de los subárboles.

Para encontrar la variable más informativa, el criterio seleccionado está basado en la ganancia de información (information gain); la información dada por un objeto o elemento depende de su probabilidad y puede ser medida en bits, como menos el logaritmo en base 2 de esta probabilidad. Siendo *C* un conjunto de objetos que contiene *p* objetos de clase *P,* y *n* objetos de clase *N*:

$$gain(A) = I(p,n) - E(A) \qquad (1)$$

$$I(p,n) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n} \qquad (2)$$

$$E(A) = \sum_{i=1}^{V}\frac{p_i + n_i}{p+n} I(p_i, n_i) \qquad (3)$$

Se asume que:

1. Cualquier árbol de decisión correcto para *C*, clasificará elementos en la misma proporción como su representación en *C*. Cualquier objeto arbitrario pertenecerá a la clase *P*, con una probabilidad *p/(p+n)* y a la clase *N*, con una probabilidad *n/(p+n)*.

2. Cuando un árbol de decisión es utilizado para clasificar un objeto, el resultado es una clase. El árbol de decisión puede ser utilizado como una fuente de un objeto "P" o "N", con la información necesaria esperada para generar este objeto dada por la ecuación (2).

Si el atributo *A* con valores *{A1,A2,…..AV},* es utilizado por la raíz del árbol de decisión, partirá *C* en *{C1,C2,…,CV}* donde C_i, contiene aquellos objetos en *C* que tienen valores A_i de A. Si C_i contiene p_i objetos de Clase *P,* y n_i de clase *N*, la información esperada requerida por el subárbol de C_i, es $I(p_i,n_i)$. La información requerida por el árbol con *A* como raíz, es obtenida por el promedio ponderado (3), donde el peso para la *i-ésima* rama, es la proporción de los objetos en *C* que pertenecen a C_i. La información ganada en la rama *A,* es por consiguiente, la dada en (1).

Este criterio de selección de variables utilizado por el algoritmo ID3 no es justo, ya que favorece la elección de variables con mayor número de valores. El algoritmo adicionalmente efectúa una selección de variables previa denominada Prepoda (preprunning), consistente en efectuar un test de independencia entre cada variable predictora y la variable de clase, de tal forma que para la inducción del árbol de clasificación, tan solo se van a considerar aquellas variables predictoras para las que se rechaza el test de hipótesis de independencia.

Algoritmo C4.5

Este algoritmo fue propuesto por Quinlan (1993), como mejora del algoritmo ID3. El algoritmo C4.5 está basado en la utilización del criterio de relación de ganancia (gain ratio), para corregir la deficiencia del algoritmo ID3 de

favorecer variables con mayor número de valores.

$$split\,info\,(A) = -\sum_{i=1}^{V} \frac{p_i + n_i}{p+n} \log_2 \left(\frac{p_i + n_i}{p+n} \right) \qquad (4)$$

$$gain\,ratio = gain(A) / split\,info(A) \qquad (5)$$

La ecuación 4, representa el potencial de información generado por dividir *C* en *V* subconjuntos, mientras que la información ganada mide la información relevante para la clasificación que resulta de la misma división. La ecuación 5 expresa la proporción de información generada por la división *(split)* qué es útil, en otras palabras, que parece provechosa para la clasificación.

Además, el algoritmo incorpora una poda *(prunning)* del árbol de clasificación al ser inducido. Esta poda está basada en la aplicación de un test de hipótesis para determinar si una determinada rama debe ser expandida. Con esta poda se evitan los problemas de sobreajuste *(overfitting)* del modelo, causados por el ruido de los ejemplos que puedan hacer crecer artificialmente el árbol. Una hipótesis $h \in H$ sobreajusta los datos del conjunto de entrenamiento (ejemplos) si existe una hipótesis alternativa $h' \in H$ tal que:

- *h* tiene menor error que *h'* sobre los ejemplos, pero
- *h'* tiene menor error que *h* sobre el universo

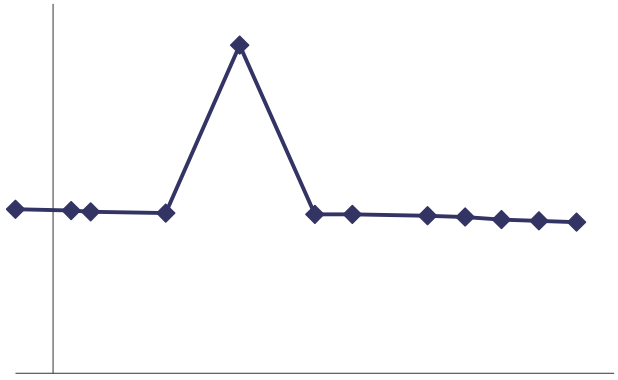


Figura II.15. Sobreajuste o Superajuste (Overfitting)

En otras palabras, el algoritmo de aprendizaje de árboles de decisión obtiene un modelo que es perfecto y robusto con respecto a la certeza, es decir,

se cubren todos los ejemplos vistos de forma correcta. Pero esto tiene el problema de que ajustarse demasiado a la certeza puede hacer que el modelo se comporte mal ante nuevos ejemplos, por el hecho de que en la mayoría de los casos el modelo es una aproximación objeto del aprendizaje, e intentar aproximar demasiado provoca un modelo poco general y muy específico. Además si los atributos contienen demasiado *ruido*, el modelo se ajustará a los errores, tal como se visualiza en la Figura II.15, y se perjudicará el comportamiento general del modelo aprendido.

Está poda puede ser temprana *(prepoda),* o tardía *(pospoda),* de acuerdo a si se realiza durante la construcción del árbol o al estar construido. Los criterios para la poda están basados en realizar un segundo conjunto de entrenamiento (validación), como la validación cruzada *(cross-validation),* o en pruebas estadísticas pera validar si la poda mejora la precisión sobre el universo.

Las principales ventajas de la técnica de árboles de decisión son las siguientes (Hernández-Orallo *et al.*, 2004):

- Se pueden aplicar a diversas tareas de minería de datos: clasificación, regresión, agrupamiento y estimación de probabilidades.
- Se pueden manejar tanto atributos numéricos (continuos) como nominales (discretos).
- Su eficiencia y escalabilidad para el manejo de grandes volúmenes de datos (tanto atributos como ejemplos).
- La facilidad para su utilización.
- Su tolerancia al ruido, a atributos no significativos y a datos faltantes.
- La facilidad en su interpretación y representación.

Entre las desventajas podemos encontrar:

- No son tan precisos como otros métodos (por ejemplo las redes

neuronales).

- Son débiles (*weak learners*) ya que dependen demasiado de los ejemplos, debido a su carácter voraz. Dos muestras distintas sobre la misma distribución pueden dar dos árboles bastante diferentes.

II.3.3.8. Métodos gráficos

Los modelos gráficos permiten realizar el proceso de KDD proporcionando flexibilidad en el sentido de que el conocimiento pueda ser codificado, representado y descubierto. Los modelos gráficos probabilísticos son una herramienta que permiten estructurar, representar y descomponer un problema utilizando la noción de independencia condicional. Algunos tipos de estos modelos son las *redes bayesianas* o *redes de probabilística causal*, *diagramas de influencia*, y *redes de Markov*, los cuales permiten acceder a la estructura del problema sin entrar en detalles matemáticos.

El modelo básico de un gráfico consiste en nudos que representan variables, y arcos que representan dependencias entre variables (o, no arcos, indicando independencias). Los nodos a los cuales apuntan los arcos representan variables aleatorias y decisiones. Los arcos apuntando dentro de las variables aleatorias indican dependencia probabilística, mientras que los arcos apuntando dentro de las decisiones, especifican la información disponible en el momento de la decisión. Por ejemplo, un nodo en una red puede representar el clima, mientras que otro puede representar la visibilidad. Un arco del clima a la visibilidad indica que esta (la visibilidad), es condicionalmente dependiente sobre el tiempo.

Redes Bayesianas

Una red *Bayesiana* es un modelo gráfico que utiliza arcos dirigidos para formar un gráfico sin ciclos, tal como se visualiza en la Figura II.16, el cual representa un modelo de competencia para un problema de patología de tuberías, en el que se presentan unas variables para explicar el problema, y los arcos

corresponden a nociones de causa o influencias.

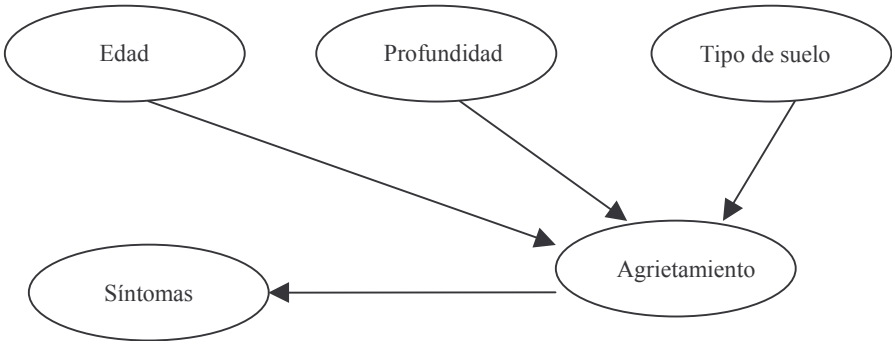


Figura II.16. Red Bayesiana

Por ejemplo, se puede pensar que los agrietamientos causan síntomas, y que la edad, profundidad y tipo de suelo causan agrietamientos. Estos gráficos también son llamados modelos causales. Otros modelos gráficos pueden representar las variables en diferente orden, dependiendo de si la gráfica es utilizada para representar el modelo de competencia, un modelo computacional para un programa, o una representación particular del punto de vista de un usuario. Los modelos gráficos pueden ser utilizados para representar las diferentes visiones del conocimiento basados en probabilidad.

Un inconveniente de las redes Bayesianas es que requieren que el espacio de las variables aleatorias (los nodos), sean definidos como discretos. En un ejemplo del clima y la visibilidad, el clima puede ser fácilmente discretizado como buen tiempo, tormentas, etc., mientras que la visibilidad podría ser definida como un espacio continuo. No obstante, el modelo de red bayesiana requiere que la visibilidad sea discretizada en rangos, como por ejemplo, 0-1, 1-2 km, etc.

Los modelos gráficos son lenguajes para expresar problemas de descomposición; presentan cómo descomponer un problema en subproblemas simples. Para gráficos de dirección sin ciclos, esto se puede realizar por medio de la descomposición condicional de la probabilidad de unión (en Cios *et al.*, 1998); *M* representa el contexto o entorno y todas las declaraciones de probabilidad son relativas al entorno. La forma general para un conjunto de variables *X* es:

$$p(X|M) = \prod_{x \in X} p(x|origen(x), M),$$

es decir la probabilidad de *X* dado *M.*

Cada variable está escrita hacia abajo condicionada en su origen, donde los orígenes *(x)*, son un conjunto de variables con un arco dirigido dentro de *x.*

Muchas técnicas para descubrimiento de conocimiento confían únicamente en los datos. En contraste, el conocimiento codificado en sistemas expertos generalmente llega únicamente de un experto. Las redes bayesianas, permiten descubrir nuevo conocimiento combinando el conocimiento de un dominio experto con datos estadísticos.

El elemento primario del lenguaje de probabilidad (bayesiano u otro) es el evento. Por evento, se entiende el estado de alguna parte del mundo en algún intervalo de tiempo en el pasado, presente, o futuro.

La interpretación de una probabilidad como una frecuencia en una serie de elementos repetitivos, se refiere tradicionalmente como el objetivo o frecuencia de interpretación, es decir dado un evento *e*, la concepción que prevalece de su probabilidad, es que es una medida de la frecuencia con la que *e* ocurre cuando se repite varias veces un experimento con *e* posibles resultados. En contraste, la interpretación de la probabilidad como un grado de creencia es llamada subjetiva o interpretación bayesiana, que es la probabilidad de que *e* represente el grado de creencia que tiene una persona de que el evento *e* ocurrirá en un único experimento.

En la interpretación bayesiana, una probabilidad o creencia dependerá de la condición de conocimiento de la persona quien da la probabilidad. La probabilidad de *e* dado ξ se denomina como $p(e|\xi)$. El símbolo ξ representa el estado de conocimiento de la persona quien da la probabilidad. También, en esta interpretación, una persona pueda valorar una probabilidad basada en información que él asume como verdadera. Se escribe $p(e_2|e_1,\xi)$ para indicar la probabilidad del evento e_2 dado que el evento e_1 es verdadero y dando el nivel de

conocimiento ξ.

Una variable representa una distinción acerca del mundo. Ella toma valores de un grupo de estados mutuamente exclusivos y colectivamente exhaustivos, donde cada estado corresponde a algún evento. Una variable puede ser discreta, teniendo un número finito de estados o puede ser continua. La notación de las variable simples es con letras minúsculas del alfabeto, mientras que con mayúsculas se representan conjuntos de variables, $x = k$ indica que la variable x está en el estado k, $X = K$ indica el estado de la variable k en el conjunto X.

La distribución de probabilidad de un conjunto de variables X, denotada por $p(X|\xi)$, es el conjunto de probabilidades $p(X = k\ |\xi)$ para todas las condiciones de X. Dando conjuntos de variables X e Y, se escribe $p(X\ |Y,\xi)$ para indicar el conjunto de distribuciones de probabilidad $p(X\ |Y = k,\xi)$ para todas las condiciones de Y.

Diagramas de Influencia

Un Diagrama de Influencia (Influence Diagram, ID) también conocido como red de decisión, corresponde a una representación gráfica y matemática de una situación de decisión. Es una generalización de las redes bayesianas, en la cual no únicamente los problemas de inferencia probabilística pueden ser resueltos, sino también problemas de toma de decisiones siguiendo el criterio de máxima utilidad esperada.

Los diagramas de influencia fueron desarrollados a mediados de 1970 dentro de la comunidad de los análisis de decisión con una semántica intuitiva de fácil comprensión.

Un diagrama de influencia está representado por un gráfico dirigido sin ciclos, con tres tipos de nodos, y tres tipos de arcos entre nodos.

Los nodos son de *decisión* y se dibujan con un rectángulo, de *incertidumbre* representados por un ovalo, y un nodo *valor* representado por un octágono o un

diamante.

Los arcos son *funcionales* si finalizan en nodos valor, *condicionales* si finalizan en nodos de incertidumbre, y de *información* si finalizan en nodos de decisión.

Los nodos de decisión y los arcos de información establecen las alternativas, los nodos de incertidumbre y los arcos condicionales modelan la información, y los nodos valor y los arcos funcionales cuantifican la preferencia, qué cosas son preferidas sobre otras.

En la Figura II.17 se puede visualizar un ejemplo de un diagrama de influencia en el que se debe tomar la decisión acerca de la actividad de ocio. Se tiene un nodo de decisión (actividad ocio), dos nodos de incertidumbre (condición de clima y predicción del clima), y un nodo valor (satisfacción). Se tienen dos arcos funcionales que finalizan en satisfacción, un arco condicional finalizando en la predicción del clima, y un arco de información finalizando en la actividad de ocio.

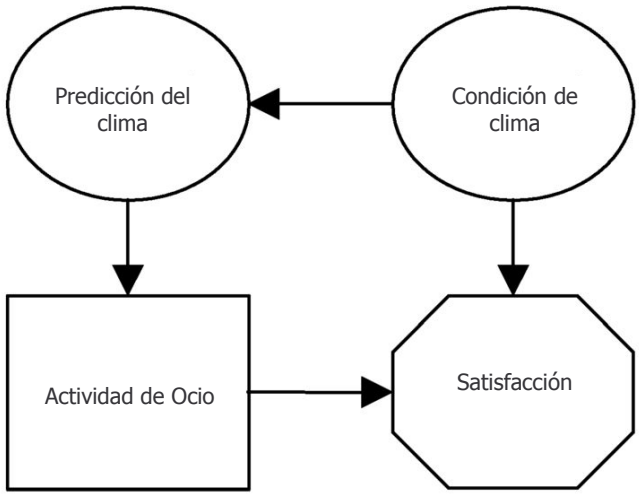


Figura II.17. Diagrama de influencia para tomar la decisión de una actividad de ocio

Los arcos funcionales finalizando en la satisfacción, indican que la satisfacción es una función de utilidad de la condición de clima y la actividad de ocio, es decir, la satisfacción puede ser cuantificada si se conoce que clima es probable, y qué actividad de ocio se tiene. El arco condicional indica la creencia de

que la predicción de clima y la condición de clima pueden ser dependientes. Y el arco de información indica que sólo se sabrá la predicción del clima cuando se tome la decisión, es decir, el clima actual solo se conocerá después de haber hecho la elección de la actividad de ocio contando solo con la predicción. Por consiguiente la actividad de ocio es independiente o irrelevante de la condición de clima.

Redes de Markov

Una red de Markov, o campo aleatorio de Markov, es un modelo de la distribución de probabilidad conjunta de un conjunto X de variables aleatorias que tienen la propiedad de Markov. Una red de Markov es similar a una red Bayesiana en su representación de dependencias, aunque en la red de Markov, se pueden representar dependencias como las cíclicas que en las redes Bayesianas no es posible representar, por otra parte no se pueden representar ciertas dependencias que son representadas por las redes bayesianas, tales como las dependencias inducidas.

Básicamente una red de Markov está representada por un grafo no dirigido *G = (V,E),* donde cada vértice *v ∈ V* representa una variable aleatoria en X, y cada lado *(u,v) ∈ E* representa una dependencia entre las variables aleatorias *u* y *v*, y por un conjunto de funciones f_K también llamados *factores o clique*, donde cada f_K tiene el dominio de algún *factor k* en *G*. Cada f_K es un mapeo de posibles asignaciones conjuntas a números reales no negativos.

La distribución conjunta representada por una red de Markov esta dada por:

$$P(X = x) = \frac{1}{Z}\prod_k f_k(x_{\{k\}}),$$

donde $x_{\{k\}}$ corresponde al estado de la variable aleatoria en el *k*-iésimo factor, y el producto repasa todos los *factores* en el grafo. *Z* es la función de partición tal que:

$$Z = \sum_{x \in X} \prod_{k} f_k(x_{\{k\}}).$$

En la práctica una red de Markov es a menudo representada como un modelo log-lineal, dado por:

$$f_k = \exp\left(w_k \phi_k\left(x_{\{k\}}\right)\right),$$

tal que

$$P(X = x) = \frac{1}{Z} \exp\left(\sum_{x} w_k \phi_k\left(x_{\{k\}}\right)\right),$$

con función de partición

$$Z = \sum_{x \in X} \exp\left(\sum_{k} w_k \phi_k\left(x_{\{k\}}\right)\right).$$

En este contexto, w_k corresponde a pesos y φ_k corresponde a funciones potenciales del factor *k* para los reales. A estas funciones también se les llama *potenciales Gibbs*; el término potencial viene de la física donde son interpretados como la energía potencial entre *vecinos cercanos*.

La *propiedad* de Marcov es que la probabilidad de que el vértice *u* del grafo esté en el estado x_u, depende solo de los vecinos próximos al vértice *u*. El vértice es condicionalmente independiente de los otros vértices del grafo. Esto se escribe como:

$$P\left(X_u = x_u \middle| X_v, v \neq u\right) = P\left(X_u = x_u \middle| X_v, v \in N_u\right),$$

donde N_u corresponde al conjunto de los vecinos próximos de *u*.

II.3.3.9. Reglas de decisión

La inducción de reglas corresponde al proceso de extraer reglas (*si-entonces*) de datos, basadas en significados estadísticos. El aprendizaje

automático (*Machine Learning (ML)*, de sus siglas en inglés) es el centro del concepto de la minería de datos, debido a su capacidad de penetración física dentro del problema, y por participar directamente en la selección de datos y en los pasos de búsqueda del modelo. Para dirigir problemas de clasificación (árboles de decisión claros y difusos), regresión (árboles de regresión), predicción temporal (árboles temporales), el campo del aprendizaje automático, básicamente se centra en el diseño automático de reglas "*si-entonces*", similares a aquellas utilizadas por los expertos humanos. La inducción de árboles de decisión es capaz de manejar problemas de gran escala debido a su eficiencia computacional, dar resultados interpretables, y en particular identificar los atributos más representativos para una tarea dada (Das *et al.*, 1998).

La inducción de reglas es utilizada para descubrir reglas automáticamente que caractericen el contenido de los datos; la representación puede ser:

$$si\ A_1 \in V_1 \wedge \ldots \wedge A_n \in V_n\ entonces\ A_{n+1} = a\ .$$

II.3.3.10. Reglas de asociación

La generación de reglas de asociación es una técnica de minería de datos utilizada para, buscar en un conjunto de datos reglas que revelan la naturaleza y frecuencia de las relaciones o asociaciones entre las entidades de los datos. Las asociaciones resultantes, pueden ser utilizadas para filtrar la información por análisis humano, y posiblemente definir un modelo de predicción basado en el comportamiento observado.

Las reglas de asociación son sentencias que pueden ser descubiertas de conjuntos con valores binarios (booleanos) 0/1. Considerando un conjunto de atributos binarios, donde cada atributo representa un grupo particular (elemento, evento), donde 1 representa presencia y 0 ausencia del grupo en una dupla, todos lo valores de 1 representan agrupamiento dentro del grupo, por ejemplo, todos los grupos fueron comprados al tiempo.

Una regla de asociación $X \Rightarrow Y$, donde $X = X_1 \wedge ... \wedge X_k$ y $Y = Y_1 \wedge ... \wedge Y_l$ significa que cuando los grupos X_1 ... X_n ocurren en una dupla con cierta exactitud, los grupos Y_1 ... Y_l ocurren también en la dupla. La exactitud de una regla es determinada como $P(Y/X)$.

Si se tiene un conjunto $R = \{A_1,..., A_p\}$ de atributos en el intervalo {0,1}, y una relación $r|R$, la regla de asociación acerca de r es una expresión de la forma $X \Rightarrow B$, donde $X \subseteq R$ y $B \in R|X$. Si $X = \{A_{i1},...,A_{ik}\}$, $A_{i1},...,A_{ik} \Rightarrow B$.

El significado de la regla es que si una fila de la matriz r tiene un 1 en cada columna de X, entonces cada fila también tiende a tener un *1* en la columna B, es decir:

$$A_{i1} = 1 \wedge . . . \wedge A_{ik} = 1 \Rightarrow B = 1$$

II.3.3.11. Métodos de agrupamiento

Son utilizados en el paso de pre-procesamiento de los datos, debido a la característica de aprender semejanzas sin supervisión entre elementos y reducir el espacio de búsqueda a un conjunto de los atributos más importantes para la aplicación, o a un conjunto finito de elementos. El método que con mayor frecuencia se utiliza para agrupar es el *K-promedios* (*K-means*) (promedio = localización media de todos los miembros de una clase particular), el cual identifica un cierto número de grupos o elementos similares; puede ser utilizado conjuntamente con el método de la *Vecindad más próxima* (*K-Nearest Neighbor K-NN*); esta técnica coloca un elemento de interés dentro de clases o grupos examinando sus atributos y agrupándolo con otros cuyos atributos son cercanos a él. *K-NN* es una técnica clásica para descubrir asociaciones y secuencias cuando los atributos de los datos son numéricos. Con atributos no numéricos o variables, es difícil aplicar esta técnica por la dificultad de definir una medida que pueda ser utilizada para cuantificar la distancia entre un par de valores no numéricos.

La clave del algoritmo iterativo *K-means* es minimizar la función objetivo

por la localización de los puntos de datos para los diferentes grupos (en *Cios et al.*, 1998). El algoritmo de *K-means* opera como sigue:

Suministrando el número de grupos (K), la función de distancia y el criterio de parada,

1. Seleccionar las semillas (grupo embrionario con un solo elemento) iniciales *c*.
2. Calcular la distancia entre el elemento (x_k) y los centroides de los grupos, d_{ik} i = 1, 2, ..., c (para la primera iteración es la semilla escogida).
3. Asignar el elemento al grupo cuyo centroide este más próximo a él.
4. Recalcular el centroide del grupo de los elementos asignados
5. Repetir los pasos 2-4 hasta que el criterio de parada este satisfecho.
6. Resultado: un grupo de promedios (prototipos) de los grupos.

Entre las debilidades del algoritmo de *K-means* se tienen que no se conocen de antemano el número de clases a ser tenidas en cuenta (se parte de un número elegido), así como, que si algunas de las medias iniciales están muy cercanas entre si no se pueden recalcular. Por el contrario los algoritmos de *agrupamiento jerárquico* no requieren el número de clases como un dato de entrada, ya que están basados en la construcción de un árbol (dendograma) en el que las hojas corresponden a los elementos del conjunto de ejemplos y el resto de los nudos corresponden a subconjuntos de ejemplos que pueden ser utilizados como particiones del espacio. Este tipo de algoritmos es utilizado cuando se tienen subclases embebidas dentro de los *clusters*.

II.3.3.12. Técnicas de visualización

Ya que es difícil imitar la intuición humana y el proceso de la toma de

decisiones mediante una máquina, se debe trasformar el conocimiento obtenido en un formato fácilmente entendible tal como las imágenes o los gráficos.

Inicialmente, la utilidad de los gráficos es para la limpieza de datos, donde se pueden visualizar conjuntos inusuales de datos, evaluar distribuciones de atributos, y sugerir transformaciones. Luego de que el modelo de minería de datos ha sido construido, se pueden utilizar técnicas similares para validar los resultados del modelo, resaltar resultados inusuales o atípicos, y presentar el modelo de forma que facilite su comprensión, Cook y Holder, 2000.

Con el surgimiento de las herramientas computacionales más económicas y con mejores prestaciones gráficas, el usuario es no sólo un reproductor de un gráfico estático, sino que, a su vez, puede interactuar con este, cambiando parámetros, expresando dudas (querying), acercamientos (zooming), y enlazando gráficos, con el fin de sacar a la luz rasgos importantes del gráfico.

Las técnicas de visualización son útiles para mostrar los resultados de los métodos de minería de datos y hacerlos más asequibles al usuario final que no tiene por qué ser necesariamente un especialista en el tema. Algunas de las técnicas empleadas para la representación gráfica de resultados generados por la minería de datos son los ***histogramas*** (estimando la distribución de probabilidad para ciertos atributos numéricos dados en un conjunto de elementos), ***gráficas de dispersión*** (proporcionan información sobre la relación entre dos atributos numéricos y uno discreto), ***gráficas tridimensionales, dendrogramas*** (análisis de correlación entre atributos u elementos), es la representación visual de los pasos de una solución de *clusters* o conglomerados jerárquica que muestra los conglomerados combinados y los valores de los coeficientes de distancia en cada paso. Los casos agrupados se indican mediante líneas verticales conectadas entre sí. El dendrograma no muestra las distancias reales sino que les aplica un cambio de escala para que sus valores estén comprendidos entre por ejemplo entre 0 y 25. Así se conserva la proporción entre las distancias de un paso a otro. La escala que aparece en la parte superior de la figura de un dendograma corresponde a estas distancias reescaladas. El eje vertical consiste en los objetos o individuos, y

el eje horizontal consiste en el número de conglomerados formados en cada paso del procedimiento.

La clasificación o agregación de un grupo de elementos se puede visualizar por medio de dendrogramas, sobre los cuales se puede evaluar el grado de similitud de objetos por medio de una distancia. Estas distancias corresponden a distancias ultramétricas que cumplen las siguientes condiciones:

$$
\begin{aligned}
&d(x,y) > 0 \quad para\ x \neq y \\
&d(x,y) = 0 \quad para\ x = y \\
&d(x,y) = d(y,x)\ \ \forall x,y \\
&d(x,y) \leq d(x,z) + d(y,z)\ \ \forall x,y,z \quad (Triángulo\ no\ equilatero) \\
&d(x,y) \leq \max\left(d(x,z), d(y,z)\right).
\end{aligned}
$$

En cuanto a la representación gráfica se puede realizar y puede ser mejorada si se utilizan diferentes formas de presentación, es el caso de los marcadores de datos en formas de cono, cilindro y pirámide que pueden mejorar la presentación de gráficos de columnas. En el capítulo IV se pueden visualizar varios tipos de representación gráfica utilizadas en esta tesis.

II.4. Notas finales

Este capítulo contiene información de utilidad para la comprensión general del paso del KDD de minería de datos, y del descubrimiento de conocimiento en bases de datos. Por otra parte, el siguiente capítulo del estado del arte del tema incluye desarrollos que hacen uso de lo expuesto anteriormente, lo que permite un mejor entendimiento de este capítulo; como es el caso de los conjuntos aproximados, la lógica difusa o los algoritmos genéticos, entre otros.

En cuanto a las generalidades acerca de las redes de abastecimiento, se ha expuesto la temática más con la finalidad de posibles aplicaciones hacia la metodología propuesta en esta tesis, que como un tratado de hidráulica urbana. Por este motivo, se inicia el capítulo con los componentes del sistema de distribución de agua; las tensiones que actúan sobre las tuberías; la problemática

de las fugas, sin duda, un tema de posible aplicación de las técnicas de tratamiento de datos expuestas como complemento de las técnicas de detección de fugas tradicionales; y el enfoque de análisis de fiabilidad del sistema, desde el punto de vista de posibles ataques o intrusiones de contaminantes en la red, ya sea debido a propios fallos del sistema o a ataques terroristas.

Capítulo III

Antecedentes y estado del arte

III.1. Introducción

El manejo de los datos y la información es quizá uno de los temas de mayor importancia en el desarrollo científico de cualquier actividad. Día a día se recopila información de diversas fuentes y en diferentes formatos, parte de la cual se estudia y analiza, pero la gran mayoría es descartada al no ofrecer, de forma aparentemente directa y obvia, repuestas al interés u objetivo por lo que se buscó o recopiló tal información.

La revisión bibliográfica de los desarrollos y aplicaciones de la búsqueda de información, por medio de lo que podríamos denominar minería de datos empleadas en temas de abastecimiento de agua resulta algo escasa; es cierto que nos encontramos ante una información que presenta cierto grado de sensibilidad debido principalmente a que las empresas de abastecimiento, siendo el agua un bien público, son en su mayor parte privadas o mixtas, con lo cual la disponibilidad de esta información se presenta renuente al dominio público, y no es fácil conseguir los datos necesarios y suficientes para el desarrollo de este tipo de investigaciones. No hay que dejar de lado la posibilidad acerca de que las disputas en el futuro no tendrán que ver con los combustibles fósiles sino con el poder del agua, y por esto la importancia en realizar estudios e investigaciones tendientes a hacer un uso más racional del recurso; pero también es cierto que en la ingeniería práctica pareciera existir cierto recelo hacia la innovación en determinados casos como el expuesto en este documento, hacia la búsqueda y aplicación de técnicas que muestran su gran potencial en sinnúmero de aplicaciones en diferentes ramas de la ciencia, y se prefiere el diseño y modelado clásico sobre la búsqueda de alternativas que puedan complementar este punto de vista.

En este capítulo se presenta una revisión del estado del arte acerca del uso del descubrimiento de conocimiento en grandes bases de datos y de la minería de datos aplicada a temas de abastecimiento de agua, y en algunos casos se hace una descripción más profunda debido al interés práctico que suscitan. Se aprecia en el desarrollo del arte encontrado que ha habido, en la última década, un

interés especial en temas relacionados con la optimización, ya sea para rehabilitación de sistemas de abastecimiento o en cuanto a la estimación de la demanda; pero para la gran cantidad de información que se genera en cualquier sistema de abastecimiento de agua está siendo poco utilizado el potencial de las herramientas y técnicas del *descubrimiento de información en grandes bases de datos*.

III.2. Estado del arte

Reich (1997) presenta una publicación en la cual se estudian aplicaciones prácticas de aprendizaje automático junto a una propuesta de proceso que puede guiar sus aplicaciones en problemas de ingeniería civil. En este artículo se asevera que no se ha implementado el uso del aprendizaje automático en problemas de ingeniería 'tal como se puede constatar hoy día', principalmente por dos razones. Los problemas prácticos son a menudo más complejos para ser manejados por un único método, y la aplicación de las técnicas de aprendizaje automático no es tan simple como tomar un programa y aplicárselo a los datos. A pesar de que la afirmación expuesta por el autor ya tiene más de una década, podemos confirmar que para el caso de los sistemas de abastecimiento de agua tiene bastante vigencia.

La revisión del estado del arte que se presenta en esta publicación, evidencia hasta cierto punto el escaso grado de interés que ha auspiciado, dentro de la comunidad científica y práctica de la ingeniería civil, la utilización de las herramientas que permitan obtener información de la "información", que a su vez pueda ser utilizada para mejorar procesos, gestionar recursos y empresas, entre otros y que, a su vez, se sigue percibiendo hoy en día.

"En un sistema de abastecimiento que requiera de bombeos, ya sea para alimentar embalses o directamente la red de distribución, es muy importante optimizar el costo energético de la utilización de las bombas", *(**An et al., 1997**)*. Muchos operadores de las estaciones de bombeo pueden utilizar heurística o

reglas para minimizar el costo de la energía utilizada por las bombas, o realizar predicciones de demandas, o ayudar a mantener el nivel de los embalses en rangos aceptables. Esta publicación corresponde a un ejemplo claro de la utilidad y el aprovechamiento que se le puede dar a las técnicas de aprendizaje automático en la ayuda de la operación y gestión de un abastecimiento de agua.

"Una de las ventajas de la utilización de técnicas de KDD (*Knowledege Discovery Databases)* para generar reglas de predicción de demanda de agua es que los modelos matemáticos sólo tienen en cuenta la parte cuantitativa del problema, mientras que con el aprendizaje automático a partir de los datos observados se consideran variables tanto cuantitativas como cualitativas".

An et al. (1995) y *An et al. (1997)* predicen demandas de agua por medio de la teoría de los conjuntos aproximados. El objetivo es el descubrimiento automatizado de reglas a partir de una muestra de datos para realizar predicciones de la demanda diaria de agua; los datos utilizados se obtuvieron del departamento municipal de agua en Regina, Saskatchewan, Canadá (*An et al.,* 1997). La base de datos contiene 306 muestras recogidas durante 10 meses que cubren la información de 14 factores ambientales y sociológicos, así como su correspondiente volumen diario de distribución de caudal.

Los posibles factores que afectan (atributos de condición) el consumo diario de agua son: el día de la semana, y factores climáticos agrupados en temperatura, humedad, precipitación, viento, y horas de brillo del sol. Los consumos diarios (atributos de decisión) totales de agua se tomaron sumando las distribuciones diarias en cada estación de bombeo de la ciudad. Las 306 muestras de entrenamiento, en el sistema de información para la predicción de demanda de agua son elementos que incluyen información diaria de los atributos de condición y decisión para 10 meses, de marzo a diciembre de 1994.

Ya que en la definición de *aproximación inferior y superior* de la teoría de los conjuntos aproximados no se tiene en cuenta la información estadística en la región límite, en esta aplicación se hizo uso del *espacio de aproximación* β para

intentar rectificar esta limitación.

El *espacio de aproximación* β AS_P es un triple <*U,R(C),P*> donde *P* es una medición de probabilidad y β es un número real en el rango (0.5,1). Este espacio de aproximación puede ser dividido en dos regiones:

Región β-*positiva* del conjunto *Y:*

$$POS_C(Y) = \bigcup_{P(Y|X_i)\geq\beta} \{X_i \in R*(C)\}$$

Región β-*negativa* del conjunto *Y:*

$$NEG_C(Y) = \bigcup_{P(Y|X_i)<\beta} \{X_i \in R*(C)\}$$

Para la obtención de las reglas de decisión probabilísticas se siguió el siguiente procedimiento:

Tomando $R*(RED) = \{X_1, X_2, \ldots X_n\}$ como un conjunto de clases de equivalencia de la relación *R*(*RED*), donde *RED* es un conjunto reducido de atributos de condición *C* en *S,* y $R*(D) = \{Y, \neg Y\}$, es una partición inducida por los atributos de decisión, se tiene que cada clase de equivalencia X_i de la relación de equivalencia *R*(*RED*) está asociada con una combinación única de valores de atributos que pertenecen a *RED*. Esta combinación de valores es llamada *descripción* de las clases equivalentes $X_i \in R*(RED)$; la descripción de X_i se puede expresar cómo:

$$Des(X_i) = \bigwedge_{a\in RED} (a = f(x_i, a)),$$

donde $\wedge$ indica el operador de conjunción y x_i es un elemento de las clases de equivalencia X_i. Las descripciones de *Y*, $\neg Y$ son:

$$Des(Y) = (D = f(x_i, D))$$

$$Des(\neg Y) = (D \neq f(x_i, D)),$$

donde *D* es el atributo de decisión y $x_i \in Y$. Las siguientes reglas describen las

relaciones entre la partición *R**(*RED*) y la partición *R**(*D*):

Para $X_i \in R^*$(RED),

$$Des(X_i) \rightarrow^{Ci} DES(Y), si\ P(Y|X_i) \geq \beta$$

$$Des(X_i) \rightarrow^{Ci} DES(\neg Y), si\ P(Y|X_i) < \beta,$$

donde $_{Ci}$ es el factor de incertidumbre, el cual es igual a P($Y|X_i$) en el primer caso y 1 – P($Y|X_i$) en el segundo. Esto significa que si el elemento x_i satisface la descripción Des(X_i) y P($Y|X_i$) $\geq$ β entonces el elemento x_i definitivamente pertenece a Y con un factor de incertidumbre; por otra parte, si P($Y|X_i$) < β el elemento x_i posiblemente pertenece al complemento $\neg Y$ con incertidumbre $_{Ci}$.

La técnica empleada para encontrar el máximo de reglas generales fue la *matriz de decisión*; ya que las reglas anteriores pueden contener atributos que son irrelevantes para determinar el concepto, se pueden obtener reglas inspeccionando qué atributos de condición pueden ser removidos en una regla sin causar inconsistencias.

Tanto la información básica como los resultados fueron agrupados en rangos discretos, tanto para su procesamiento como para su interpretación. En total se generaron 149 reglas para los diferentes conceptos o rangos de predicción de demanda de agua.

Algunos ejemplos del tipo de reglas generadas son los siguientes:

Para el rango D = [53 – 60],

$(a_0 = (D \text{ or } L \text{ or } MA) \wedge (a_5 > 64) \wedge (a_{10} <= 10.84) \wedge (a_3 <= -3.36) \rightarrow^1 (53 < D <= 60)$.

Esta regla cubre el 66.7% de los elementos de entrada que incluyen el rango, y establece que sí el día de la semana es domingo, lunes o martes, y la humedad mínima es mayor que 64, y el promedio de velocidad del viento es menor o igual a 10.84, y la temperatura media es menor o igual que –3.36,

entonces la demanda de agua está entre 53 y 60 con una probabilidad 1, es decir que la totalidad de los elementos seleccionados para el rango cumplen la regla.

Para el rango D = (89 – 90],

$$(a_1 <= 23.18) \wedge (a_{12} <= 36.88) \wedge (a_3 > 10.78) \wedge (50 <= a_5 <= 64) \rightarrow^{1} (80 < D <= 90).$$

Esta regla cubre el 10.5% de los elementos de entrada que incluyen el rango. Establece que sí la máxima temperatura es menor o igual que 23.18, y la máxima velocidad del viento es menor o igual que 36.88, y la temperatura mínima es mayor de 10.78, y la humedad mínima está entre 50 y 64 inclusive, entonces la demanda de agua está entre 60 y 90 con una probabilidad de 1.

Para el rango D = (100 – 110],

$$(a_2 > 10.78) \wedge (a_{12} > 27.03) \wedge (a_5 <= 31) \wedge (a_{13} > 9.60) \rightarrow^{1} (100 < D <= 110).$$

Esta regla cubre el 33.3% de las muestras que incluyen el rango, y establece que si la temperatura mínima es mayor de 10.78, y la máxima velocidad del viento es mayor de 27.03, y la mínima humedad es menor o igual a 31, y el número de horas de brillo del sol es mayor de 9.60, entonces la demanda de agua está entre 100 y 110 con una probabilidad de 1.

"Las redes de distribución de agua corresponden a sistemas geográficamente distribuidos con gran heterogeneidad en cuanto a sus estructuras de control, las estrategias de manejo, y la variabilidad geométrica en continua expansión y cambios en la demanda a lo largo de su vida útil. Por estas características, las empresas de distribución de agua deben encarar el problema de la integración de los datos y el conocimiento en relación con el control y la explotación óptima" ***(Afsarmanesh et al., 1997)***. Muchas veces el control local de los sistemas es llevado a cabo basándose en la experiencia de los operadores; no hay un control coordinado que garantice a la vez el suministro continuo, que coloque estándares de calidad, el ahorro de energía, la optimización de tuberías y

la reducción de pérdidas. Los autores presentan el esquema de un modelo (el sistema WATERNET) para el manejo y control de la red de abastecimiento que permite disminuir los costos de explotación, garantizar el suministro continuo de agua con un monitoreo de mejor calidad, el ahorro de energía y minimizar las pérdidas del recurso natural. El sistema está compuesto por cinco subsistemas (manejo de la información, optimización, calidad del agua, modelado y simulación, aprendizaje), y un sistema de supervisión que integra estos subsistemas para asistir en la decisión y operación óptima de la red. Se presentan algunos experimentos preliminares con datos históricos de una red aplicando técnicas de aprendizaje automático. El artículo presenta una descripción de la situación en el momento en cuanto a cómo se realizaba el control y manejo de la información en los sistemas de abastecimiento, y podría tomarse como base de en qué dirección marchar para organizar y facilitar el control del sistema.

De este mismo proyecto, ***Camariha-Matos y Martinelli (1999)*** presentan una publicación enfocada en los aspectos de aprendizaje automático, principalmente los procesos de adquisición de conocimiento (minería de datos) y el uso del conocimiento adquirido para mejorar la operación de la red. La justificación se basa en los procedimientos limitados utilizados para operar la red y en los factores que influencian esta operación (bajo la disponibilidad e información histórica), y en la característica de expansión continua y modificación de las redes de distribución. La información de la que se dispone corresponde a variables físicas (caudales, presiones, niveles), variables químicas (PH y cloro), estado de dispositivos (bombas encendidas o no, grado de abertura de válvulas), operaciones (algoritmos de control u operaciones efectuadas por trabajadores) y alarmas.

Las tareas identificadas por los autores como posibles para la utilización de técnicas de aprendizaje de datos incluyen: (a) planeamiento como soporte a la producción, (b) identificación de factores que influencian acciones, (c) monitoreo y manejo de alarmas, (d) mantenimiento preventivo, (e) mejoramiento de la satisfacción del usuario. Como resultado de una fase de selección e identificación de prioridades teniendo en cuenta factores tales como la restricción de tiempo,

viabilidad y utilidad, los autores decidieron implementar dos tareas de aprendizaje, *la predicción de demanda de agua, y el error en monitoreo y manejo de alarmas*. La predicción de la demanda tanto horaria como diaria se realiza haciendo uso de la técnica de redes neuronales artificiales. El conocimiento adquirido acerca de errores en el monitoreo y manejo de armas es presentado en forma de reglas, las cuales se generan por un algoritmo de aprendizaje inductivo.

Savic y Walters (1999) realizan un estudio piloto para determinar los beneficios de la minería de datos en la identificación de las causas de la rotura de tuberías en un abastecimiento. Los resultados obtenidos indican que la aplicación de las técnicas utilizadas (árboles de decisión y análisis estadístico) daría buenos resultados en la identificación de relaciones-causa de roturas en tuberías, así como la presentación de una serie de recomendaciones para el trabajo posterior.

"Los principales ítems de desembolso en la operación de un sistema de abastecimiento de agua corresponden a los costos de mantenimiento, incluyendo la rehabilitación, reemplazo y expansión del sistema existente para mantener la función principal del sistema en cuanto a la distribución continua de agua a una presión adecuada" (*Savic y Walters 1999*). Por tanto, parte integral de la estrategia en la gestión de las redes debe estar orientada hacia planeamientos que permitan tener los volúmenes de caudal apropiados para la situación presente y futura, así como, la garantía de las presiones en el sistema, y la reducción de costos futuros de mantenimiento.

Los autores introducen el uso de la *"hidroinformática"* como una herramienta de técnicas innovadoras para la gestión de redes de abastecimiento, que tiene origen en las ciencias de cálculo computacional y la inteligencia artificial. El uso de las técnicas propuestas como particularmente adaptables en aplicaciones de la industria del agua, corresponde a los sistemas de información geográfica (SIG), y la minería de datos (las redes neuronales artificiales y los algoritmos genéticos específicamente). El trabajo se enfoca hacia la atención en la necesidad de las estrategias de mantenimiento en las redes de distribución de agua; se presenta una revisión del estado del arte de cómo se ha enfocado el problema

(optimización compleja). "La clave del éxito en la gestión de la red consiste en garantizar el desempeño óptimo técnico en un riesgo acordado y asociado al costo mínimo en cada una de las llamadas Áreas Críticas". Estas áreas son identificadas por la compañía de abastecimiento (calidad del agua, balance entre suministro y demanda, manejo de activos, gestión de la información, servicio al cliente, por ejemplo).

Reich y Barai (1999) presentan una publicación crítica de artículos publicados entre 1992 y 1997 en *"Artificial Intelligence in Engineering"*, acerca de la evaluación de modelos de aprendizaje automático en problemas de ingeniería. Concluyen que, para estimar correctamente la exactitud de los modelos de aprendizaje automático, se debe seleccionar cuidadosamente el método apropiado, ya que los resultados obtenidos por diferentes modelos pueden ser muy diferentes; se presenta un resumen acerca de la teoría de estimadores para la evaluación de modelos, así como los principales métodos de estimación del error en modelos de aprendizaje automático.

Las afirmaciones realizadas sobre las cuales sustentan su artículo son que la elección del método de evaluación es crítica, los resultados obtenidos por diferentes métodos puede ser muy diferente, y que las relaciones entre diferentes resultados de evaluación varía con las propiedades de los datos y el objetivo y sistema de aprendizaje, y que, por tanto, no es posible determinar aquellas relaciones a priori. Los métodos expuestos para estimar el desempeño de los modelos de aprendizaje automático son: *resustitución, hold-out, validación cruzada (leave-one out o k-fold),* y *bootstrap*. En la *resustitución,* la serie completa de datos se utiliza para entrenar la red, lo cual tiende a hacer que sea un método altamente optimista. En *Hold-out,* los datos son aleatoriamente divididos en datos de entrenamiento y de prueba, generalmente 2/3 de los datos para entrenamiento y 1/3 para prueba. La *validación cruzada,* divide los datos en *k* subconjuntos de aproximadamente igual tamaño; el programa de aprendizaje automático es entonces entrenado *k* veces dejando fuera cada vez uno de los subconjuntos de entrenamiento para utilizarlo como prueba y estimando el error como el promedio de todos los entrenamientos. En *bootstrap,* una muestra de *n*

ejemplos es ejecutada con reemplazo de los *n* ejemplos originales, donde cada ejemplo tiene igual probabilidad a ser muestreado; en promedio, $1 - 1/e = 0.632$ de los ejemplos originales son corridos dentro de la muestra, la nueva muestra es utilizada para entrenamiento y la muestra anterior para prueba.

El interés de este trabajo radica en el énfasis que realiza en el cuidado que se debe tener para la evaluación de los métodos de aprendizaje automático, en especial las redes neuronales, aunque desde la fecha de publicación del trabajo el software de modelado de técnicas de minería de datos ha mejorado en cuanto a la implementación de diferentes métodos de evaluación.

Tachibana y Ohnari (1999) como resultado de observaciones horarias durante tres años de datos de consumos de agua en una planta de tratamiento en el área metropolitana de Japón, y haciendo uso de *agrupamientos (clustering)* de acuerdo con la forma de ondas que tiene el consumo diario de agua, extraen regularidades de los datos y proponen un modelo para la predicción de la demanda, en especial los días festivos para los cuales varía el patrón de consumo. Agrupando los gráficos de *ondas* por atributos, tales como el día de la semana, clima y rangos de temperatura, se generan 504 grupos o *clusters* y 504 patrones base del modelo de predicción, lo cual no parece apropiado para una estructura de modelo de predicción. Para reducir esta cantidad de patrones y las discrepancias que se presentan para las condiciones (día, estado del tiempo y temperatura), se hace uso de curvas acumuladas para agrupar el consumo horario de agua. Con esto se obtiene qué los gráficos de ondas pertenecientes a diferentes *clusters* sean clasificados en el mismo grupo, reduciendo el número de patrones base.

Un inconveniente del modelo, tal como lo mencionan los autores, es que el consumo de agua se ve afectado también por factores que fluctúan a largo plazo, tales como la densidad de población y el tipo de consumo si es residencial, industrial, comercial, institucional, etc. Otro inconveniente es que no se define claramente cómo se realiza el agrupamiento de *clusters* para hacer la curva acumulada de consumos; se da a entender que se acumulan días diferentes con

condiciones climáticas diferentes e iguales temperaturas, obteniendo patrones de consumo casi iguales.

Lertpalangsunti et al. (1999) desarrollan una herramienta para realizar predicciones; se presentan los beneficios de utilizar módulos múltiples de redes neuronales artificiales para la predicción de la demanda. Los resultados de esta aproximación se comparan con regresión lineal y un programa de razonamiento basado en casos. La herramienta fue utilizada para realizar predicciones de las demandas de agua en el sistema de abastecimiento de agua de Regina (Canadá); presenta el inconveniente de que su manejo no resulta sencillo si no se tiene experiencia en el sistema bajo el cual fue implementada la herramienta (Real-Time Expert System Shell G2), aunque se da una explicación extensa acerca del funcionamiento de la herramienta. La información utilizada corresponde a datos del día de la semana, operación de las bombas, caudales horarios y nivel de embalses, combinado con factores ambientales (temperatura, humedad, lluvia, nieve y velocidad del viento). Para estimar el error se utilizó el *porcentaje de error medio absoluto*. Se realizaron cinco experimentos previos para mejorar la estimación de la demanda; el primero para la elección del número de series temporales y el número de neuronas ocultas de la red; el segundo para probar qué otras variables se deben incluir además del aspecto temporal; el tercero para mejorar la estimación aumentando el número de datos de entrenamiento; el cuarto para intentar mejorar la estimación separando los datos en múltiples series; y el quinto para comparar las técnicas de razonamiento basado en casos, regresión lineal y redes neuronales.

Se utilizaron 3 y 5 días previos de demanda de agua para hacer la estimación de la demanda del día siguiente, y 50% de los datos de entrenamiento y los otros 50% de prueba. Los mejores resultados se obtienen para 3 días de demanda previa y 5 neuronas ocultas. En cuanto a los factores ambientales se encontró que sólo la temperatura reduce el error tanto en los datos de entrenamiento como de prueba, mientras que los otros tienden a sobreajustar los datos de entrenamiento. Los datos de entrenamiento se aumentaron al 90% para mejorar la estimación, haciendo uso sólo de la temperatura como variable

adicional a las series temporales de demandas; con esto se logró reducir el máximo error en los datos de prueba del 36.87% al 25.61%. En cuanto a la creación de múltiples series de datos, no se obtienen buenos resultados si se separa la temperatura, en cambio se nota un comportamiento diferente en la demanda los sábados, por tal motivo se dejó una serie de domingo a viernes y otra para los sábados reduciendo el error de la estimación. En cuanto al último experimento, las redes neuronales se presentan como mejores estimadoras comparándolas con las otras dos técnicas

"Grandes bases de datos se han incrementado en aplicaciones de infraestructura civil. Aunque es relativamente fácil "realizar preguntas" de bajo nivel en estas bases, no es tan sencillo encontrar respuestas a, por ejemplo, ¿cómo afectan las condiciones ambientales la rotura de pavimentos?", basándose en métodos tradicionales. Las técnicas de minería de datos pueden ofrecer una solución acerca de la obtención de conocimiento de las bases de datos en la infraestructura de obras civiles; ***Buchheit et al. (2000)*** hacen uso de la técnica de Conjuntos Aproximados (*Rough Sets*), para evaluar la operación de un *Espacio Inteligente de Trabajo*, en cuanto a su consumo de energía, el confort, la flexibilidad organizacional y la adaptabilidad tecnológica.

Para realizar la evaluación del desempeño operacional del sistema inteligente se toman mediciones del sistema de aire acondicionado (temperatura, flujo, estado en intervalos de 30 minutos), del consumo eléctrico (cantidad de kilovatios hora en cada terminal cada 30 minutos) y de las condiciones ambientales del espacio (temperatura, humedad relativa, radiación solar, presión, punto de rocío, velocidad y dirección del viento, y lluvia en intervalos de 5 minutos). Se obtiene un total de 400 atributos y 600MB de información durante un año.

Los resultados obtenidos se expresan por medio de reglas del tipo (SI... ENTONCES... Y... ENTONCES...), para clasificar el confort dentro del edificio. Este artículo es una buena aplicación básica de la utilidad de la minería de datos para la ayuda de la toma de decisiones, además de presentar de una forma clara los

pasos seguidos en la elaboración de un modelo de descubrimiento de conocimiento a partir de datos.

Le Gat y Eisenbeis (2000) hacen uso de la técnica estadística clásica del análisis de supervivencia con un modelo Weibull para predecir fallas en redes de agua, basándose en registros de mantenimiento de dos compañías de agua. Entre las variables tenidas en cuenta para la primera de las compañías se tiene la longitud de la tubería, la edad del tubo, el diámetro, el tipo de unión (junta o pegado), tipo de suelo, nivel de tráfico, clase de suministro (gravedad o bombeo). Para uno de los modelos se cuenta con información de 9 años y para el otro desde 1926, en la cual no se tiene información del tipo de suelo y tráfico. Para el análisis la información se estratificó de acuerdo con los materiales de la tubería. Los resultados obtenidos para el caso en que se dispone de poca información pero se tienen datos ambientales, son bastante buenos en la estimación de las fallas; para el caso en que se tiene mayor información, se cree que se podría mejorar la estimación, aunque no es mala, si se tuviera información ambiental.

Sforna (2000) haciendo uso de redes neuronales de Kohonen (Self-organizizng Kohonen map) y lógica borrosa (Fuzzy Logic), identifica valores anómalos de consumo, es decir, que superan los límites contractuales y el factor de potencia, en una base de 91020 clientes comerciales de una compañía eléctrica, entre 10 kW y 1 MW de demanda máxima contractual, con un tamaño de la base de datos de 150 Mbytes. La utilización del sistema indujo sorpresa entre los expertos ya que automáticamente se encontraron características escondidas en una gran cantidad de datos. Adicionalmente, la técnica permite la creación de *clusters* de clientes permitiendo la posibilidad de introducir nuevas estrategias de mercado.

León et al. (2000) presentan un sistema híbrido experto (EXPLORE) con el objetivo de satisfacer la demanda de agua (decisión acerca del caudal diario en la planta de tratamiento prediciendo las demandas, manteniendo los niveles de los tanques dentro de márgenes seguros), reduciendo los costos energéticos generados por el bombeo, manteniendo la calidad del agua tratada. El sistema

experto está basado en reglas que contienen el conocimiento necesario para la operación del sistema de abastecimiento. Estas reglas son combinadas con conocimiento de expertos y métodos matemáticos. La aplicación se hace sobre el sistema de abastecimiento de Sevilla (España), obteniendo un 25% de reducción de costos energéticos, además de estructurar la experiencia de los operadores y simplificar nuevas estrategias por simulación. En el trabajo no queda claro, al realizar la estimación de las demandas de agua, a qué porcentaje de error corresponde la demanda, es decir; la variabilidad de consumo a lo largo del día no se refleja en el error estimado al comparar la demanda estimada con la real, hecho que podría afectar las horas punta; aunque es cierto que según lo reportado para solo un 9% de la información el error de estimación está entre el 10% y 15%.

Utilizando redes neuronales con el arquetipo de aprendizaje con resolución múltiple (NNMLP) se han estimado los caudales mensuales de junio de 1985 a octubre de 1986 en la cuenca de Clear Boggy en Oklahoma con base en el registro histórico de caudales diarios de 1948 a 1986 ***(Liang, X.; Liang, Y., 2001)***.

La estructura de red neuronal utilizada fue sin ciclos 12-5-1 (red de tres capas con 12 nodos de entrada, 5 neuronas en la capa escondida y una neurona de salida), las neuronas escondidas utilizan la función de activación sigmoidea y la de salida una función lineal. El proceso de entrenamiento se comenzó con pesos aleatorios y se utilizó un procedimiento de back-propagation en todas las actividades de aprendizaje. Se utilizaron 440 datos mensuales para entrenar la red y el periodo estimado corresponde a junio de 1985 a octubre de 1986.

Fueron utilizados tres criterios para evaluar el desempeño de la NNMLP; el error de la raíz cuadrática media (RMS), el coeficiente de determinación (E_coef) y la diferencia del volumen total anual entre los caudales estimados y los observados.

El coeficiente de determinación es utilizado para medir la habilidad de estimación de diferentes métodos y se define como:

$$E\_coef = 1 - \frac{\sum_{i=1}^{n}(Q_{i,f} - Q_{i,obs})^2}{\sum_{i=1}^{n}(Q_{i,obs} - \overline{Q_{obs}})^2},$$

donde *n* es el número de meses con datos, $Q_{i,f}$ y $Q_{i,obs}$ son respectivamente los caudales estimados y observados para el mes *i*, y $\overline{Q_{obs}}$ es el caudal climatológico mensual es cual se obtiene basado en los datos mensuales de 1948 – 1984.

Los resultados obtenidos presentan que el error cuadrático medio es de 19.9 mm/mes, y el coeficiente de determinación es de 0.84. La diferencia del volumen total anual entre el caudal estimado y el observado es de 159.7 mm/año.

Babovic et al. (2001a) hacen uso de la programación genética para generar una serie de ecuaciones en temas ecológicos, que son comparadas con formulaciones presentes en la literatura, específicamente para determinar la velocidad de sedimentación de materia fecal (de importancia en procesos de tasas de sedimentación, ciclos geoquímicos y disponibilidad de nutrientes) de organismos marinos. Los autores resaltan el hecho de que la programación genética es capaz de crear una ecuación de velocidad de sedimentación con fenómenos similares a aquellos identificados por expertos, aún con componentes no lineales ligeramente diferentes. Se concluye que estas formulaciones creadas a partir de la programación genética son tan exactas sino más que las formulaciones tradicionales.

"Los costos económicos y sociales asociados a la rotura de tuberías y al problema de fugas en los sistemas de abastecimiento de agua se incrementan rápidamente a niveles inaceptables" ***(Babovic et al., 2001b y 2002)***. La red de tubos de una ciudad y todos los componentes asociados con esta red (válvulas, bombas, reservorios, etc.) constituyen los activos de un suministro de agua, y como en cualquier otro activo, es importante invertir en su mantenimiento para que cumplan con su tarea. La motivación de este trabajo se produjo debido a la política implementada en la ciudad de Copenhague (Dinamarca), en los años 80 de reemplazar un 1% de longitud de la tubería de la red de abastecimiento de

agua por año debido a razones económicas, pérdidas de agua, capacidad, calidad del agua, reclamaciones y compensaciones, cooperación en trabajos de construcción y visión a largo plazo.

Esta es una de las pocas publicaciones acerca de la posible utilización práctica de las herramientas del manejo de datos aplicadas a los sistemas de abastecimiento de agua. Haciendo uso de técnicas donde la comprensibilidad de los resultados obtenidos así como de la metodología utilizada es clara, se obtienen resultados llamativos acerca de la utilidad y posible utilización de las herramientas en la ayuda a la gestión de un sistema de abastecimiento.

El artículo propone la utilización de técnicas de minería de datos para determinar el riesgo de rotura de tuberías. Por ejemplo, el análisis de bases de datos de los alrededores de donde ocurren los eventos de rotura pueden ser utilizados para establecer un modelo de riesgo como función de las características asociadas del tubo roto (su edad, diámetro, material de que esta construido, etc.), tipo del suelo sobre el que yace la tubería, factores climatológicos (como la temperatura), cargas de tráfico, etc. La base de datos de las tuberías corresponde a redes principales (300mm a 1200mm) y redes de distribución (50mm a 250mm) de diferentes materiales (hierro colado, hierro dúctil, PE de alta, media y baja densidad, hormigón armado, asbesto cemento, acero, hormigón armado pretensado, y PVC).

El trabajo presenta dos métodos para el análisis del riesgo de rotura de tubos en una red de suministro. Las técnicas utilizadas son los modelos de punteo y las redes bayesianas. Los modelos de punteo corresponden a técnicas de manejo de los datos mientras que las redes bayesianas representan una combinación de los modelos determinísticos con las técnicas de minería de datos.

La base de datos utilizada en el estudio está compuesta por los tubos de la red en el año 1995 (ubicación, diámetro, material, año de colocación, longitud, identificación, método de renovación, diámetro – material y fecha de renovación) así como una tabla de trabajos de reparación (identificación de la tubería, localización, causa de rotura, fecha de rotura, tipo de suelo, en o no uso) llevados

acabo entre diciembre 25 de 1928 y enero 17 de 1995, que contiene 3175 trabajos de reparación.

El modelo de punteo intenta enlazar casos que presentan un comportamiento similar. Esto se lleva acabo asignando un puntaje (un valor entre 0 y 100) a cada caso y agrupando casos en clases de puntajes similares; la bondad o efectividad del modelo se mide de acuerdo con el coeficiente de concordancia, el cual establece el nivel de concordancia entre la clasificación hecha por el modelo y el orden de los datos.

La orientación dada al problema consiste en series temporales; por tanto se toma una "instantánea" de las condiciones de la red para cada año repitiendo las propiedades de los tubos en cada instantánea para producir series de datos coherentes; el resultado es la ocurrencia (o no ocurrencia) de una rotura durante el año siguiente de la instantánea. Entre otros, los datos utilizados corresponden a la fecha de la instantánea, identificación del tubo, año de instalación, edad, roturas, localización, diámetro, longitud, material, tráfico sobre el tubo y, ocurrencia de por lo menos una rotura durante el siguiente año.

La estrategia de modelado consistió en crear dos modelos de puntos, uno que separe los casos de no rotura o extremado riesgo bajo de rotura de aquellos que producen incertidumbre, y un segundo modelo que maneje estos casos. Para la construcción del modelo se establece la serie de datos a utilizar, se realiza un pre-procesamiento de los datos, los operadores utilizados son binarios no lineales y la búsqueda del mejor modelo se realiza utilizando algoritmos genéticos.

La red bayesiana es alimentada con parámetros acerca del tubo y del suelo que lo rodea y de la presión dentro del tubo como "inputs". Como "output", el modelo produce un estimativo de la historia del tubo y el valor de las funciones de los tres estados límite; esfuerzo circunferencial, esfuerzo de corte y el estado límite de fatiga.

La técnica empleada para finalizar el modelo del proceso de rotura son los árboles de clasificación.

Los resultados obtenidos con el modelo de puntaje muestran que, aun cuando la edad está entre las variables utilizadas por el modelo, su poder de predicción es relativamente bajo. El mejor estimador de predicción para la primera partición (casos con baja probabilidad de rotura) es la longitud del tubo; a mayor longitud mayor número de roturas. Otro elemento de predicción utilizado fue el inicio o comienzo, es decir, el número de la casa donde el tubo inicia, la cual no parece ser una relación obvia, pero el modelo es más sensible a esta variable que a la edad del tubo.

El análisis de sensibilidad del modelo refinado (modelo de incertidumbres) toma como elemento de predicción principal el número de roturas en el modelo anterior; parece más probable que un tubo que se ha roto anteriormente se vuelva a romper. El segundo elemento de predicción es el momento en que se toma la "instantánea", el cual tiene que ver con el instante en el tiempo (año, mes, día, hora); este elemento de predicción es un indicador de que en algunos periodos existe más probabilidad de rotura que en otros.

Los modelos de puntaje dan un método para clasificar los tubos de acuerdo a su riesgo de rotura, lo cual es necesario para presentar un esquema de rehabilitación de la red. Este modelo es totalmente de manejo de datos; no existe ningún esquema para introducir conocimiento general acerca del sistema en cuestión en el algoritmo de inducción. La calidad del modelo depende de la calidad de los datos utilizados. La fuerza de los modelos de puntaje radica en revelar relaciones entre variables que no son obvias para la mente humana; por ejemplo, la relación entre el número de la casa en la que está ubicado un tubo y su riesgo de rotura.

Los modelos de redes bayesianas son más conocimiento que datos orientados, es decir se combinan los datos con el conocimiento acerca del problema. El coeficiente de concordancia es de 65.6% para el modelo con redes bayesianas. Como resultado se establece en el artículo que el modelo de puntos es más homogéneo en su desempeño, ya que clasifica un alto número de casos correctamente.

Coulibaly et al. (2001) hacen uso de cuatro tipos diferentes de configuraciones de redes neuronales para simular las fluctuaciones del nivel de agua en el acuífero Gondo de la región Sahél en Burkina Faso, y obtienen buenos resultados en la predicción del nivel mensual de fluctuaciones en el acuífero, con la ventaja adicional de la escasa cantidad de información con la que se dispone en el estudio, importante en zonas donde no se dispone de mucha instrumentación de medida. El modelo puede ser útil para mejorar el planeamiento del suministro de agua en áreas semiáridas.

Dandy y Engelhardt (2001) presentan el uso de la técnica de los algoritmos genéticos para encontrar un esquema óptimo para el reemplazo de tuberías en un sistema de abastecimiento de agua; adicionalmente, este artículo presenta una buena revisión bibliográfica acerca de diferentes modelos de rehabilitación de sistemas de abastecimiento a lo largo de las tres últimas décadas. El trabajo realizado tiene datos de un sistema real en una zona de Adelaida (Australia), en el cual se optimizan los costos (criterio económico) de reemplazar tuberías (no se tiene en cuenta la reparación de tuberías, por su alto costo); este análisis se realizó para tres modelos separados; uno para decidir si una tubería debe ser reemplazada o no (con esto se verifica la eficacia de los algoritmos genéticos para identificar tuberías a reemplazar); un segundo modelo teniendo en cuenta un plazo de 20 años de acuerdo con las limitaciones de presupuesto, con restricciones en previsión de posibles fondos en bloques de 5 años; y un tercer modelo que incluye el diámetro de la nueva tubería como variable de decisión. La estimación de la tasa de fallos se realiza por medio de una regresión, aplicando la metodología de mínimos cuadrados teniendo en cuenta la edad de la tubería. Uno de los inconvenientes del trabajo es que solo se tienen datos para los materiales de asbesto cemento y hierro fundido revestido con cemento, así que para los materiales de acero revestido con cemento, hierro dúctil revestido con cemento, y PVC la tasa de fallos se estimó a partir de los datos de los materiales para los cuales se cuenta con información. Del trabajo realizado se destaca el tercer modelo en el que se incluye el diámetro a reemplazar; aunque aumenta el tiempo requerido de entrenamiento del algoritmo genético, se logran

reducciones en diámetros que satisfacen las condiciones hidráulicas.

Revelli y Ridolfi (2002) presentan un trabajo en el cual describen la posibilidad de aplicar los conceptos de la teoría difusa (*fuzzy*) en el análisis de redes hidráulicas. Al presentarse incertidumbre en ciertas variables que afectan el funcionamiento real de una red de abastecimiento, tal como puede ser el factor de rugosidad utilizado en las ecuaciones de diseño y modelado de las redes con el paso de los años, o las demandas en los nudos, los autores proponen hacer uso de la lógica borrosa para tratar con estas variables. El método sugerido por los autores transforma las ecuaciones difusas en problemas de optimización no lineal. Aunque tal como se menciona en el texto, no se ha probado el tiempo computacional requerido, los resultados que se obtienen parecen ser bastante prometedores, cuando la información de partida con la que se cuenta es escasa y no permite realizar un estudio estadístico del comportamiento de las variables.

Los autores realizan un primer ejemplo en una red con dos mallas, cinco tuberías y cuatro nodos, en el cual se cuenta con valores conocidos de longitudes, diámetros, demandas en los nudos, y la carga piezométrica de entrada, el material es acero para el cual se sabe experimentalmente que el coeficiente de rugosidad *c* varía entre 55 y 65 $m^{1/3}s^{-1}$, con un valor probable de 60 $m^{1/3}s^{-1}$ (para entre 10 y 20 años de operación). Para efectos de comparación se calculan las variables incógnitas (caudales circulantes y presiones en los nudos), para el valor más probable de 60 $m^{1/3}s^{-1}$ de coeficiente de rugosidad. Para el caso de los caudales circulantes, haciendo uso de funciones de pertenencia triangulares, se obtienen diferencias de hasta el 30% entre el valor probable de *c* (pico) y los mínimos; para las presiones es de un 10% que equivale a entre 6 y 7 metros. Esto muestra, como, pequeñas incertidumbres en los datos iniciales son significativamente amplificados en las variables no conocidas, debido a, la no-linealidad y combinación de incertidumbres.

Si la tubería ha sido utilizada por muchos años, se pueden presentar incrustaciones y diferencias entre tuberías, con lo cual la función de pertenencia está entre 55 y 65 $m^{1/3}s^{-1}$, aún cuando para los picos está entre 50 y 70 $m^{1/3}s^{-1}$;

para este caso se obtienen funciones trapezoidales, con las cuales se incrementa la diferencia en los caudales circulantes en un 40% y, para las presiones, entre 7 y 8 metros. Si se agregan incertidumbres en las demandas se tienen diferencias de hasta el 50% para los caudales circulantes, y en las presiones de hasta 11 metros.

Con una red un poco más compleja (4 mallas, 12 líneas, 9 nodos), los autores obtienen incertidumbres en los valores de caudales circulantes de entre el 15 y el 20% para caudales circulantes altos, mientras que para pequeños caudales hasta un 200%. En cuanto a las presiones, entre 3 y 4 metros. Con este trabajo se muestra la variabilidad de los resultados obtenidos en el modelado de la red debido a las incertidumbres en las variables consideradas. Los autores proponen su utilización conjunta con modelos de calidad de agua, debido a la exactitud que estos requieren en cuanto a los caudales circulantes y las presiones.

Karpenko y Sepehri (2002) presentan un esquema para la detección e identificación de fallos en el mecanismo de accionamiento de una válvula de control neumático, basado en una red neuronal multicapa. Los resultados obtenidos muestran que la red entrenada es capaz de detectar e identificar presiones incorrectas de suministro, fugas en el diafragma, y fallos en el mecanismo de accionamiento. La función de activación de las capas ocultas es una tangente hiperbólica.

Los datos de entrenamiento corresponden a parámetros de funcionamiento obtenidos experimentalmente para predecir la condición normal de funcionamiento de la válvula y/o doce posibles condiciones de fallo. No se escogen ejemplos donde dos o más fallos ocurran simultáneamente. Los pesos de la red se actualizaron por retropropagación, haciendo uso del algoritmo del gradiente descendiente con momento. El número ideal de neuronas en la capa oculta se determinó por la observación del comienzo de sobre-ajuste, que a su vez fue determinado cuando la suma de una neurona en la capa oculta no lleva a un incremento en la exactitud de la red. Se entrenaron cien redes para cada tamaño de capa oculta para minimizar los efectos de la asignación aleatoria de los pesos

iniciales. Los resultados corresponden al promedio de estos entrenamientos.

Bessler et al. (2003) desarrollan políticas de operación en embalses siguiendo dos metodologías: para un sistema de embalse, y para varios embalses de agua. Se aplican metodologías de modelado de redes, optimización con programación lineal, extracción de reglas con técnicas de minería de datos (árboles de decisión), y regresión lineal (estadística clásica), predicción de caudales, simulación y análisis. El modelado realizado con la técnica de minería de datos da muy buenos resultados en el caso del embalse único; para el sistema de múltiples embalses los resultados obtenidos no son malos pero se presentan dificultades en cuanto a la poca diversidad de información, lo cual provoca pocas clasificaciones para algunas clases. La bondad de los modelos de C5.0 (árboles de clasificación) se estima basándose en matrices de confusión.

Bhattacharya et al. **(2003)** aplican técnicas de aprendizaje automático (redes neuronales artificiales y aprendizaje por refuerzo) combinándolas con un sistema de soporte a la decisión basado en programación lineal (*Aquarius*), para realizar un óptimo control dinámico en tiempo real de recursos hídricos; no se obtienen muy buenos resultados al aplicar las metodologías por separado. Construyendo un sistema híbrido de redes neuronales y aprendizaje por refuerzo obtienen mejores resultados. El modelo fue utilizado en Holanda, ya que sus condiciones particulares de tener ciudades por debajo del nivel del mar y la creación de *pólder*, les requiere de tener sistemas de control y decisión en escalas temporales cortas (15 minutos).

La tarea consistía en minimizar el tiempo de respuesta a una situación de posible inundación del *pólder*. Las variables de entrada elegidas fueron los niveles de agua y la precipitación. La red neuronal tiene que clasificar el estado del bombeo, es decir, si se activa o apaga. Los resultados obtenidos no fueron satisfactorios. Por tanto, se contempló un sistema de aprendizaje basado en aprendizaje por refuerzo para desarrollar un algoritmo híbrido que redujera el error de los modelos de redes neuronales, eligiendo una generalización del aprendizaje *Q* a través de una red neuronal, donde cada vector de entrada en el

conjunto de entrenamiento de la red neuronal representa un par de estado-acción para el aprendizaje *Q*.

El aprendizaje por refuerzo es una aproximación de aprendizaje automático para modelar una estrategia de control óptima. Los elementos del aprendizaje por refuerzo son: un agente, un ambiente o conjunto de estados que describan el entorno, un conjunto de acciones que el agente puede realizar para modificar su entorno, y una función de refuerzo que indica al agente el resultado obtenido al realizar la acción sobre el entorno. Para aprender la relación entre la política de control óptima y el estado se hace uso de la técnica del aprendizaje *Q* (*Q-learning*); para cada estado hay un valor *Q* asociado con cada posible acción.

Van Zyl et al. (2004) utilizan el método "*hillclimber Hooke and Jeeves*" para mejorar la búsqueda con algoritmos genéticos del óptimo dentro de una región de búsqueda, ya que mientras éstos últimos son eficientes encontrando la región de solución óptima son menos eficientes para encontrar el punto óptimo dentro de ésta región; por esta razón y pensando en una situación de emergencia en un abastecimiento de agua, el método híbrido fue desarrollado con el énfasis principal en la velocidad de convergencia más que en la fiabilidad. Este método híbrido fue probado en un sistema de distribución hipotético, y en un gran sistema de distribución existente en el Reino Unido. Se concluye que este método híbrido tiene un mejor desempeño que un algoritmo genético puro, tanto en velocidad de convergencia como en la calidad de la solución encontrada.

Uno de los inconvenientes de los algoritmos genéticos es que, a pesar de su gran poder para encontrar óptimos, el tiempo computacional requerido es bastante largo, lo cual no los hace tan atractivos para problemáticas de gestión y operación de la red de abastecimiento de agua. Los algoritmos de *ascenso de colina* (hillclimber) están basados en optimización local. Son llamados también estrategia irrevocable porque no permiten regresar a otra alternativa, es decir, es un método no exhaustivo, ya que no explora todo el espacio de estados; como máximo sólo encuentra una solución. Su funcionamiento es de la siguiente forma:

- Utilizan una técnica de mejoramiento iterativo.

- Inician a partir de un punto (punto actual) en el espacio de búsqueda
- En cada iteración, un nuevo punto es seleccionado de la vecindad del punto actual.
- Si el nuevo punto es mejor, se transforma en punto actual, si no otro punto vecino es seleccionado y evaluado
- El método termina cuándo no hay mejorías, o cuándo se alcanza un número predefinido de iteraciones.
- Del procedimiento de prueba, existe una realimentación que ayuda al generador a decidirse por cuál dirección debe moverse en el espacio de búsqueda.
- En estos procesos se abandona la búsqueda si no existe un estado alternativo razonable al que se pueda mover.
- Los algoritmos de ascenso de colina son típicamente locales, ya que deciden qué hacer mirando únicamente a las consecuencias inmediatas de sus opciones, lo cual los hace fuertes donde los algoritmos genéticos son débiles.
- Puede que nunca lleguen a encontrar una solución, si son atrapados en estados que no son el objetivo, desde donde no se puede hallar mejores estados.

Los autores comparan la optimización realizada aplicando algoritmos genéticos y el método híbrido del consumo energético sobre un sistema de abastecimiento de agua. Con la utilización del método híbrido reducen tanto los costos de operación del sistema, como el tiempo computacional requerido para realizar la optimización. La variable a optimizar fue el nivel de los tanques de almacenamiento a lo largo del día, que permite minimizar los costos anuales de bombeo en todo el sistema, teniendo en cuenta los periodos de tarifas energéticas diarios.

Dandy y Engelhardt (2006) presentan un algoritmo basado en un algoritmo genético multi-objetivo para desarrollar curvas de equilibrio entre el costo económico y la fiabilidad en esquemas de reemplazo de tuberías de suministro de agua. Los costos económicos se expresan en términos del valor presente de reemplazar la tubería más el esperado por los daños y reparaciones asociados a la rotura de la tubería. La fiabilidad se mide como el número esperado

de interrupciones por año. La bondad del modelo expresado por los autores radica en que se realiza sobre un caso real y que se identifica un programa de reemplazo en un horizonte de 20 años restringido por los fondos disponibles. En el estudio se utilizó un modelo de tasa de roturas basado en la edad de la tubería para cada material y diámetro.

La zona de estudio corresponde a un suburbio de la zona norte de Adelaida (Australia), que abastece principalmente zona urbana; aunque hay un par de zonas industriales y comerciales. Se tienen 488 secciones separadas de tubería de diámetros superiores a 150mm. El historial de roturas es de 10 años, por lo cual se aplicó un análisis de regresión para ampliar el historial. El número de usuarios que podría verse afectados por una interrupción se asumió como función del uso del suelo. Los resultados de la optimización se presentan en función del costo mínimo de reemplazo de una tubería por otra de igual diámetro, y el número mínimo esperado de interrupciones a los clientes (criterio de fiabilidad establecido), bajo dos horizontes de planeamiento, uno en el que el reemplazo se realiza en el instante inicial, y otro en el que el reemplazo se puede realizar en cualquier momento dentro de un periodo establecido limitado por un presupuesto.

Dawsey et al. (2006) proponen un método de redes bayesianas para expresar relaciones entre eventos y observaciones de contaminación, en una red de abastecimiento hipotética. El objetivo del trabajo es el poder identificar falsos positivos de contaminación, útil en temas de seguridad y planes de acción de respuestas en tiempo real. La metodología propuesta se implementa en un caso hipotético de estudio, con una red que consta de 235 nodos y 261 tramos de tubería, entre 50.8 mm y 304.8 mm de diámetro. Las presiones para el modelo se configuran entre 276 kPa y 552 KPa, y una velocidad no menor de 1.53 m/s. Se identificaron tres localizaciones de sensores de contaminantes, basadas en una inspección cualitativa de los patrones de flujo en el modelo con simulaciones en periodo extendido, pero sin ningún tipo de optimización en cuanto a su ubicación. Como propósito de ilustración, la metodología propuesta por los autores a partir de la red Bayesiana consta de tres posibles escenarios: una detección de falso positivo, y dos de positivos verdaderos. Para estimar las probabilidades previas de

detección en cada uno de los sensores se realizaron simulaciones repetidas con Epanet, basándose en una única inyección. La concentración de soluto es medida durante un periodo de simulación de 36 horas para cada combinación de localización y curso de masa, para garantizar que se capture el pico de la concentración en cada sensor. Una detección positiva es asumida arbitrariamente cuando se exceden 100 mg/L.

El principal inconveniente que tiene este trabajo es que es sobre un modelo no real; además los puntos de posible contaminación se toman como uno solo, dejando la red vulnerable a diferentes ataques.

El-Baroudy y Simonovic (2003,2004,2006) utilizan la técnica de la lógica difusa (Fuzzy Logic) para evaluar el funcionamiento de una red de abastecimiento basándose en la estimación de los parámetros de vulnerabilidad (severidad de las consecuencias de fallo), fiabilidad (qué tan probable es que falle el sistema), robustez (habilidad del sistema para adaptarse a un amplio rango de condiciones futuras de carga), y resilencia (qué tan rápido se recupera) de la red. El cálculo de las mediciones de rendimiento del sistema depende de la identificación exacta del estado insatisfactorio del sistema. Es difícil llegar a una definición precisa del evento de fallo, no obstante, por la incertidumbre en la determinación del suministro, demanda, y el umbral de desempeño insatisfactorio aceptado. El análisis de fiabilidad difuso cuantifica esta incertidumbre a través del uso de funciones de pertenencia apropiadas para describir el estado de seguridad del sistema difuso, y de los eventos de fallo difusos.

En los análisis de fiabilidad de los sistemas, se hace uso de la *carga* y la *resistencia* como conceptos fundamentales para definir el riesgo de fallo del sistema, que en este caso fueron reemplazados por la *demanda* y el *suministro* respectivamente. En cuanto a la función de pertenencia, se presentan varias posibilidades en cuanto a la forma que éstas pueden adoptar; sin embargo para el problema aquí manejado, los componentes del sistema tienen una capacidad máxima y una mínima que no pueden ser excedidas; por tanto, cualquier forma de función de pertenencia debe tener dos límites extremos con valores de

pertenencia cero. Las formas triangular y trapezoidal corresponden a las funciones de pertenencia más simples que cumplen estos requisitos. En el trabajo presentado por los autores, estos formulan las componentes de las funciones de pertenencia en términos del margen de seguridad *difuso,* definido como la diferencia entre el suministro *difuso* y la demanda *difusa.*

Para totalizar la representación múltiple de los componentes del sistema de abastecimiento que transportan agua con aquellos que no lo hacen, se optó por la representación integrada de componentes del sistema, cada una con diferentes relaciones de fallo; es decir, si se tiene una representación de dos componentes en serie, si una componente falla el otro también; por el contrario, si dos componentes están conectados en paralelo, el fallo de uno no conduce al fallo del otro. Esto facilita los cálculos de las mediciones del funcionamiento *difuso* del sistema, haciendo uso de funciones de pertenencia normalizados que tienen un valor máximo de uno. Estas funciones de pertenencia son agregadas para calcular las medidas de funcionamiento del sistema difuso.

La función de pertenencia de fallo del sistema es utilizada para calcular el índice difuso de resilencia. Agregando estas funciones de cada uno de los componentes del sistema, se obtiene la función de pertenencia para todo el sistema, que corresponde a la función de pertenencia máxima de recuperación del sistema. La metodología se utilizó para estudiar el comportamiento del abastecimiento de la ciudad de Londres (Ontario) en Canadá, con dos componentes principales de estudio; el sistema de suministro de agua del lago Hurón, y el sistema de suministro de agua del área de Elgin. Los resultados obtenidos muestran que la medición fiabilidad-vulnerabilidad difusa combinada para el sistema del lago Hurón, es mayor que la del sistema del área de Elgin tanto para la forma triangular como trapezoidal en 10 veces, aunque se apreció que la forma de la función de pertenencia es bastante sensible. La medida de la resilencia difusa para el área de Elgin, es cuatro veces mayor que la del lago Hurón. Por tanto, se encontró que el primer sistema es más fiable y robusto que el segundo.

La metodología propuesta por los autores, además de permitir identificar componentes críticos dentro del sistema, sirve como soporte para la determinación de fallos en un sistema de abastecimiento de agua, permitiendo la incorporación de todos sus componentes ya sea de forma paralela o en serie. El inconveniente que presenta en cuanto a su aplicabilidad general, es la disponibilidad de la información necesaria para la construcción de las funciones de pertenencia.

Zhang et al. (2007) haciendo uso de la técnica de reconocimiento artificial inmune (inspirado en el sistema inmune humano), exploran la aplicación de su uso para la extracción de reglas de operación de un embalse en un sistema hídrico. Los sistemas inmunes artificiales, implementan una técnica de aprendizaje inspirada en el sistema inmune humano, que corresponde a su mecanismo natural de defensa, aprendiendo acerca de sustancias extrañas y protegiendo el cuerpo de invasiones de microorganismos, bacterias, o virus.

El sistema inmune humano es altamente distribuido, altamente adaptativo y auto organizativo por naturaleza, manteniendo una memoria de sustancias extrañas (antígenos), y tiene la habilidad de aprender continuamente de nuevos antígenos. Una vez el cuerpo sufre un ataque de un antígeno, la respuesta inmune del sistema será provocada, y las células inmunes del cuerpo (linfocitos B) producidas por la médula ósea, serán estimuladas para secretar anticuerpos, los cuales sufren evoluciones continuas, y pueden identificar y neutralizar o eliminar los antígenos.

El trabajo propuesto por los autores en la operación de un embalse, se basa en el sistema de reconocimiento inmune artificial (AIRS), en el cual los genes representativos del antígeno y el anticuerpo del reconocimiento artificial B (ARB), son representados como la formulación de la regla *"si (N,S,I,D,P),* entonces *C"*, donde *N* corresponde al número de periodo de operación (mes de 1 a 12), *S* es el almacenamiento de agua en el embalse, *I* las entradas de agua al embalse, *D* la demanda de agua, *P* la condición hidrológica anual (húmedo, por encima de la media, promedio, por debajo de la media, seco, extremadamente seco), y *C* la

clase del patrón de operación. Básicamente, el algoritmo está compuesto de la presentación de los antígenos para el entrenamiento; luego se generan las células de memoria y ARBs; luego los anticuerpos compiten por adquirir los recursos para saber si sobreviven; los anticuerpos de la misma clase y valor de simulación alta como el antígeno son seleccionados para entrar en la célula de memoria; una vez un antígeno es repartido por el AIRS, el procedimiento continua con el próximo antígeno hasta que todos entren al sistema. Las clasificaciones obtenidas a partir de este modelo fueron comparadas con otro que tiene en cuenta la entropía para la estimación de las condiciones hidrológicas anuales, y una metodología de redes de base radial para extraer las reglas de operación del embalse.

Entre las conclusiones dadas por los autores, señalan que aunque los resultados obtenidos con el AIRS para la extracción de reglas de operación del embalse son bastante satisfactorios, se obtienen redundancias de reglas en el sistema de memoria con el aumento de células de memoria, la escasez de consideración de la importancia de los diferentas rasgos o características entre atributos, y la carencia de representaciones asociativas entre los rasgos de los atributos y los patrones de operación o entre las características. Ellos proponen regular los pesos de las características de las reglas de operación del AIRS, haciendo uso de una estimulación conjunta entre linfocitos T y B, además de unos mecanismos de auto adaptación de los Arbs, para generalizar la capacidad del AIRS, y la mejora de la interpretabilidad de las reglas.

Díaz et al. (2008) aplican la técnica de búsqueda heurística basada en la población *(particle swarm optimizatión - PSO)*, para un problema de optimización de *clusters*, para la determinación de la influencia del material de la tubería y de las longitudes de tramos en una red de abastecimiento, en cuanto a daños ocurridos en la red. Los datos corresponden a la red descrita como aplicación práctica de esta tesis para el municipio de Calarcá. Se aplicó una técnica derivada de PSO, que considera tanto variables discretas como continuas. Por otra parte, uno de los principales inconvenientes asociados a la PSO, radica en el hecho de que es difícil mantener niveles adecuados en la diversidad de la población, y balancear las búsquedas locales y globales; con esta formulación se es capaz de

encontrar la solución óptima, de forma más eficiente y con un esfuerzo computacional menor, debido a la riqueza de diversidad en la población introducida (Izquierdo *et al.,* 2009a, Montalvo *et al.,* 2008b).

Del capítulo anterior, numeral II.3.3.3 se tienen las siguientes expresiones:

$$V_i^{n+1} = wV_i^n + c_1 r_1 \times \left(pbest_i - X_i^n \right) + c_2 r_2 \times \left(gbest - x_i^n \right) \quad (1)$$

$$X_i^{n+1} = X_i^n + V_i^{n+1} \quad (2)$$

$$V_i^{n+1} = wV_i^n + c_1 r_1 \times \left(pbest_i - X_i^n \right) + c_2 r_2 \times \left(lbest_i - X_i^n \right) \quad (3)$$

$$w = w_{\max} - \frac{\left(w_{\max} - w_{\min} \right) \times n}{iter_{\max}} \quad (4)$$

En este caso, el peso de inercia decreciente se estimó de acuerdo con la siguiente ecuación (Jin *et. al., 2007)*, la cual mejora el desempeño de la PSO:

$$w = 0.5 + \frac{1}{2\left(\ln\left(k \right) + 1 \right)}, \quad (5)$$

Adicionalmente, la formulación inicial de Kennedy y Eberhart (1995) fue dotada con la habilidad de ajustar sus parámetros para optimizar la forma en que las partículas se acomodan a las mejores condiciones. Para conseguir esto, los tres parámetros de la formulación original, se consideran como tres nuevas variables que son incorporadas a los vectores de posición X_i (Montalvo *et al., 2010)*. En general si *D* corresponde a la dimensión del problema, y *P* es el número de parámetros auto adaptativos, el nuevo vector de posición para la partícula i será:

$$X_i = \left(x_{i1}, ..., x_{iD}, x_{iD+1}, ..., x_{iD+P} \right) \quad (6)$$

Es claro que las primeras *D* variables corresponden al vector de posición real de la partícula en el espacio de búsqueda, mientras que las últimas *P* dan cuenta de sus parámetros; éstas nuevas variables no forman parte de la función de idoneidad, pero son operadas haciendo uso del mismo paradigma de

aprendizaje utilizado en la PSO.

Igualmente, la velocidad y la mejor posición hasta el momento para la partícula i, incrementan su dimensión, con el correspondiente significado:

$$V_i = \left(v_{i1},...,v_{iD},v_{iD+1},...,v_{iD+P}\right) \quad (7)$$

y,

$$Y_i = \left(y_{i1},...,y_{iD},y_{iD+1},...,y_{iD+P}\right). \quad (8)$$

Haciendo uso de las expresiones 1 o 9 y 2, cada partícula es dotada con la habilidad de ajustar sus parámetros teniendo en cuenta los que tenía en su mejor posición anterior, así como los parámetros del líder lo cual facilita su movimiento hacia una posición privilegiada. En consecuencia, las partículas hacen uso de su conocimiento del pensamiento individual y de la cooperación social para mejorar sus posiciones, y la forma como optimizan estas posiciones acomodándose a las condiciones (las propias y las del líder).

Por otra parte, para abordar el problema de las variables discretas, el algoritmo toma las partes enteras del vector de velocidad de las componentes discretas tomadas en cuenta; por lo tanto, las nuevas componentes de velocidad discretas V_i son enteras y, por consiguiente, el nuevo vector de componentes discretas será también entero (puesto que los vectores iniciales fueron generados con valores enteros). De esta forma, la actualización de velocidades para variables discretas se realiza por (Izquierdo *et al.,* 2009c):

$$V_i = \text{int}\left(\left(wV_i + c_1r_1\right)\times\left(pbest_i - X_i\right) + c_2r_2\times\left(gbest - X_i\right)\right) \quad (9)$$

donde int denota que solo se toma la parte entera del resultado.

Uno de los principales inconvenientes de la PSO es la dificultad en mantener niveles aceptables de diversidad de población, mientras se realiza el balance de las búsquedas locales y globales; en Montalvo *et al.,* (2008b) se muestra cómo son detectadas frecuentes colisiones de pájaros en el espacio de búsqueda, especialmente con el líder. Esto hace que el tamaño efectivo de la población decrezca y, por consiguiente, la efectividad del algoritmo. En Izquierdo

et al. (2009a) se introduce una derivada de la PSO en la cual unos pocos de los mejores pájaros son seleccionados para controlar la colisión y, si esta ocurre, los pájaros que colisionan son regenerados aleatoriamente. Con esto se evita la generación de convergencias prematuras y la clonación de poblaciones, con lo cual se incrementa la diversidad de la población, así como la mejora de las características de convergencia del algoritmo y la calidad de las soluciones finales.

A continuación se presenta el pseudocódigo del algoritmo modificado, siendo *k* el número de iteraciones:

- $k = 0$
- Generar una población aleatoria de *M* partículas: $\{X_i(k)\}_{i=1}^{M}$ de acuerdo con (6)
- Se evalúa la idoneidad de las partículas (sólo las primeras *D* variables entran en la función de aptitud)
- Se registra la mejor localización local $\{Y_i(k)\}_{i=1}^{M}$; así como los correspondientes parámetros
- Se registra la mejor localización global, $Y^*(k)$, y la lista de las *m* partículas mejores para controlar las colisiones (incluyendo sus correspondientes parámetros)
- Mientras no exista una condición de parada se ejecuta lo siguiente:
  - Determinar el parámetro de inercia w, de acuerdo con (5)
  - Iniciar el ciclo desde 1 hasta el número de partículas *M*
    - Inicio
      - Calcular la nueva velocidad para la partícula i de acuerdo con (1), y tomar su parte entera (para la optimización discreta) para las primeras *D* variables, de acuerdo a (9)
      - Actualizar la posición $X_i(t+1)$ de la partícula i, de acuerdo a (2)
      - Calcular la función de aptitud para la partícula i
      - Si la partícula i tiene mejor valor de aptitud que el valor de aptitud de la mejor partícula histórica, entonces dejar la partícula i como la nueva mejor partícula y actualizar la lista.
      - Si la partícula i no es una de las mejores partículas, pero coincide con una de las m mejores partículas, entonces regenerar la particular i aleatoriamente (incluyendo sus parámetros)
    - Fin
  - k = k + 1

- Mostrar la solución dada para la mejor partícula.

Durante las ejecuciones del algoritmo se obtuvieron resultados idénticos. Se utilizó un tamaño de población de 30 partículas. Las velocidades máxima y mínima se establecieron como:

- Máxima velocidad para las variables discretas = 50% del rango de la variable.
- Velocidad mínima = - Velocidad máxima.

La condición de parada del algoritmo es si después de 20 iteraciones no se obtienen mejoras en la solución. Los resultados se obtienen para un promedio de 40 iteraciones. Los *clusters* se realizaron utilizando la PSO y considerando diferentes posibilidades (2, 3 y 4 *clusters*). El espacio de búsqueda se estableció multidimensional.

Para realizar el clúster se hizo uso del algoritmo CLARA (Clustering Large Applications), que es una combinación de algoritmos de muestreo y el algoritmo PAM (Partitioning Around Medoids). Con este algoritmo, en vez de encontrar los medoides[3], se representa una muestra del conjunto de datos y se utiliza el algoritmo PAM, para seleccionar un conjunto óptimo de medoides de la muestra. Los problemas de sesgo son tratados repitiendo el proceso de muestreo y clustering múltiples veces y seleccionando el mejor conjunto de medoides como clustering final.

La calidad de los medoides resultantes es medida por la disimilaridad promedio entre cada objeto en todo el conjunto de datos *D* y el medoide de su clúster, definida por la siguiente función de coste:

[3] Los medoides, son objetos representativos de un conjunto de datos o un clúster con una serie de datos cuya disimilaridad promedio para todos los objetos en el cluster es mínima. Los medoides son similares en concepto a las medias y centroides, pero los medoides son siempre miembros de un conjunto de datos. Son comúnmente utilizados con datos donde la media o el centroide no pueden ser determinados.

$$Cost(M,D)=\frac{\sum_{i=1}^{n} dis\left(O_i, rep\left(M,O_i\right)\right)}{n},$$

donde *M* corresponde a un conjunto de medoides, $dis\left(O_i,O_j\right)$ es la disimilaridad entre los objetos O_i y O_j, y $rep\left(M,O_i\right)$ da un medoide en *M* el cual es cercano a O_i.

Se obtienen como resultados *clusters* que dividen los daños de acuerdo con los materiales de las tuberías de la red, principalmente dos; el primero representado principalmente por tubos de hormigón armado y PVC; mientras que para el segundo la mayoría de tubos corresponden a PVC. Las longitudes de las tuberías se establecieron como cortas, medias y largas, basándose en los valores de la base de datos; para el primer grupo la mayoría de las tuberías son largas, teniendo las cortas menor incidencia; en el segundo grupo la cantidad de tubos medios y largos fueron siempre los mismos, siendo los tubos cortos más significativos. En cuanto a la identificación de problemas, se concentran en tubos de tamaño medio hechos de hormigón armado, y en tamaños grandes de PVC. Igualmente, se identifican problemas en tuberías cortas de PVC. En cualesquiera del resto de materiales o en tubos de tamaño medio no se tienen influencias relevantes sobre los problemas detectados.

La metodología aquí propuesta no presentó mejoría en cuanto a los resultados expuestos en los capítulos de resultados de esta tesis respecto a la clasificación de los tipos de daños; no obstante ha mostrado su eficacia como optimizador en diversidad de problemáticas en la literatura (Montalvo *et al.,* 2008a, Izquierdo *et al.*, 2009c). Sin embargo, es de resaltar las mejoras que se le han hecho al algoritmo para que trabaje tanto con variables discretas y continuas, y al enriquecimiento de la diversidad introducido, además del manejo auto adaptativo de los parámetros. Esto permite que en general, la herramienta de optimización sea útil, ya que se pueden agregar más términos a la función de ajuste sin otorgar más complejidad conceptual al problema.

Finalmente, es de destacar, que el algoritmo de PSO otorga ventajas en problemas de optimización en la industria del agua, debido a la cualidad de manejar formulaciones multi-objetivo con soluciones buenas en cortos espacios de tiempo. Todo esto gracias a la habilidad de las partículas para decidir como grupo, cómo moverse dentro del espacio de búsqueda y, cambiar su comportamiento durante el proceso de búsqueda.

***Alonso et al.* (2009)** establecen un modelo de optimización para decidir la cantidad de tubería en metros lineales a rehabilitar, basándose en una restricción presupuestaria, y en la valoración de factores de influencia, los costos de la rehabilitación y los beneficios obtenidos con la misma. La ventaja de la metodología propuesta es que al existir la restricción presupuestaria, se optimizan tramos a rehabilitar sin necesidad de actuar sobre la red en general.

***Herrera et al.* (2009b)** haciendo uso de las *máquinas de soporte vectorial* para clasificar anomalías y un algoritmo *kernel* para entender el comportamiento de las demandas de agua, prueban una metodología para detectar anomalías en el comportamiento de las demandas de agua.

Christodoulou et al. (2009) realizan un estudio de un sistema de soporte a la decisión, basado en una técnica neuro-difusa para un análisis multifactorial del riesgo de fallos y el manejo de recursos, relacionados con las redes de distribución de agua. El diseño sostenible, las operaciones, y manejo de las redes de distribución de agua, requieren una evaluación y cuantificación (de ser posible) de la degradación estructural de las tuberías, así como una evaluación de los factores socioeconómicos relacionados con las redes de tuberías (tal como el número de usuarios servidos, la proximidad de tuberías a zonas residenciales o comerciales). La inspección directa y la evaluación de muchos de estos factores son dispendiosas en tiempo y costosa. Por tanto, un medio de inferir los valores de los factores a través de reglas expertas, muestreo de datos, y sistemas neuro-difusos, parece ser una forma efectiva para deducir conclusiones sobre la importancia comparativa de los factores, su severidad en el tiempo, y sus posibles resultados en términos del riesgo de falla. Los sistemas neuro-difusos pueden

procesar y enlazar no solo datos históricos y patrones de datos resultantes (a través del componente de red neuronal del sistema), sino también el conocimiento de expertos en el campo particular (a través del componente de lógica difusa del sistema). La base de datos utilizada para este estudio corresponde a 20 años de series de datos de la ciudad de Nueva York (USA), y 5 años de series de datos de la ciudad de Limassol (Chipre).

Los atributos tenidos en cuenta para la elaboración del modelo fueron *roturas previas, material de la tubería, longitud, diámetro, carga de tráfico, proximidad de la red a una autopista, proximidad de la red al metro y proximidad de la red a intersecciones de vías* para el caso de Nueva York, en el caso de Limassol se tiene los mismos atributos con excepción del metro, autopista e intersecciones. En el análisis de resultados se presenta una primera etapa de utilización de las redes neuronales, las cuales son utilizadas para determinar el o los atributos más representativos para el análisis de riesgo de rotura, obteniéndose que las roturas previas presentan un factor de riesgo dominante. También se tienen factores dominantes para el tipo de material, el diámetro y la longitud. En cuanto al conocimiento deducido (reglas difusas), es obtenido examinando diferentes combinaciones de entradas y salidas registradas (vectores de entrenamiento), y por la combinación de patrones en comportamiento con reglas desarrolladas por expertos. A cada factor de riesgo se le asigna una función de pertenencia (triangular o trapezoidal) asociada a una caracterización lingüística dentro de un rango de valores (pequeño, medio, grande). De acuerdo a esto, se obtienen reglas del tipo "*para un número de roturas previas pequeño, material = 1 (hierro dúctil), diámetro grande y longitud grande, el ciclo estimado de vida es de 21,850 días (alrededor de 59 años).*

Guo y Ding (2009) destacan el hecho de que la aplicación de la minería de datos es poco frecuente en las redes de abastecimiento de agua, así como que su aplicación tendrá efectos positivos en la resolución de problemas existentes. Esto los autores lo plasman en dos aspectos: la carencia general de alta predicción para problemas posibles, y la posibilidad de obtener información comprensible no

numérica para la toma de decisiones.

Huang y McBean (2009) hacen uso de un modelo de minería de datos con el fin de identificar las posibles localizaciones de intrusión de contaminantes en una red de abastecimiento de agua, tal como se presenta en el capítulo anterior de fiabilidad del sistema. El problema que plantea este artículo se basa fundamentalmente en: la instrumentación de la red, la aplicación de la metodología y la valoración de la localización del punto de intrusión. La red de estudio es inventada (inconveniente en cuanto a su aplicación real) y la red de sensores corresponde a menos del 2% del total de nodos en la red. El modelo de minería de datos hace uso de un término *difuso* definido como, *estado del dato,* que tiene dos valores; no detectado (ND) que corresponde a la mínima concentración de contaminación detectable, con lo cual un valor superior a ND se considera como un estado *anormal*, o contaminación detectada, mientras que una concentración menor a ND o 0 se considera como un estado *normal,* o contaminación no detectada. Para valorar el punto de intrusión hacen uso del método de máxima verosimilitud, y a la vista de los resultados estadísticos obtenidos no queda claro el poder definir un único punto de intrusión.

Herrera et al. (2010) presentan una variedad de alternativas de aprendizaje automático para efectuar la predicción de demanda de agua en un abastecimiento del sur de España. Los datos corresponden a mediciones horarias de caudales durante los meses de enero a abril de 2005. Adicionalmente, se cuenta con información climatológica de temperaturas, velocidad del viento, lluvia y presión atmosférica. Los algoritmos utilizados para la predicción son: redes neuronales artificiales *(ANN)*, *projection pursuit regression (PPR)*, regresiones adaptativas multivariadas por splines *(MARS)*, regresiones de soporte vectorial *(SVR)* y, *random forests*. Los mejores resultados se obtienen con *SVR;* se utilizó un experimento de simulación de Monte Carlo para manejar el desempeño de las series temporales garantizando el no sesgo de las estimaciones.

III.3. Contribuciones propias

Del tema de pre y pos procesamiento de la información se realizó una contribución (Díaz *et al.*, 2010), a presentar en el *International Congress on Environmental Modelling and Software* (5 a 8 de julio de 2010 en Ottawa Ontario - Canadá), basado en parte del trabajo realizado en esta tesis. Adicionalmente, durante el trascurso de la investigación se han presentado una serie de publicaciones, principalmente en congresos, con las cuales se ha querido siempre mostrar la aplicabilidad de las herramientas de minería de datos en la gestión de los sistemas de abastecimiento de agua, tal como se puede ver en Díaz *et al.* (2002a, 2002b, 2002c, 2004a, 2004b, 2004c, 2004d, 2005, 2007) e Izquierdo *et al.* (2008a). La temática de estas publicaciones ha estado fundamentada en los pasos del proceso de descubrimiento de conocimiento en bases de datos; el paso de minería de datos más utilizado ha sido el aprendizaje automático.

III.4. Notas Finales

Aunque se encuentran publicaciones relacionadas con técnicas de manejo de datos aplicadas a los sistemas de abastecimiento de agua, tal como concluyen Gou y Ding (2009), se aprecia una carencia en la predicción de posibles problemas y en la posibilidad de obtener información comprensible no numérica para la toma de decisiones. Adicionalmente, se ha enfocado el análisis desde una perspectiva no conjunta, abordando problemáticas individuales como la predicción de demandas o la optimización de diámetros en las redes, pero no como un todo en cuanto a la gestión integral de los sistemas de abastecimiento de agua.

No obstante, es interesante la variedad de técnicas que han sido utilizadas para resolver las diferentes problemáticas, lo cual permite confirmar las posibilidades que brindan estas herramientas para conformar un sistema de gestión integral del abastecimiento. No podemos abstraernos en este punto de la problemática en la consecución de la información para desarrollar el sistema de gestión integral y, aquí, si no se cuenta con la colaboración y el interés de los

encargados de gestionar el servicio, poco se puede desarrollar desde fuera, porque el mayor interés planteado se basa en poder contar con información real y fiable, para poder desarrollar la metodología del descubrimiento de conocimiento en bases de datos.

Capítulo IV

Aplicación práctica

IV.1. Introducción

Como ya se ha mencionado en el Capítulo II, para efectuar un estudio en el que se apliquen los procedimientos de descubrimiento de conocimiento en bases de datos y específicamente la tarea de minería de datos, es esencial contar con una información suficiente y fiable.

Por otra parte, tal como ha podido verse en el estudio del estado del arte, se presentan diferentes posibilidades de aplicación del manejo de datos e información aplicados a los sistemas de abastecimiento de agua, con el fin de obtener patrones y conocimiento nuevo a partir de esta información. Tal como lo destacan Gou y Ding (2009), las técnicas de minería de datos se presentan como herramientas que ofrecen efectos altamente positivos en la resolución de problemáticas en los sistemas abastecimiento de agua. Una razón es que estas técnicas ofrecen modelos predictivos, con la ventaja de que éstos pueden ser no numéricos. También que permiten una mejor comprensión de los fenómenos que no pueden ser caracterizados o cuantificados numéricamente, como es el caso de los daños que se presentan en la red; lo cual beneficia el tener sistemas sostenibles, más aún, cuando se está haciendo uso de un recurso natural de vital importancia.

Muchas empresas encargadas de la gestión de los abastecimientos pueden tener sus reportes de daños en la red; pero podemos preguntarnos que utilidad le dan más allá de tener almacenada esta información en formato de papel como un simple anecdotario histórico de lo ocurrido en la red, sin plantearse la utilidad que dicha información representa al momento de gestionar el abastecimiento.

Aunque estamos tratando con un bien público, necesario y escaso en algunos casos, tal como lo es el agua, para la aplicación de la metodología en temas relacionados con los sistemas de abastecimiento de agua no es tarea fácil conseguir la información necesaria y adecuada, debido al recelo en el manejo de esta información por parte de los operadores de los sistemas de abastecimiento. El interés que nos planteamos con esta tesis es presentar esta variedad de

herramientas como soporte en la gestión y operación de los sistemas; por tanto forma parte importante el hecho de contar con una información real y veraz de lo que sucede en la problemática diaria de los abastecimientos.

No obstante, aunque no en demasiada cantidad (temporalidad y variables), nos ha sido facilitada una información para el desarrollo de este trabajo por parte de la empresa MULTIPROPÓSITO DE CALARCÁ S.A., E.S.P., que maneja y opera el sistema de abastecimiento de agua potable del municipio de Calarcá, en el departamento del Quindío en la región cafetera de Colombia. Esta información será utilizada con el fin de obtener conocimiento a partir de la aplicación de ciertas técnicas de minería de datos, basándonos en el esquema de la obtención de conocimiento en base de datos descrito en el Capítulo II.

En un principio el interés del trabajo se establece como abierto a los resultados que puedan ser obtenidos tal como se plantea en los objetivos de desarrollo de esta tesis, siendo ésta una de las grandes ventajas de las técnicas de minería de datos con respecto de las técnicas estadísticas clásicas; ya que no se tienen necesariamente predeterminadas variables dependientes de variables independientes; aunque, como objetivo final, se establece la predicción de los daños reportados basándose en el resto de información disponible, ya que corresponde a uno de los puntos críticos tanto de funcionamiento como de gestión económica en el manejo de la red. El poder caracterizar la diversidad problemática presente en la operación de una red de abastecimiento a partir de su comportamiento, es una tarea que indudablemente presenta beneficios tanto para el conocimiento de la misma, como a la hora de plantearse actuaciones a realizar sobre ella.

Como se ha indicado, la aplicación práctica de esta tesis se realiza con información obtenida del sistema de abastecimiento de agua potable de un municipio de la región cafetera colombiana; esta información básicamente corresponde a reportes de daños obtenidos de la red de abastecimiento durante el año 2006, junto con el modelo hidráulico del sistema para un periodo de un día. Tal como se establece en los objetivos de la tesis, se quiere encontrar la

aplicabilidad del descubrimiento de conocimiento en bases de datos a partir de esta información obtenida, con el fin de aportar nuevo conocimiento del sistema que permita contribuir con una mejor operación de la gestión de la red de abastecimiento de agua potable.

En este capítulo se hace una descripción detallada de la zona de estudio, así como del manejo y tratamiento que se le dio a la información suministrada como base del siguiente capítulo de resultados y discusión.

IV.2. Descripción de la zona de estudio

El sistema de acueducto (abastecimiento) del Municipio de Calarcá de acuerdo con información suministrada por la empresa operadora, está compuesto principalmente por:

- *Sistema de captación:* actualmente se abastece de cuatro fuentes; Río Santo Domingo, Quebrada el Salado, Quebrada San Rafael y Quebrada Naranjal.

- *Sistema de Conducción:* Santo Domingo – Salado; una línea de conducción de longitud 4,2 Km en Hierro Dúctil de 12" (304.8 mm). San Rafael; una línea de conducción de longitud 2,9 Km en tubería de Acero al Carbono 12" (304.8 mm). Naranjal; una conducción de longitud de 800m en PVC 8" (203.2 mm).

- *Producción de Agua Potable:* La Planta de potabilización es de tipo convencional con técnicas de filtración a tasa constante, y lavado con tanque y sedimentadores convencionales. Las unidades que componen la planta son: canal de aproximación y mezcla rápida, floculadores mecánicos, sedimentadores convencionales, filtros de tasa constante, tanque de cloración, laboratorio, y zona de productos químicos.

- *Sistema de Almacenamiento:* está conformado por cuatro tanques enterrados de hormigón armado con aproximadamente 5.600 m^3 de volumen de

almacenamiento.

- *Sistema de Distribución de Agua Potable:* está conformado por aproximadamente 120 Km de redes entre principales y secundarias.

El sistema de distribución tiene a su vez sistemas de macromedición, micromedición, sectorización, hidrantes, y estaciones reductoras de presión, entre otros componentes, mediante los cuales se garantiza la prestación del servicio de abastecimiento de agua potable. Este servicio en el área urbana del municipio de Calarcá tiene una cobertura del 100%.

En cuanto a las carencias que se presentan en la red, se tiene el cumplimiento de la vida útil de materiales como: hierro fundido (HF), hierro galvanizado (HG), y asbesto cemento (AC); al igual que la insuficiencia hidráulica en diámetros menores a 3" (1/2", 3/4", 1", 1 1/2", 2").

IV.2.1. Ubicación Geográfica

Figura IV.1. Colombia en Sudamérica y ubicación del departamento del Quindío en Colombia

Colombia se encuentra ubicada al norte de Sudamérica teniendo como

fronteras por el norte el mar caribe, por el este los estados de Venezuela y Brasil, por el sur Ecuador, Perú y Brasil, y por el oeste el océano pacífico y el estado de Panamá. El departamento del Quindío (Figura IV.1) y (Figura IV.2) cuya capital es el municipio de Armenia se encuentra ubicado en la denominada región andina de Colombia en el costado occidental de la cordillera central. Limita por el norte con el departamento de Risaralda, por el este y sur con el departamento de Tolima, y por el sur y oeste con el departamento del Valle del Cauca.

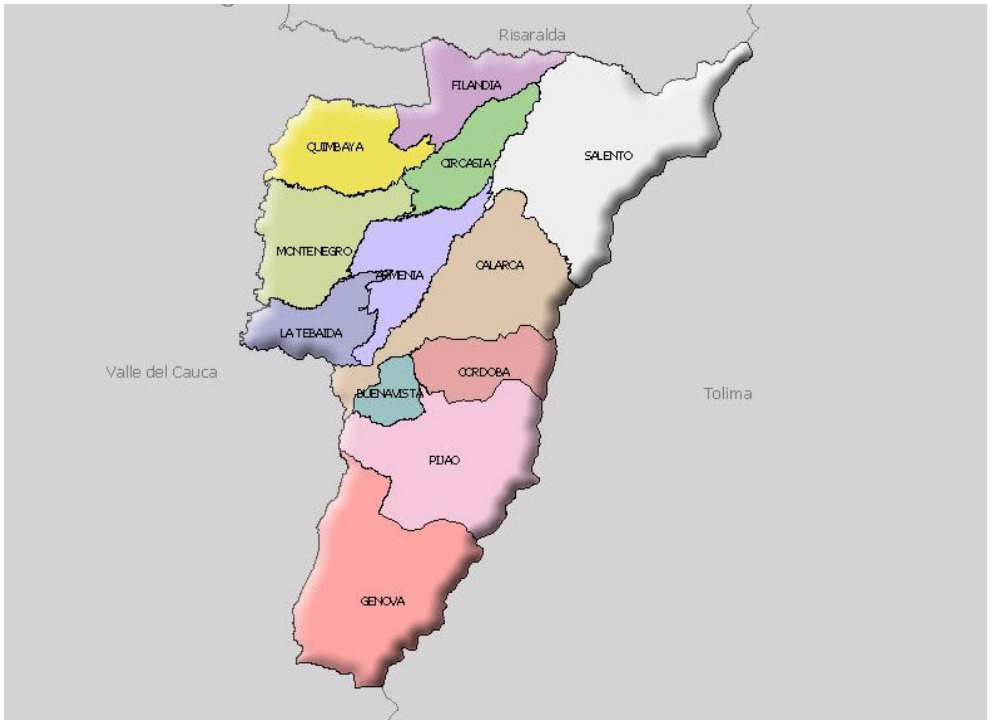


Figura IV.2. Municipios del Departamento del Quindío

El área municipal de Calarcá tiene las siguientes fronteras:

- Norte: el municipio de Salento.
- Oriente: el municipio de Cajamarca (departamento del Tolima).
- Sur: los municipios de Córdoba, Buenavista, Pijao en el Quindío, y Caicedonia en el departamento del Valle del Cauca.
- Occidente: los municipios de La Tebaida y Armenia.

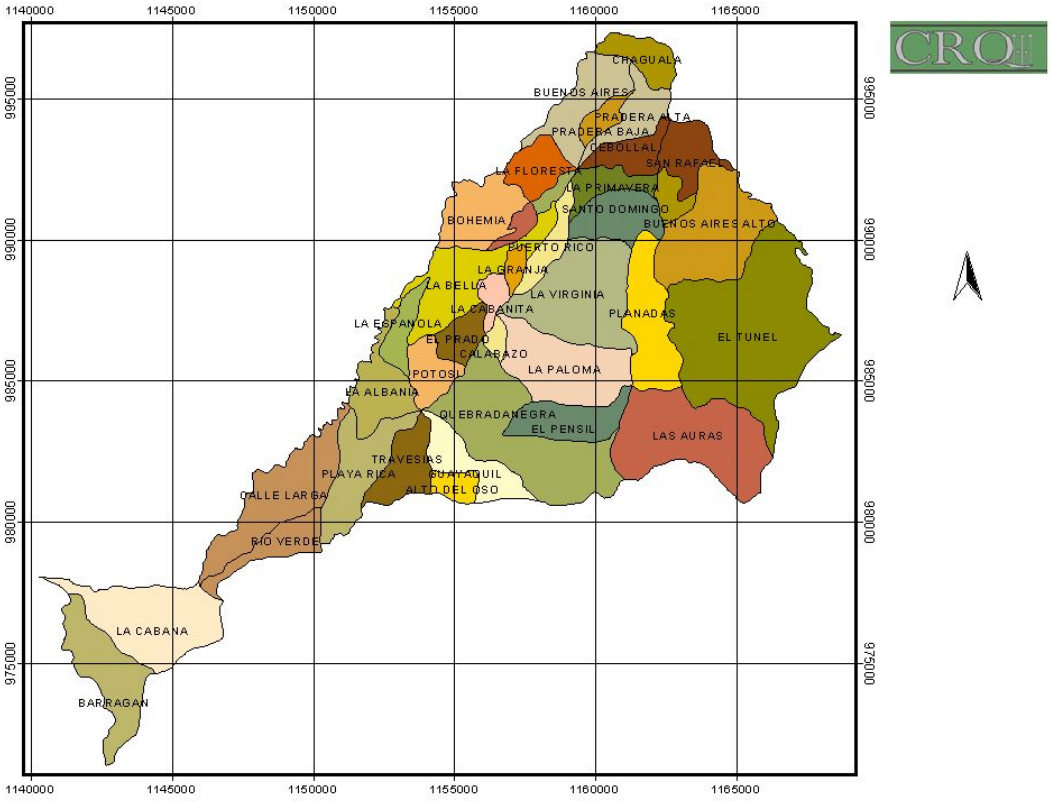
La cabecera municipal está ubicada en los 4º 04' 56,57" de latitud Norte y 74º 04' 51,03" de longitud Oeste, respecto al meridiano de Greenwich, y a una

altura de 1536 metros sobre el nivel del mar.

El municipio tiene una extensión territorial de 21923 hectáreas (has.), de las cuales 244 has. corresponden al perímetro urbano, y 21679 has. corresponden al sector rural.

La población municipal es de 73741 habitantes (censo 2005), distribuida en 56200 habitantes en el sector urbano y 17541 habitantes en el rural.

IV.2.2. Localización General del municipio de Calarcá



Fuente: http://www.crq.gov.co

Figura IV.3. División política del Municipio de Calarcá

El municipio de Calarcá (Figura IV.3) está ubicado al oriente del Departamento del Quindío (Colombia), sobre el costado occidental de la Cordillera Central. La altitud sobre el nivel de mar varía entre los 1000 metros, en la confluencia de los ríos Quindío y Barragán, formación del río La Vieja, y los 3640

metros, en el Alto del Campanario en la vereda El Túnel.

Las coordenadas planas del municipio de Calarcá con origen Bogotá son:

X: 997500 mE Y: 836096 mN

X: 971862 mE Y: 809938 mN

IV.2.3. Geología Y Geomorfología

Las características geomorfológicas presentes reflejan procesos de origen endógeno asociados con el complejo volcánico sedimentario Quebrada Grande del período Cretáceo; la secuencia sedimentaria de origen volcánico está conformada por: derrames de lavas intercaladas con pizarras arcillosas y silíceas, limolitas y liditas.

En su composición geomorfológica se tiene paisaje de montaña compuesto por filas y vigas de clima medio húmedo y muy húmedo; de relieve ondulado y fuertemente ondulado; y paisaje de piedemonte formado por relieve ondulado y plano con vallecitos. Se determina el modelado torrencial que forma el Glacis del Quindío, originado en cenizas volcánicas y meteorización de las rocas ígneas y metamórficas.

El departamento del Quindío se localiza sobre la zona de influencia del sistema de fallas de Romeral, el cual tiene inicio al sur del Golfo de Guayaquil en el Ecuador y se interna en el mar Caribe al norte en la región de Barranquilla (departamento del atlántico). Comprende un numeroso conjunto de fallas paralelas que, en el departamento son conocidas con los nombres de: Falla de Silvia – Píjao, - Falla Cauca – Almaguer, y Falla Campanario – San Jerónimo, cuya actividad se remonta desde el periodo Paleozoico hasta la actualidad.

IV.2.4. Climatología

Por su situación geográfica con marcadas variaciones de altura sobre el

nivel del mar, el territorio municipal tiene gran diversidad climatológica; comprende las clasificaciones de las zonas de vida: bosque húmedo tropical (bh-T), bosque muy húmedo premontano (bmh-PM), bosque muy húmedo montano (bmh-M), bosque húmedo montano (bh-M), y páramo. La temperatura oscila entre los 8ºC en la zona de páramo, y los 25ºC en la zona de valle.

Para el análisis climático del área del municipio de Calarcá, se utilizó información del anuario meteorológico de Cenicafé 2006, cuya estación meteorológica (La Bella) se encuentra ubicada dentro del límite municipal (vereda La Bella) al sur de la cabecera del municipio, como se visualiza en la Figura IV.4. Dicha información corresponde a la recolección de datos en un periodo de más de veinte años, por lo cual se considera confiable.

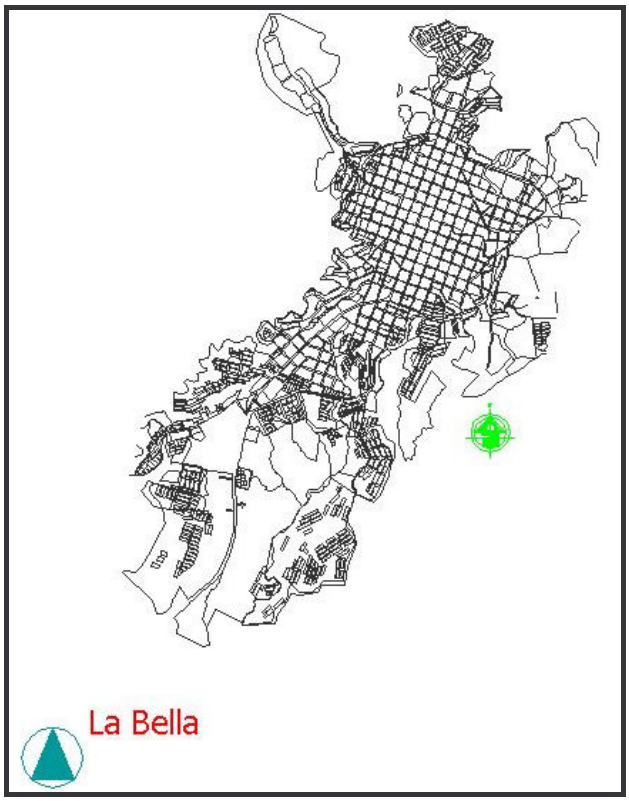


Figura IV.4. Localización estación La Bella

NOMBRE	*MUNICIPIO*	*LATITUD NORTE*	*LONGITUD OESTE*	*ALTITUD (M)*	*CUENCA HIDROGRÁFICA*
La Bella	Calarcá	4°- 31’	75°- 40’	1450	Quindío

Tabla IV.1. Localización de la Estación Meteorológica La Bella

Fuente: CENICAFÉ, Anuario Meteorológico, 2006

El clima en el municipio de Calarcá se clasifica de acuerdo a los niveles de altitud y pisos térmicos como sigue:

CLASIFICACION DEL CLIMA	***ALTITUD m.s.n.m***	***TEMP ºC***	***LLUVIA mm año***	***CARACTERÍSTICAS***
Extremadamente frío pluvial (EF-P)	3500-3600	6-9	1800	La actividad agropecuaria es inexistente por la agresividad climática, dificulta el establecimiento de núcleos humanos.
Muy frío pluvial (MF-P)	3000-3500	9-12	2200	Las bajas temperaturas y la humedad en estas zonas están condicionadas por la neblina constante y los vientos; originadas por la abundancia de precipitación y transpiración vegetal.
Frío muy Húmedo (F-MH)	2000-3000	12-18	2200	Condiciones ideales para la ganadería de leche y cultivos forestales.
Medio húmedo y muy húmedo (M-MH)	1300-2000	18-24	2100	Zona Cafetera hasta los 1800 m.s.n.m. , condiciones óptimas para los cultivos de esta zona
Medio húmedo transicional a medio seco.	1000-1300	18-24	1500	Zona Cafetera, condiciones óptimas para los cultivos de esta zona

Fuente: Acuerdo No. 015 2000.

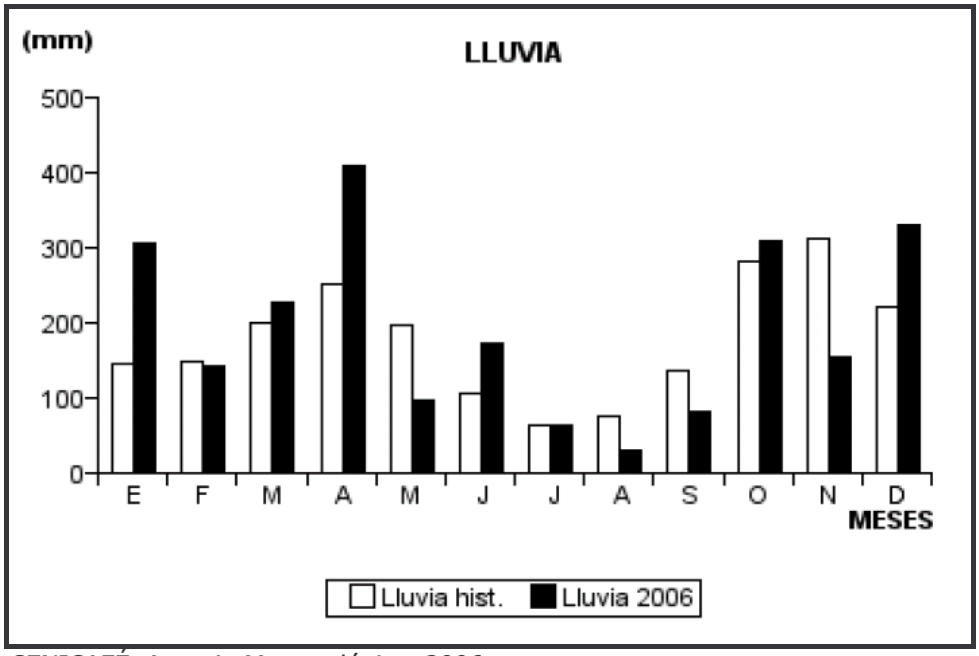
Tabla IV.2. Clasificación del Clima en Calarcá

IV.2.4.1. Precipitación

El comportamiento de la lluvia presenta una tendencia bimodal, con dos periodos de lluvias más intensas (marzo – abril – mayo, y octubre – noviembre – diciembre), y dos menos lluviosos (enero – febrero, y junio – julio – agosto – septiembre).

El clima del municipio está regido por el comportamiento de la zona de

convergencia inter-trópical (CIT), que se desplaza en forma cíclica de sur a norte y de norte a sur. Este desplazamiento se debe a que una mayor temperatura en la zona ecuatorial aumenta la evaporación de las masas de agua, con lo cual, se forman una mayor cantidad de nubes. La insolación en el Ecuador es permanente, pero hay dos periodos en el año durante los cuales el sol incide perpendicularmente (equinoccios en marzo 21 y septiembre 21), produciendo la mayor temperatura.



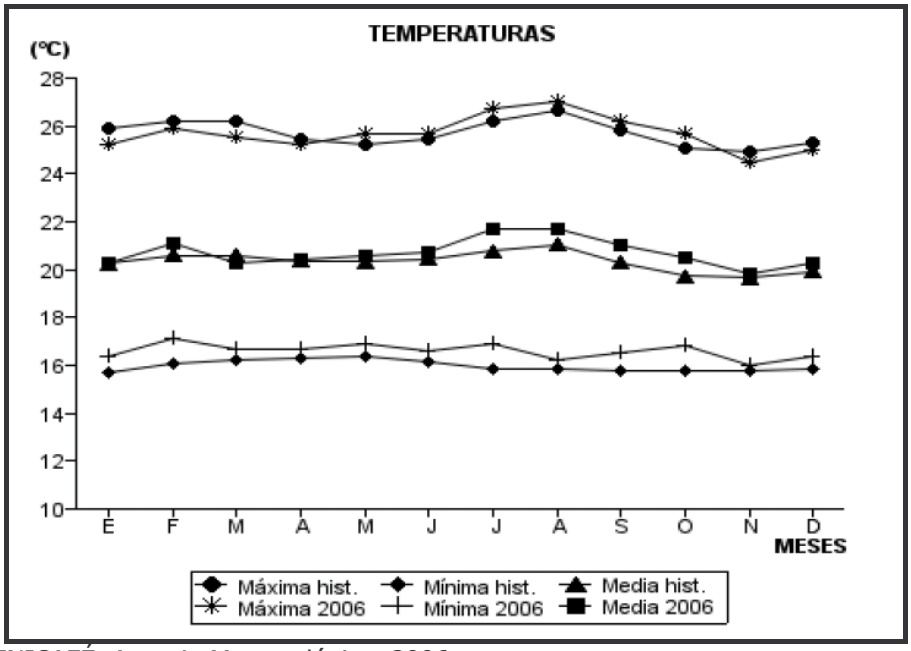
Fuente: CENICAFÉ, Anuario Meteorológico, 2006.

Figura IV.5. Precipitaciones Estación La bella

Se aprecia en esta figura un comportamiento atípico para el año 2006 de acuerdo con el registro histórico de precipitaciones para la estación, presentándose tanto periodos secos como lluviosos más pronunciados.

IV.2.4.2. Temperatura

El municipio presenta una temperatura promedio de 20.3°C, obteniendo los menores registros en la zona montañosa debido a la altura, influencia de vientos, lluvia, y deficiencia de rayos solares como consecuencia de la alta nubosidad durante gran parte del año.

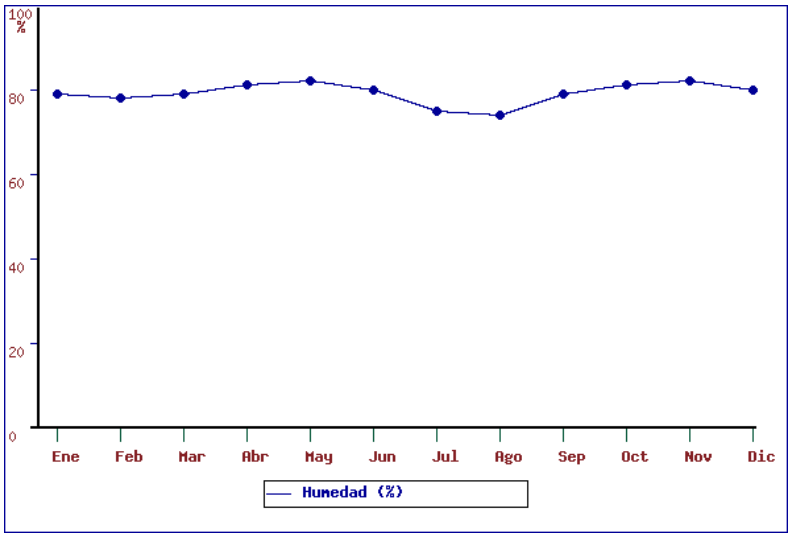


Fuente: CENICAFÉ, Anuario Meteorológico, 2006.

Figura IV.6. Temperaturas Estación La bella

IV.2.4.3. Humedad relativa

El municipio de Calarcá presenta un promedio anual del 84.37% de humedad relativa, su rango de variación está entre un 5% entre periodos secos y periodos de lluvia.



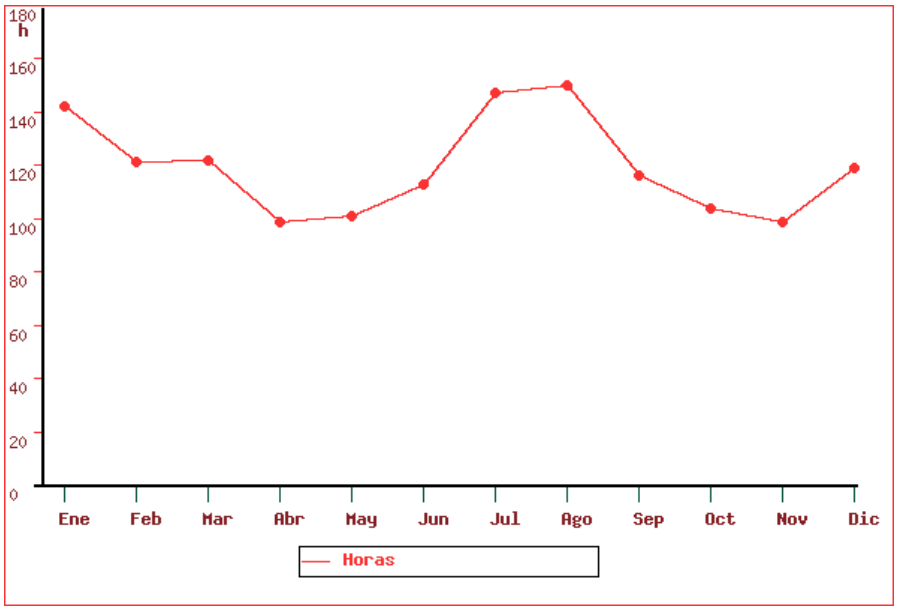
Fuente: www.cenicafe.org

Figura IV.7. Distribución anual de la Humedad Relativa en la Estación La Bella

IV.2.4.4. Brillo solar

En el municipio las zonas con menor brillo solar son las que se encuentran sobre los 2500 m.s.n.m., específicamente en las veredas Buenos Aires Alto, El Túnel, Las Auras, y Planadas, todas ellas sobre el área montañosa de la Cordillera Central. Esto se debe a la alta nubosidad y constante presencia de neblina, lo que no permite la entrada directa de la luz solar durante gran parte del día, disminuyendo por lo tanto los periodos de brillo solar.

La duración del brillo solar en el municipio de Calarcá es aproximadamente de 1500 a 2000 horas/año, siendo más corta en la mañana que en la tarde.



Fuente: www.cenicafe.org

Figura IV.8. Distribución mensual del Brillo Solar en la Estación La Bella

IV.2.4.5. Vientos

Se puede estimar la velocidad del viento en el municipio de Calarcá de la siguiente forma:

0.86 m/s dirección norte - oeste durante el día, y 0.33 m/s dirección oeste en la noche; la máxima velocidad del viento puede ser considerada aproximadamente entre 15 -20 m/s.

IV.2.5. Hidrografía

IV.2.5.1. Red de Drenaje

La red hidrográfica del municipio de Calarcá está compuesta principalmente por la cuenca del río Santo Domingo, ocupando aproximadamente el 64% [155 km^2] del área total del municipio; seguido en importancia por el área que comparten con el municipio parte de las cuencas hidrográficas de los ríos Quindío, río Verde, y río Barragán.

La gran red de drenaje compuesta por los ríos mencionados anteriormente, se caracteriza por la dependencia de su forma y uniformidad de la litología por la que atraviesa, y de los ejes tectónicos o fallas presentes, y por su intensa actividad de disección. Debido a estos factores, la hidrografía del municipio tiene una forma predominantemente dendrítica, es densa y ha actuado intensamente sobre la superficie del abanico de Armenia, moldeándolo hasta presentar la actual morfología.

IV.2.5.2. Usos del Recurso Hídrico

En general, las corrientes de agua presentan dos características principales respecto a los servicios, por una parte son fuentes de abastecimiento de los acueductos municipales y veredales, y por otra, funcionan como los colectores de los alcantarillados urbanos, aguas negras de viviendas aisladas, centros poblados, vertimientos del matadero municipal, actividades agroindustriales, agrícolas, y pecuarias.

Adicionalmente, en el ámbito regional, se le suma la recepción de lixiviados y escorrentía superficial de sitios de disposición de los residuos sólidos.

IV.2.5.3. Cuencas Hidrográficas

Todo el departamento del Quindío drena sus aguas superficiales a la cuenca

del río La Vieja, que corre por el costado occidental sirviendo de límite departamental. Se divide en dos cuencas menores: la del río Quindío que recoge las aguas de la mitad del norte del departamento, a lo largo de un recorrido transversal desde el extremo nororiental hacia la parte media del costado occidental de éste, y la del río Barragán que recoge las aguas de la mitad sur del departamento, en un recorrido de sur a norte sirviendo de límite departamental hasta encontrarse con el río Quindío.

Calarcá está ubicado dentro de la cuenca del río Quindío y por su jurisdicción corre uno de los principales afluentes de éste, el río Santo Domingo, teniendo una gran ventaja sobre otros municipios del departamento en cuanto a la oferta y demanda de agua debido a que, éste nace y desemboca dentro de la jurisdicción municipal.

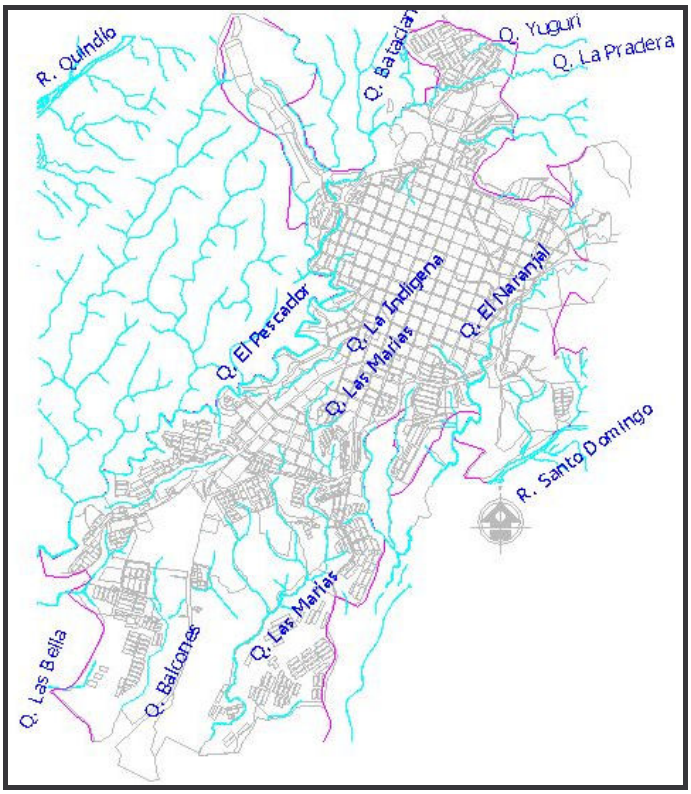


Figura IV.9. Red de drenaje en el municipio de Calarcá

Cuenca del río Quindío

Los principales ríos que pasan por Calarcá presentan una característica en

común; y es que, sus corrientes convergen hacia el suroeste del municipio.

El río Quindío nace al norte del departamento en el municipio de Salento, a una altitud aproximada de 3800 m.s.n.m. sobre la Cordillera Central. Parte de la margen izquierda de la cuenca del río Quindío pertenece a Calarcá, ocupa un área aproximada de 36 kms^2, a partir de la quebrada El Castillo hasta el sitio en que convergen sus aguas con las del Río Verde; sirviendo de límite noroeste con el municipio de Armenia.

El río Quindío en el sector del centro poblado La María, pierde totalmente su potabilidad y cualquier signo de vida debido al alto grado de contaminación, producto del proceso del cuero en las curtiembres, además de los desechos del matadero municipal de Armenia.

Cuenca del río Verde

Nace a 3500 m.s.n.m. en la cordillera central en el municipio de Córdoba y cruza hacia el oeste del municipio por un relieve montañoso escarpado hasta llegar al asentamiento denominado Río Verde, a partir de ese lugar ocupa una pequeña parte del área del municipio de Calarcá, atravesando la zona en sentido este - oeste hasta confluir al río Quindío; el área que ocupa dicha cuenca es de aproximadamente 31 kms^2, ocupando la parte alta de las veredas La Cabaña y Río Verde.

Cuenca del río Barragán

Tiene su origen en el Municipio de Génova a una altura aproximada de 2800 m.s.n.m., corre hacia el norte del departamento por un relieve montañoso hasta llegar a la altiplanicie donde sirve de límite del sector sur - oeste del municipio de Calarcá con el departamento del Valle; finalmente deposita sus aguas en el río Quindío formando el río La Vieja. El área que ocupa en el sector es de 20 kms^2, cartográficamente solo se identifica una quebrada que la compone dentro de Calarcá, la quebrada La Picota. El sector de la cuenca ocupa las veredas

Barragán y la parte baja de la vereda La Cabaña.

Cuenca del río Santo Domingo

Como se mencionó anteriormente, esta cuenca representa la mayor importancia para el desarrollo económico y social del municipio de Calarcá. La cuenca del río Santo Domingo ocupa un área de 155,14 km^2, abarcando la mayoría de la zona rural (veredas Las Auras, El Túnel, Buenos Aires Alto, Planadas, San Rafael, Santo Domingo, La Primavera, Puerto Rico, La Virginia, La Paloma, El Calabazo, El Pensil, Quebradanegra, Playa Rica, Alto del Oso, Guayaquil, Travesías, Potosí).

El acueducto urbano del municipio de Calarcá se alimenta de cuatro fuentes: el río Santo Domingo y las quebradas El Salado, San Rafael, y Naranjal, afluentes del Santo Domingo. Se tienen cuatro bocatomas en las fuentes citadas así:

- Una sobre el río Santo Domingo a 1613 m.s.n.m.
- Una sobre la quebrada El Salado a 1635 m.s.n.m.
- Una sobre la quebrada San Rafael a 1630 m.s.n.m
- Una sobre la quebrada el Naranjal a 1620 m.s.n.m.

Las áreas de captación de las microcuencas que permiten el abastecimiento del acueducto urbano se localizan aguas arriba de las alturas relacionadas. Las microcuencas se ubican sobre una vertiente de cordillera con relieve escarpado a muy escarpado (50 al 75%, y más), con predominio de rocas magmáticas del complejo Quebrada Grande (Litodema volcánico) en la parte baja, y sedimentarias del complejo Quebrada Grande (Pizarras arcillosas grauvacas y limolitas) en la parte alta. Los suelos corresponden principalmente a la asociación Chinchiná - El Cedral en la parte alta, y la asociación Santa Isabel-Herbeo.

Morfometría de la cuenca

Los parámetros morfométricos más importantes de la cuenca indican su comportamiento geomorfodinámico. La cuenca se caracteriza por presentar rocas relativamente blandas, que dan lugar a una red de drenaje de moderada densidad, desarrollada en un área grande. La alta descarga de sedimentos debida al relieve presenta tres causas principales: el predominio de pendientes entre 30 y 60%, que facilita el escurrimiento superficial y el arrastre de gran cantidad de sedimentos; el flujo sub-superficial de agua que también arrastra material, influenciado por la alta capacidad de infiltración; y por último, el uso actual del suelo en el cual predomina la ganadería extensiva sin técnica alguna, dejando áreas amplias con una cobertura vegetal mínima (gramíneas).

La forma semicircular de la cuenca y el hecho de que el río presenta un valle en "V" profundo cerca a Calarcá determinan la Pendiente de la cuenca; de acuerdo con el mapa de pendientes para la cuenca se definen 5 rangos:

Rango de Pendiente	*Procesos Característicos y condiciones del Terreno*	*Vegetación y cultivos comunes en la Zona*
0-15 % (0.8º)	Plano a Inclinado: denudación no apreciable, registrándose socavamiento de orillas. Es una zona laborable con maquinarias agrícolas.	Terrenos con plátano, frutales, pastos, guadua, café con y sin sombrío. Área 867 Has (10%).
15-30% (8-16º)	Moderadamente empinado a empinado: a partir de este rango se generan procesos erosivos superficiales y algunos de remoción en masa. Es posible utilizar maquinaria agrícola pesada en las pendientes más bajas, se recomienda arar en forma perpendicular a la pendiente.	
30-60% (16-32º)	Empinado: se presenta la mayor cantidad de procesos erosivos de la cuenca. Generalmente movimientos en masa como: Deslizamientos, reptación, y carcavamiento; erosión superficial como: graderías y terracetas. Presenta posibilidades limitadas para el arado.	La cobertura vegetal corresponde a bosques de galería, secundario, primario, e implantado, y pastos. Área 3468 has. (40%).

Rango de Pendiente	*Procesos Característicos y condiciones del Terreno*	*Vegetación y cultivos comunes en la Zona*
60-90% (32º-42º.9)	Empinado a muy empinado, se presentan procesos erosivos principalmente de origen gravitatorio, como caídas de roca, derrumbes en los afloramientos rocosos. Es imposible realizar cultivos tecnificados.	Predomina la cubierta vegetal de bosque primario de páramo, pastos, cultivos de papa, y bosque implantado. Área 1734 has. (20%).
> 90% (<42%)	Extremadamente empinado: afloramientos rocosos conformando escarpes, los procesos denundacionales son las caídas de roca. Es imposible cualquier labor agrícola.	Cubierta vegetal de rastrojo y, bosques primarios de páramo. Área 1300.5 has. (15%).

Tabla IV.3. Características de los rangos de pendientes del río Santo Domingo

Procesos erosivos

Los procesos erosivos de la cuenca son consecuencia de los factores hídricos, antrópicos (mal manejo de suelos), y estructurales influenciados por la tectónica; debido a que la cuenca se encuentra sobre un terreno ligado a zonas de debilidad, asociados a fallas y fracturamientos. Los procesos erosivos corresponden a superficiales (carcavamiento y generación de surcos, erosión laminar, graderías y terracetas, golpes de cuchara, socavamiento de orillas), y movimientos en masa (deslizamientos planares, rotacionales, derrumbes, caídas, flujos de lodo y escombros, reptación, subsistencia de carreteables).

La situación después del movimiento telúrico del 25 de enero de 1999 relacionada con los procesos erosivos presentes en el municipio de Calarcá, es una condición especial, por cuanto, se observa mayor incidencia de ellos, en la presentación de múltiples "golpes de cuchara" y deslizamientos en mayor grado. Este fenómeno se ha intensificado por la ocurrencia de la continua y alta precipitación (alta frecuencia de lluvias en la zona).

Riesgos derivados de los procesos naturales

A. Carcavamiento Y Generación de Surcos (C-S)

Se presenta en las áreas utilizadas para la ganadería extensiva, sobre los 1800 m.s.n.m., en suelos inconsolidados susceptibles a la erosión; corresponde a la pérdida del suelo en zonas de terracetas producto del sobrepastoreo, las cuales se unen y profundizan formando surcos por concentración del escurrimiento superficial.

Son causales de estos procesos la intensidad de las lluvias sobre áreas con sobrepastoreo, cultivos limpios, mala labranza, y deforestación en pendientes mayores al 30%.

B. Erosión Laminar

Es el arrastre uniforme de las capas de suelo por la escorrentía, generado por la poca o nula cobertura vegetal; se observa principalmente en los potreros, márgenes de la quebrada Pinares, escarpes del río Santo Domingo, y en las divisorias de aguas de las quebradas La Gata y Uritá, con pendientes mayores al 30%.

C. Graderías Y Terracetas

Son microformas originadas por el pisoteo continuo del ganado (sobrepastoreo) en áreas pendientes, formando una especie de zanjas con coberturas vegetal sobre el talud; los cultivos con surcos trazados paralelamente a curvas de nivel, como cultivos tecnificados, café, papa, y hortalizas, totalmente descubiertos, generan las terracetas debido al corte del horizonte A.

D. Golpes de Cuchara

Corresponden a deslizamientos, derrumbes, y flujos que conforman los movimientos de masa superficiales de pequeñas dimensiones, poca profundidad, y producto de la acción antrópica (tala de bosques para ampliar las fronteras agropecuarias), que unido a las características del suelo (piroclastos inconsolidados y abundancia de arenas), no permite un buen anclaje de la vegetación y, al ser deforestada, se genera un desequilibrio e inicia el proceso erosivo.

E. Socavamiento de Orillas

Es el producto de las corrientes de agua al golpear contra las orillas, socavando y desestabilizando las riveras por la pérdida del soporte y, generando derrumbes, ampliando el cauce, y cambiando la morfología del terreno.

F. Movimiento de Masa

Son aquellos en los que la capa activa y el plano de cizallamiento están más profundos que el sistema radicular de los árboles, y se relaciona estrechamente con las condiciones estructurales, la disección, y evolución natural de las vertientes.

G. Deslizamientos Planares

Son deslizamientos en que la masa se mueve hacia afuera y abajo de una superficie más o menos plana, controlada por planos de estratificación de esquistocidad o de diaclasamiento, que concuerda aproximadamente con la pendiente y, puede progresar indefinidamente a lo largo de la pradera. Abarcan aproximadamente unas 17.5 has. Son producto del corte en las vías con pendientes muy fuertes, cultivos limpios con aguas subterráneas y de escorrentía, la gravedad, y la poca cohesión superficial, unida a la alta permeabilidad.

H. Deslizamientos Rotacionales

Son los movimientos de material a lo largo de una superficie cóncava con movimiento lateral y rotacional en sentido opuesto a la pendiente, ocasionados por el efecto del agua, la pendiente, y la gravedad. Se amplían por erosión remotamente y caídas, ocupan una área aproximada de 19.6 has, son frecuentes en las rocas de la unidad volcánica al oeste de la zona.

I. Derrumbes

Se producen cuando hay movimiento de materiales, los cuales no siguen una superficie de deslizamiento definido por tener movimientos caóticos en pendientes mayores del 35%, se encuentran asociados con los cortes de carretera que desestabilizan las formaciones superficiales produciendo el desplazamiento rápido.

J. Caídas

Son desprendimientos o desplomes de roca, suelo, o escombros, de forma rápida; en que la masa se desplaza su mayor tiempo en caída libre. Son producto de la gravedad, diaclasamiento, foliación, meteorización, y socavamiento en la base de los taludes.

K. Flujo de lodo y escombros

Son movimientos de material no consolidados con un alto contenido de agua, que se desplaza a lo largo de una pendiente o por un drenaje natural; están ligados al exceso de agua en suelos de origen volcánico, ocupan una extensa área aproximada de 22 Has.

L. Reptación

Es un flujo lento de formación superficial que se da sobre materiales con planos favorables de deslizamientos o, sobre zonas con materiales metamórficos en avanzado estado de meteorización. Son generados por cambios bruscos de pendientes, el peso del estrato superior saturado del agua, los movimientos tectónicos, y la acción negativa del hombre, en suelos compuestos por cenizas volcánicas.

M. Subsistencia de carreteables

Esta puede ser rápida; causada por, el lavado diferencial y los hundimientos lentos por consolidación natural o sobrepeso.

N. Erosión antrópica

En la cuenca, las actividades humanas que más influyen en la generación de procesos erosivos son:

- Ganaderías intensivas sin ningún manejo de conservación de suelos; en terrenos pendientes moderados a empinados, sobre cenizas volcánicas inconsolidadas, producen graderías y terracetas.

- La explotación del horizonte arcilloso del perfil de meteorización de la Unidad Volcánica, para la fabricación de ladrillos, desarrollada en las márgenes de la quebrada El Naranjal.

- Cultivos Limpios (café tecnificado, papa, y hortalizas) en los cuales se deja al descubierto la capa superficial, para que la acción sea más intensa; la deforestación, las quemas, y el trazo de vías de comunicación sin estudios previos, en zonas donde el diaclasamiento como la estratificación y foliación están a favor de la pendiente y, donde la cubierta de ceniza volcánica por efecto del agua y la pendiente se puede desplazar.

Contaminación Hídrica

La causa principal de la contaminación hídrica en Calarcá la determina las vertientes finales de las aguas residuales a los cauces de las quebradas y ríos, (ríos Santo Domingo, y Quindío, quebradas El Naranjal, La María, y El Pescador). En segundo lugar, la contaminación proveniente del beneficio del café que dispone su agua contaminante al cauce hídrico más cercano.

Las centrales de sacrificio, como el matadero municipal que dispone sus residuos en el río Santo Domingo, y las granjas avícolas, que desaguan en los cauces más próximos, y las curtidoras de cuero, que además de crear contaminación orgánica, son los principales contaminadores químicos del río Quindío.

Otras contaminaciones menores, pero no poco vulnerantes, se dan eventualmente en las arterias viales cuando los vehículos que transportan químicos se accidentan, los productos transportados son recibidos por las fuentes hídricas, además del uso de plaguicidas y abonos químicos en las labores agrícolas.

Finalmente la contaminación por residuos sólidos, que se presentan en las márgenes de las quebradas, debido a que sus moradores depositan sus basuras y otros elementos inservibles en los cauces de las quebradas

IV.3. Descripción de la información

IV.3.1. Generalidades del abastecimiento de Calarcá

La información utilizada en esta tesis fue proporcionada por la *EMPRESA MULTIPROPÓSITO DE CALARCA S.A. ESP*, cuya actividad económica está dedicada a la administración y operación de los servicios públicos domiciliarios de acueducto, alcantarillado, aseo y generación de energía eléctrica para su venta en bloque. Esta empresa está localizada en el municipio de Calarcá, departamento del

Quindío en Colombia (sur América). La sociedad anónima se creó en octubre de 2002 en la cual el socio mayoritario es Empresas Publicas de Calarcá EMCA ESP con el 40% de las acciones, y los restantes provenientes de capital privado. Los datos utilizados para el desarrollo de esta tesis no son de accesibilidad pública, por tanto se realizó una solicitud escrita para su obtención.

El sistema de acueducto presta sus servicios en el área urbana del municipio de Calarcá, que equivale 14.229 usuarios. Según datos ofrecidos por la empresa, para el año 2006 el índice de agua no contabilizada en la distribución correspondía al 49.03%, y para el año 2007 este índice tenía un valor del 42.44%.

- **Proceso de producción de agua potable**

La planta de potabilización de agua tiene una capacidad de diseño de 260 litros por segundo (l/s), distribuidos así: planta convencional 100 l/s, y planta compacta (decantador pulsator) 160 l/s, que actualmente se encuentra fuera de servicio (información suministrada por la Empresa el 13 de junio de 2008).

1. *Captación:* el agua se toma de fuentes naturales superficiales, el río Santo Domingo, quebrada El Salado, quebrada El Naranjal, y quebrada San Rafael a través de unas rejillas (bocatomas) que impide que elementos como hojas, palos, y otros de mayor tamaño queden retenidos y no entren a través de ellas. Todas las fuentes de abastecimiento son utilizadas y de igual importancia por sus características organolépticas, físicas, químicas, microbiológicas, y su aporte en caudal en las diferentes épocas del año.

2. *Conducción:* actualmente se cuenta con tres líneas de conducción de agua cruda, una que conduce el agua proveniente del río Santo Domingo y la quebrada El Salado en tubería Hierro Dúctil (HD) de diámetro 12”, otra que conduce el agua proveniente de la quebrada San Rafael en tubería de diámetro de 10” en Acero al Carbono, y otra que conduce el agua proveniente de la Quebrada Naranjal en tubería de diámetro 8”. Las conducciones de agua cruda se vigilan diariamente y se realiza mantenimiento a las ventosas y válvulas existentes, con una periodicidad semestral.

3. *Desarenador:* corresponde a un tratamiento primario y consiste en una cámara destinada a la remoción de las arenas y sólidos que están en suspensión en el agua, mediante un proceso de sedimentación por gravedad. Actualmente se cuenta con tres sistemas de desarenación, uno al inicio de cada conducción. El desarenador de Santo Domingo – Salado cuenta con dos compartimientos con 8 tolvas de autolavado.

4. *Sistema convencional:* está compuesto por tratamiento primario y secundario. El primario corresponde al tratamiento en el que se remueve una porción de los sólidos suspendidos y de la materia orgánica del agua residual. Esta remoción normalmente es realizada por operaciones físicas como la sedimentación. El efluente del tratamiento primario usualmente contiene alto contenido de materia orgánica y una relativamente alta DBO (Demanda Bioquímica de Oxígeno). El tratamiento secundario corresponde a aquel directamente encargado de la remoción de la materia orgánica y los sólidos suspendidos. En la actualidad la planta convencional tiene una capacidad de 160 l/s. La planta cuenta con dos unidades, cada una conformada por floculadores mecánicos, hidráulicos, sedimentador de placas paralelas, filtración, desinfección y almacenamiento.

5. *Coagulación y Floculación:* la coagulación es la aglutinación de las partículas suspendidas y coloidales presentes en el agua mediante la adición de coagulantes; la floculación es la aglutinación de partículas inducida por una agitación lenta de la suspensión coagulada. El proceso de coagulación se realiza empleando sulfato de aluminio tipo B. Durante la aplicación se realiza una mezcla rápida con el fin de que el sulfato se disuelva totalmente en el agua, luego se inicia el proceso de floculación mediante agitación lenta en 8 floculadores mecánicos, lo que ocasiona la formación de *flocs* (partículas de mayor tamaño).

6. *Sedimentación:* corresponde al proceso en el cual los sólidos suspendidos en el agua se decantan por gravedad, previa adición de químicos coagulantes. Se cuenta con dos unidades de sedimentadores de alta tasa a donde llega el agua después de la floculación, quedando retenidos en el proceso de sedimentación los

flocs ya formados.

7. *Filtración:* corresponde al proceso mediante el cual se remueven las partículas suspendidas y coloidales del agua al hacerlas pasar a través de un medio poroso. El agua sedimentada pasa a través de un lecho filtrante (mixto) conformado por tres capas: antracita, arena, y grava; donde se terminan de retener todas las partículas que le hayan podido quedar al agua, para posteriormente ser desinfectadas.

8. *Desinfección:* corresponde al proceso físico o químico que permite la eliminación o destrucción de los organismos patógenos presentes en el agua. Se cuenta con un sistema de cloración por medio de la utilización de cloro líquido, el cual se encarga de eliminar los microorganismos presentes en el agua, inclusive en las tuberías de distribución del agua potable. La Empresa Multipropósito de Calarcá, implementó un sistema de dosificación, control y neutralización de fugas de cloro, con lo cual, garantiza una adecuada aplicación del producto y, minimiza los riesgos que puede causar tanto al personal que labora en la planta como a la comunidad aledaña.

9. *Almacenamiento:* la planta de tratamiento de agua potable cuenta con cuatro tanques de almacenamiento con una capacidad total de 5155,9 m3.

10. *Control de calidad de agua:* según el decreto 475/98 y el reglamento técnico de agua potable (RAS) 2000, la Empresa debe realizar un control de la calidad del agua, el cual se lleva a cabo: en las fuentes de abastecimiento a la entrada de la planta de potabilización, durante el proceso de potabilización y, en la red de distribución.

El laboratorio de calidad de agua ha participando consecutivamente en el programa de control de calidad interlaboratorios (PICCAP), liderado por el Instituto Nacional de Salud, recibiendo la autorización para la realización de análisis *organolépticos, físicos, químicos, y microbiológicos* al agua potable; resoluciones No. 3689 de Noviembre 4 de 2004 y, No. 004645 de 15 de diciembre de 2005, emanadas por el Ministerio de la Protección Social. Como control de la

calidad del agua suministrada, se realizan muestreos adicionales con entidades externas, para la verificación del agua entregada; es así, como durante los años de 2004 y 2005, el laboratorio de aguas de la universidad Tecnológica de Pereira (departamento de Risaralda – Colombia), realizó la caracterización del agua en tres puntos de la red de distribución, analizando parámetros especiales exigidos por el decreto 475/98, certificando que: "La empresa multipropósito de Calarcá s.a. e.s.p. suministra agua potable de calidad". Certificado Calidad de Agua: 2004 – 2005.

11. *Sistema de Macromedición:* la empresa adelantó un completo programa para dotar de macromedición su sistema de acueducto. Dicho programa busca la optimización de instrumentos de medición en: cuencas hidrográficas, tuberías de entrada y salida de la planta de tratamiento de agua potable, en los tanques de almacenamiento, y en la red de distribución.

- **Sistema de distribución de acueducto**

El sistema de distribución en la parte urbana del municipio es a gravedad con una longitud aproximada de 120 kilómetros lineales.

1. *Sistema de Toma de presiones:* con el fin de realizar un seguimiento permanente al comportamiento de las presiones en la red de distribución, se realiza la toma de presión en la red de acueducto en 10 puntos en las frecuencias establecidas en el modelo hidráulico.

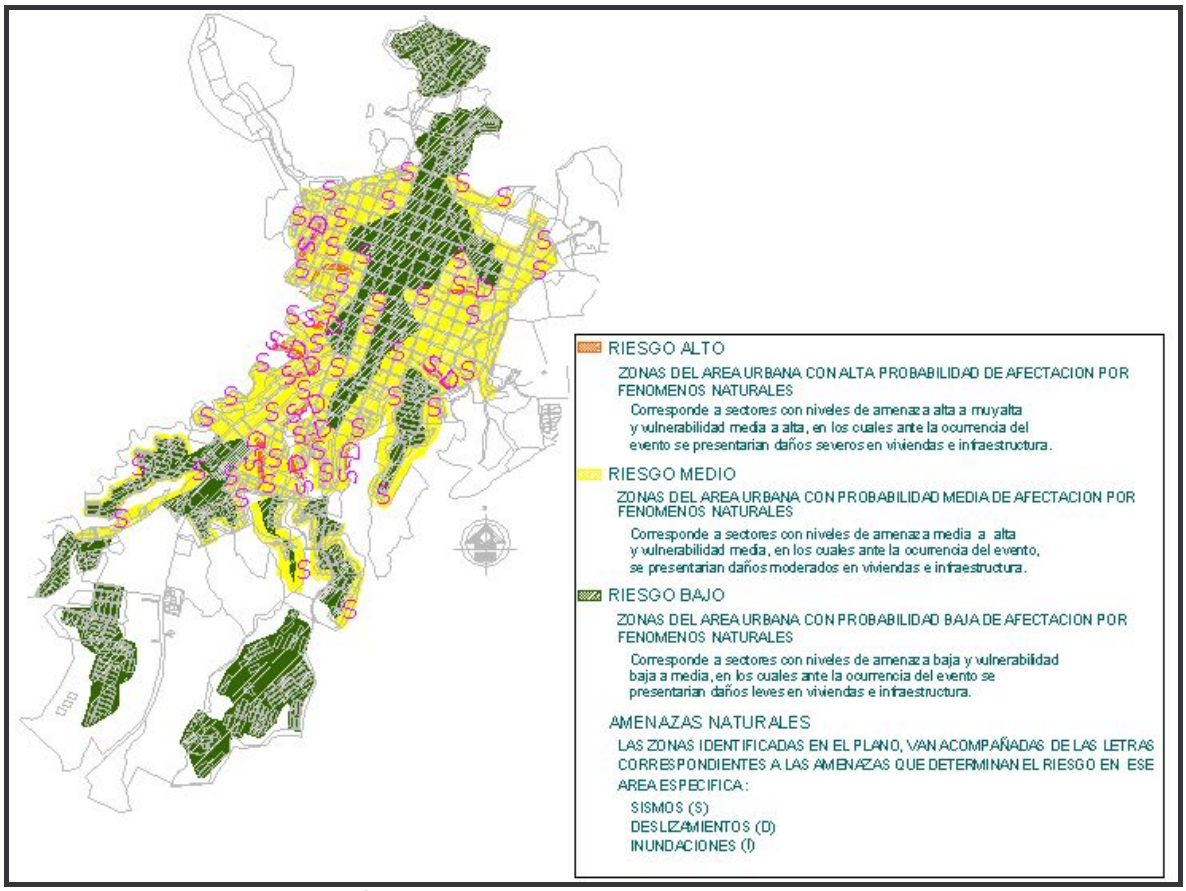
2. *Estaciones reductoras de presión:* se cuenta con 12 puntos de toma de presión tal como se puede apreciar en la Tabla IV.4 más adelante.

IV.3.2. Peticiones, Quejas y Reclamos PQR's

La información básica para el desarrollo del trabajo está basada en los formularios de *Peticiones, Quejas y Reclamos (PQR's)* para el año 2006, que son llevados por parte de la empresa y, en los cuales, se refleja cualquier incidencia que ocurra en la red tanto principal como domiciliaria durante las 24 horas del día.

Adicionalmente, se cuenta con: el modelo hidráulico de la red para este mismo año, datos de mediciones de presión en la red y, un plano de riesgos para el municipio de Calarcá que contiene la caracterización de riesgos y vulnerabilidades de amenazas geológicas; tal como se visualiza en la Figura IV.10.

También se hace uso del Anuario Meteorológico Cafetero (Cenicafé, 2006), en el cual se encuentra la información climatológica utilizada. Uno de los atractivos de las herramientas de minería de datos, es que no es necesario conocer ni intuir una relación previa entre las variables a ser tenidas en cuenta por el modelo, tal como se puede constatar en Babovic *et al.,* 2002. Por esto hemos utilizado la variable climatológica del brillo solar dentro de los datos utilizados en esa tesis, la cual previamente no parecería tener una relación con los resultados obtenidos.



Fuente: INGEOMINAS y Alcaldía Municipal (2003)

Figura IV.10. Plano de riesgo en el Municipio de Calarcá

A continuación, se presenta una definición y descripción de algunos de los términos encontrados en los reportes.

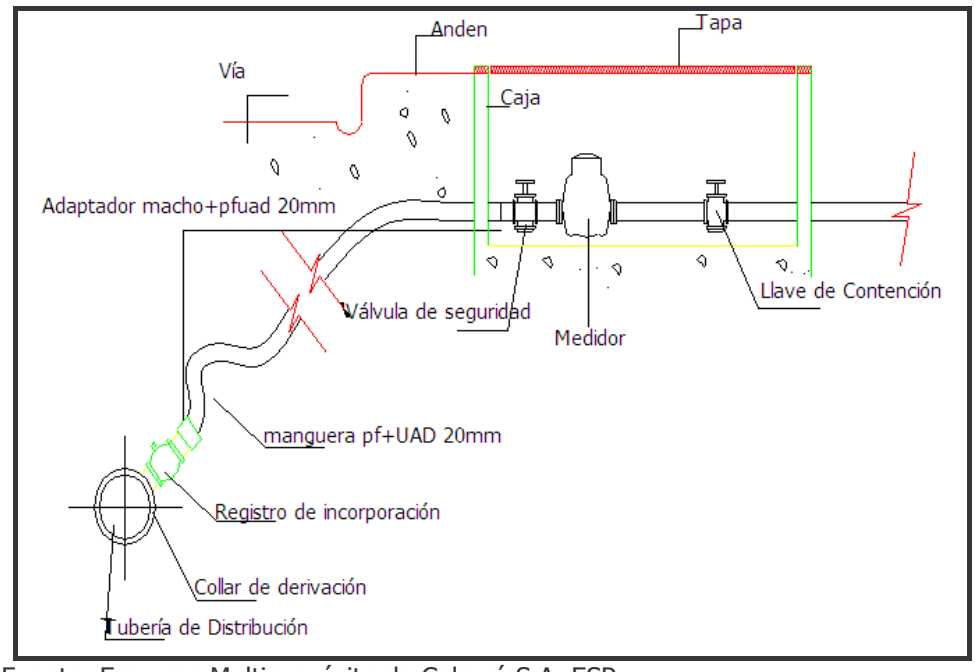
IV.3.2.1. Definiciones

- *Adaptador macho y hembra:* son acoples utilizados para conectar generalmente mangueras de polietileno para acometidas (en diámetros de ½" y 1"). Su funcionamiento es similar al de un tornillo y una tuerca. Es el caso de los micromedidores que vienen con terminal macho, por lo cual, se requiere un adaptador hembra para empalmarlo a la domiciliaria (tubería).

- *Collarín:* es un accesorio que se emplea para acoplar una acometida a una red principal, generalmente vienen de diámetros 2" a ½", 3" a ½", 4" a ½", etc.

- *Universal:* también se refiere a un accesorio generalmente para efectuar empalmes entre un tramo de tubería de PVC y otro.

- *Fichero del medidor:* es la parte del medidor que permite visualizar la medida.



Fuente: Empresa Multipropósito de Calarcá S.A. ESP

Figura IV.11. Esquema de una acometida típica.

- *Pitorra:* es el terminal que tiene el medidor para la conexión con la red domiciliaria; "es una especie de válvula, por donde entra el agua, normalmente es

la que se tapa por sedimentos". Aunque otros fontaneros llaman pitorra a una arandela, una universal, la que une el tubo con el contador, entonces es por la arandela que se produce la fuga.

- *Tubería AC:* tubería de asbesto cemento.
- *Tubería AP:* tubería de hormigón armado AP, CCP, ó ACCP (American Cilinder Concrete Pipe).
- *Tubería HG:* tubería de hierro galvanizado.
- *Tubería PVC:* tubería de cloruro de polivinilo.
- *Tubería PE:* tubería de polietileno.
- *Tubería PAD - PEAD:* tubería de polietileno de alta densidad.
- *Tubería PF:* tubería de plástico flexible.
- *Tubería HF:* tubería de hierro fundido.

IV.3.2.2. Descripción de la base de datos

En este apartado se presenta el análisis descriptivo y pretratamiento de la información utilizada para el desarrollo de la aplicación práctica de esta tesis; y como fundamento del siguiente capítulo de resultados y discusión.

La base de datos utilizada consiste en los registros en papel para el año 2006 de los registros de *Peticiones, Quejas y Reclamos (PQR's)* que contienen información acerca de daños ocasionados en la red, así como el detalle de la inspección técnica y la solución adoptada.

La empresa (Multipropósito de Calarcá S.A. ESP.) recibe reportes de PQR's las 24 horas del día, para esto se tiene habilitado en las oficinas centrales de la empresa el servicio de atención al usuario, en horarios de 7:30 a.m a 12:30 m. y de 1:30 p.m a 6:30 p.m de lunes a sábado, adicionalmente, se cuenta con una

línea de atención al usuario habilitada las 24 horas del día durante todo el año en la que se reciben reportes.

Para el año 2006 se tienen un total de 846 registros de PQR's. En este reporte se consigna la fecha de presentación, el nombre de la persona que presenta el PQR, la dirección de ubicación del incidente, la descripción acerca del PQR dada por la persona que lo presenta, la visita o concepto técnico y la solución adoptada, y la fecha y el nombre del funcionario que realiza la solución, tal como se visualiza en la Figura IV.12.

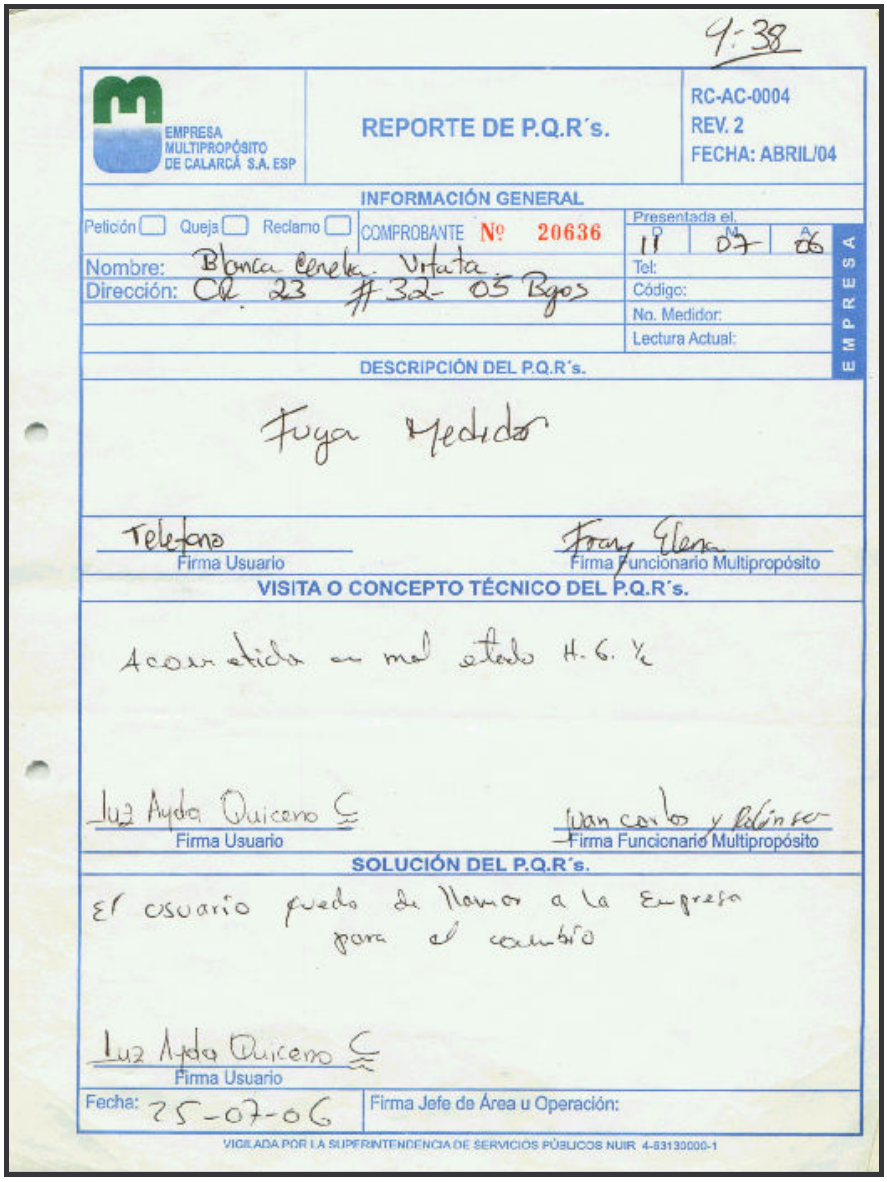


Figura IV.12. Formato de un PQR

Esta información fue digitada en una base de datos para su posterior

tratamiento. En la Figura IV.13 se puede visualizar el número de reportes PQR's por mes para el año 2006, presentándose en el mes de mayo el mayor número de incidencias; así como, al final del año, en los meses de noviembre y diciembre, una menor cantidad de avisos sobre daños en la red.

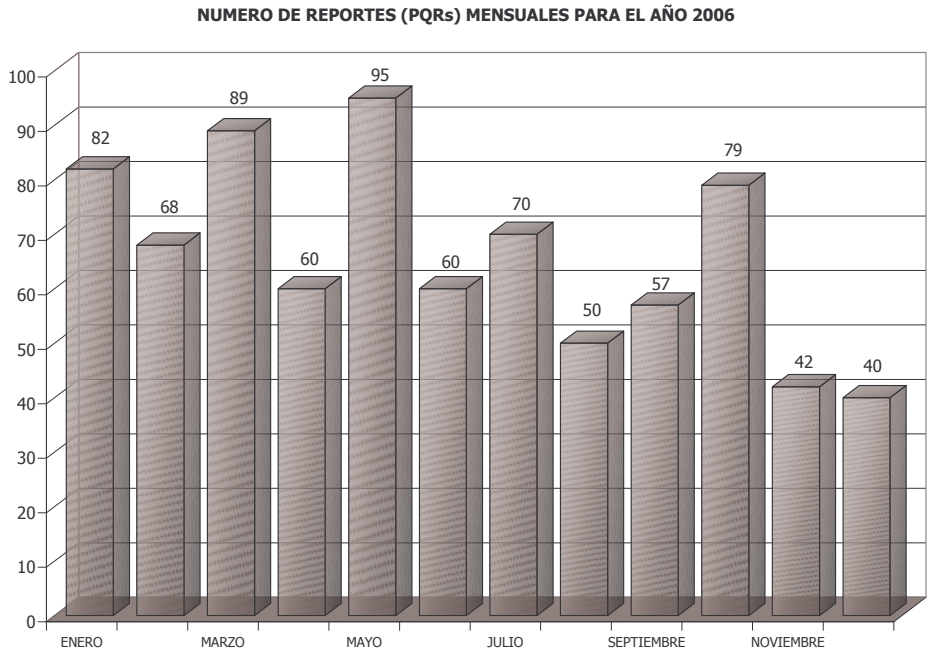


Figura IV.13. Reportes PQR's para el año 2006

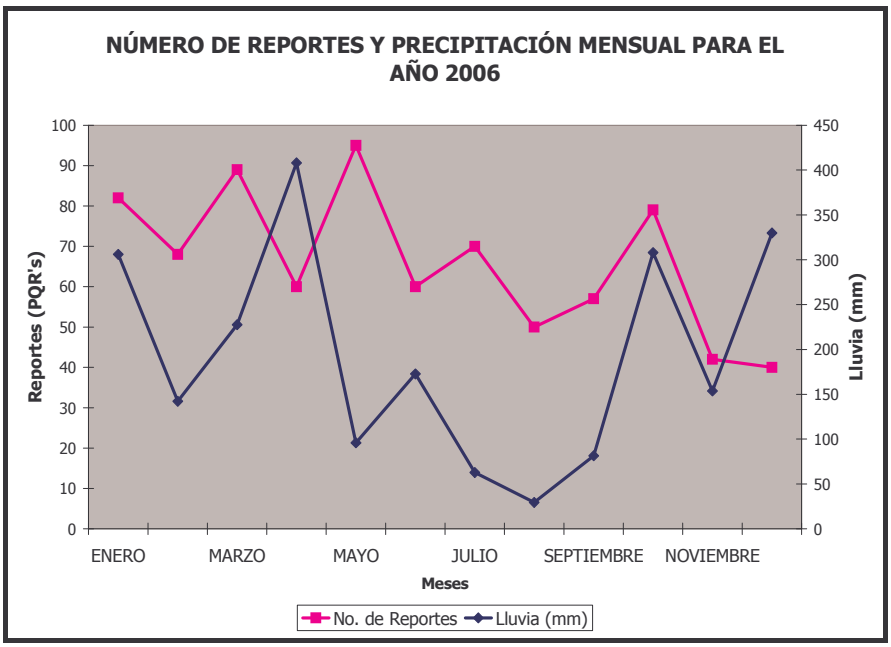


Figura IV.14. Número de reportes PQR's y Precipitación mensual

En esta figura no se aprecia una relación de dependencia directa entre las

precipitaciones mensuales para el año 2006 en la estación La Bella, ya que el mes de mayor precipitación (abril) no corresponde al mes con mayor número de incidencias (mayo), sino, al contrario, es uno de los meses con menor cantidad de precipitación, Figura IV.14. Sólo enero, que en este año tuvo un alto nivel de precipitaciones comparado con su serie histórica, presenta también un elevado número de reclamaciones por daños. No obstante se aprecian ciertos periodos de correspondencia entre ambas variables.

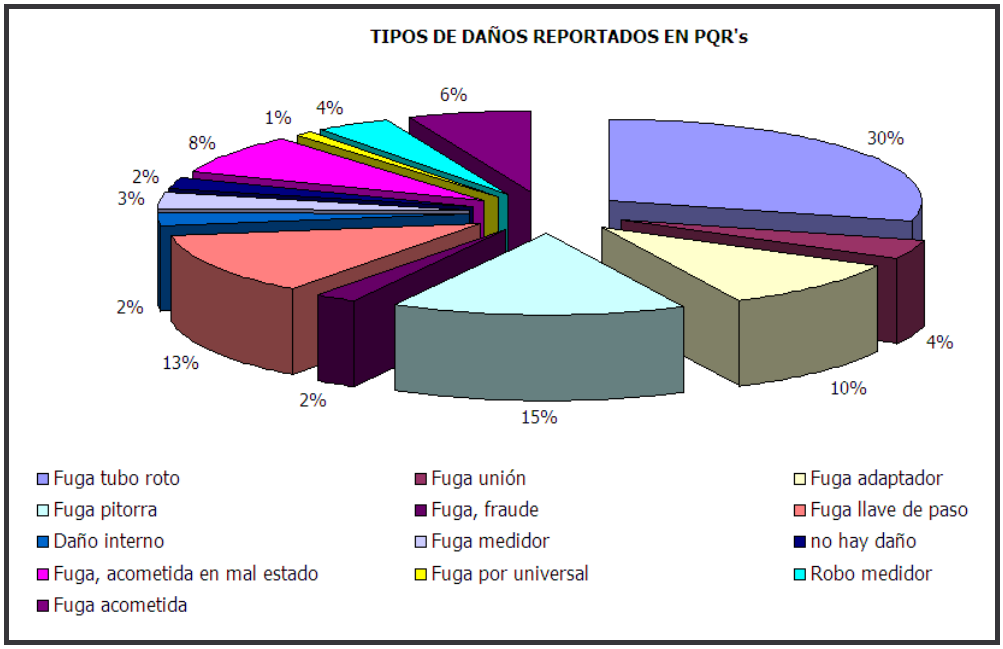


Figura IV.15. Tipos de Daños Reportados

En cuanto a los tipos de daños reportados, tal como se puede visualizar en la Figura IV.15, un 30% corresponden a roturas en tuberías, seguido de fugas en las pitorras con un 15%, y fugas en las llaves de paso con un 13% de reportes. Las fugas debidas a fraudes representan un porcentaje bajo (2%), entendiéndose por fraude a reconexiones ilegales después de suspender el servicio. El porcentaje de daños internos tiene igualmente un valor bajo, correspondiente al 2% de PQR's. Como se ha mencionado anteriormente, el sistema de abastecimiento presenta deficiencias en cuanto a, la vejez de ciertos materiales que ya han cumplido su vida útil, y a problemas de insuficiencia hidráulica para ciertos diámetros; por tanto, esto podría tener incidencia en la cantidad de fugas presentes en los reportes, aunque, por otra parte, se tiene un mayor número de incidencias en las tuberías plásticas que en las metálicas o las de asbesto

cemento, siendo estas dos últimas más antiguas que las primeras. Una debilidad del trabajo, que se realizó con estos datos, es no poder contar con más información acerca de las tuberías, especialmente en lo referente a su edad, y los tipos de tuberías en cuanto a su presión nominal.

En la Figura IV.16 se visualiza la distribución en cada mes de los tipos de daños reportados, se puede apreciar que, en el mes de enero hay un incremento en la cantidad de fugas por tubo roto, siendo este un mes que corresponde al período vacacional; también, se puede apreciar, que independientemente de la numerización de daños adoptada se presentan en general bastantes reportes por pérdidas de agua en las acometidas, factor que influye directamente en la buena gestión tanto técnica como económica del sistema; y que, comparado con el fraude detectado por sustracción ilegal del líquido, afecta en mayor proporción las fugas ocasionadas al sistema.

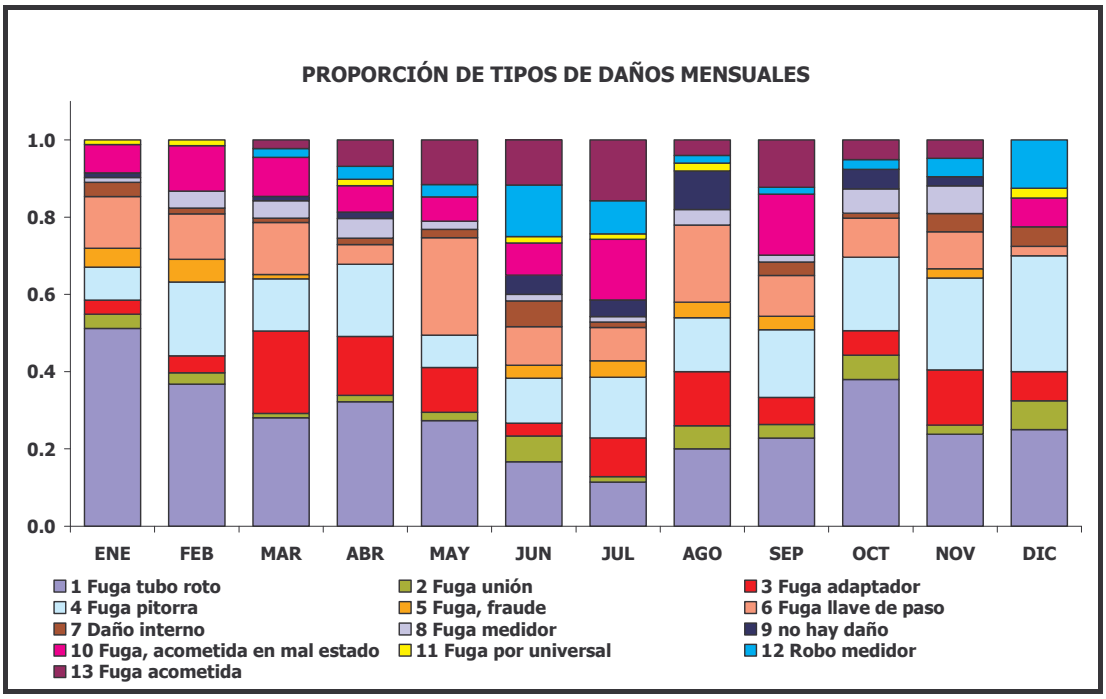


Figura IV.16. Relación de tipos de daños reportados en cada mes

Igualmente, la proporción de daños internos reportados, es bastante menor a la correspondiente de daños reportados en las acometidas en general. Los meses de junio y julio presentan un comportamiento típico de mayor problemática dentro de la gestión técnica del sistema, ya que se puede apreciar una mayor

variedad de tipificaciones de daños, reduciéndose las fugas por tubos rotos.

En cuanto a la distribución horaria de los reclamos se puede visualizar en la Figura IV.17 que la mayor parte de los reclamos se presentan en la mañana entre las seis y las doce del mediodía.

Por redes, la mayor parte de PQR's corresponden a la red domiciliaria (799) por (32) de la red principal (Figura IV.18). El mayor porcentaje de averías por diámetro en la red, corresponde a los de media pulgada (12.7 mm) con casi un 60% de reportes, seguido por tuberías de 16 mm con un 20.6% de reportes. En la Figura IV.19 se pueden visualizar las distribuciones de daños según los diámetros.

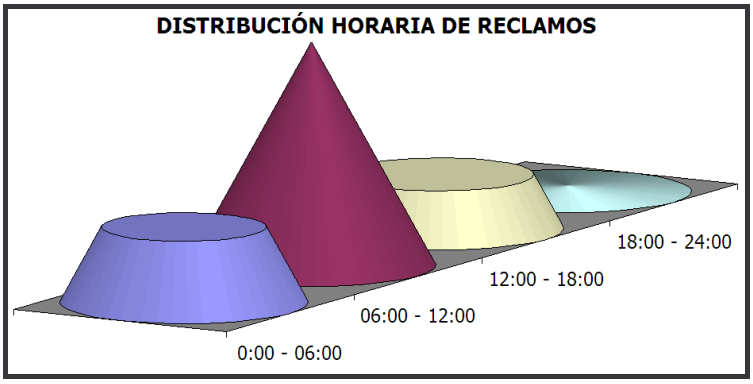


Figura IV.17. Distribución de los reclamos en el día

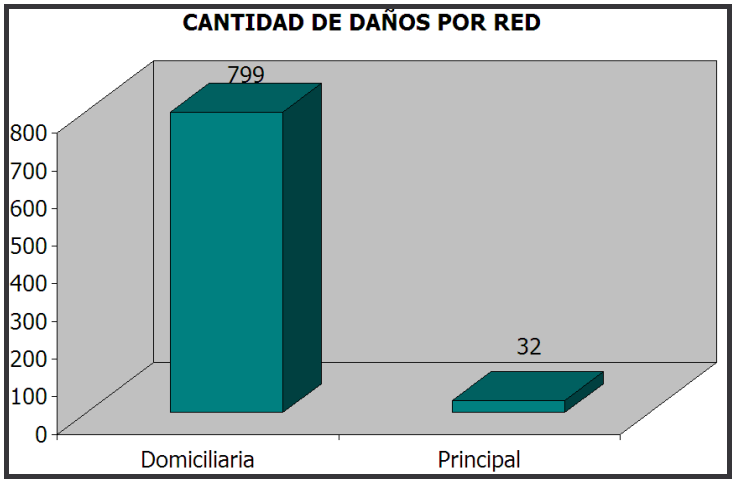


Figura IV.18. Daños por red

En cuanto a los materiales reportados (no siempre se completa todo el formulario para cada PQR), un 41.7% de los reportes corresponden a tuberías de PVC, seguido por un 37.4% de tuberías de PE, un 20.4% de tuberías de HG, y un

0.4% de tuberías de HF; tal como se visualiza en la Figura IV.20.

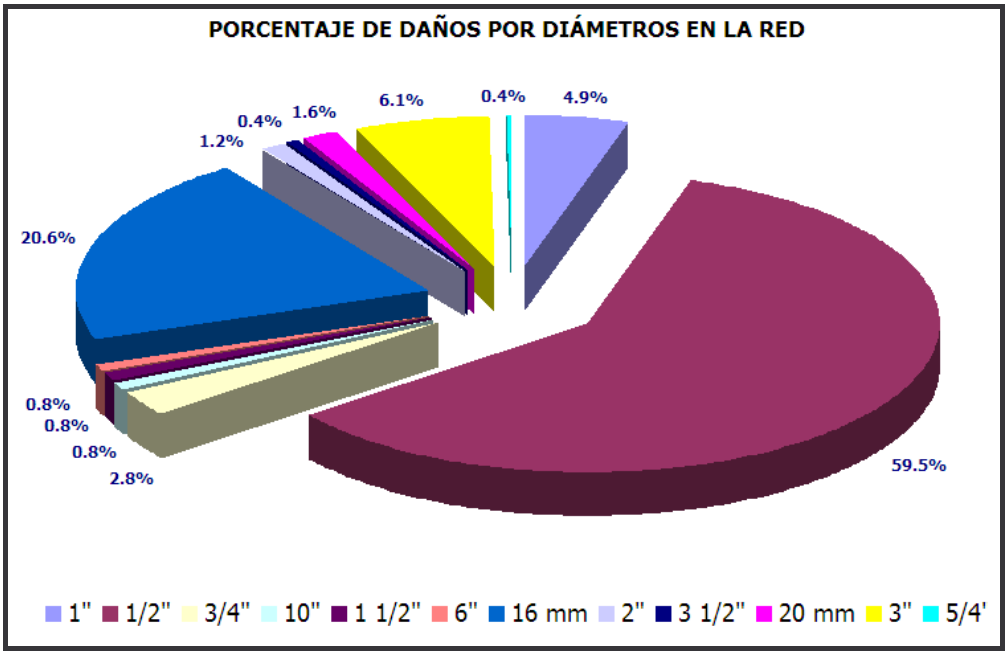


Figura IV.19. Distribución de daños por diámetros

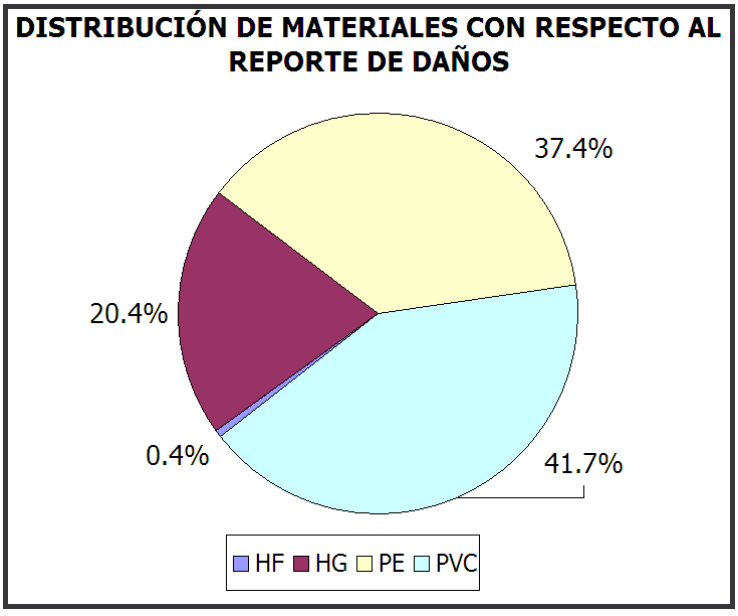


Figura IV.20. Distribución de materiales

El que se vea más afectada la red domiciliaria que la principal puede ser indicio de una gestión más hacia el mantenimiento de la segunda; aunque se requeriría de mayor información para poder corroborar esta afirmación. No obstante, la distribución de diámetros y materiales para los daños reportados de la red dejan vislumbrar que el mantenimiento de las redes domiciliarias, ya sea por que se requieren cambios debido a su edad, o problemas derivados de su

instalación, debe ser mejorado para disminuir los reportes.

En relación con la gestión de los reportes, el 69.5% de los registros de PQR's son atendidos y solucionados el mismo día y, el 16.6% por ciento de los reportes se atiende y soluciona con una demora de un día (Figura IV.21); cabe mencionar al respecto que se considera como erróneo el dato de valores menores a un día (diferencia entre la fecha en que se recibe el reporte y la fecha de solución); del total de reportes se solucionan una cantidad de 768 (Figura IV.22); en la Figura IV.23 se visualiza la conformación de cuadrillas (trabajadores de campo de la empresa), que es tomada como una de las variables para el análisis.

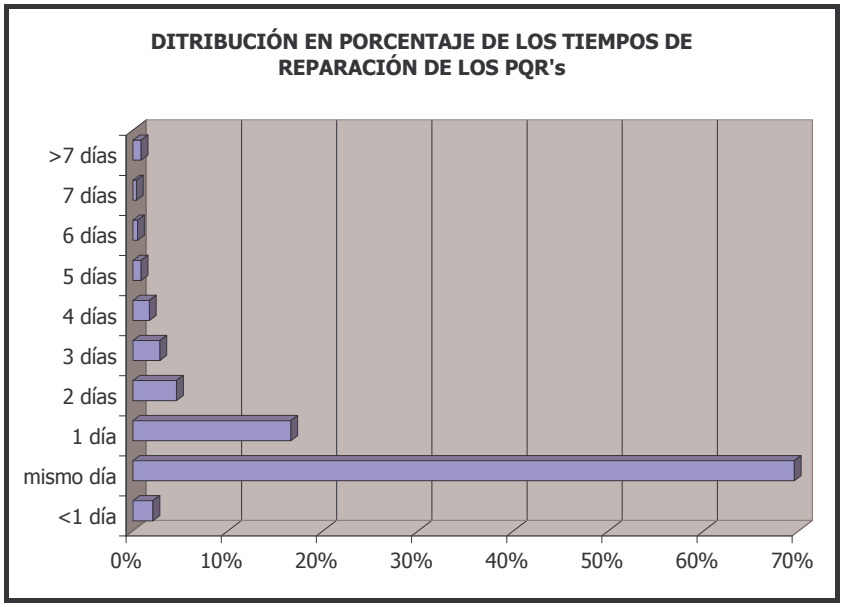


Figura IV.21. Tiempos de reparación

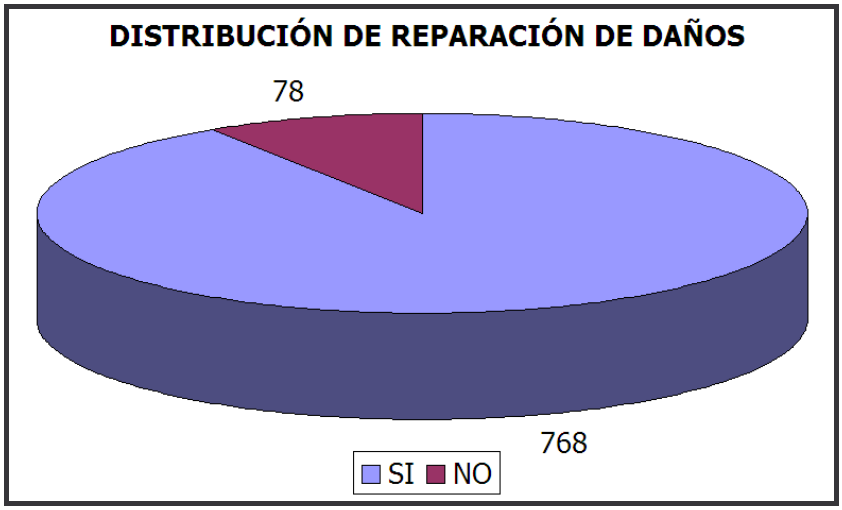


Figura IV.22. Reparación de daños

En cuanto al nivel de riesgo debido a fenómenos naturales, un 0.71% de las localizaciones de los registros de PQR's corresponden a zonas con un nivel alto, un 38.77% a zonas con un nivel medio, y un 52.72% de los datos están ubicados en zonas de riesgo bajo. Un 7.8% de los datos no está ubicado en ninguna zona de riesgo.

Por amenazas, el 0.71% de los registros de PQR's están ubicados en zonas de amenaza de sismo y deslizamiento y, un 38.77% en zonas de sismos. El restante 60.52% de los registros no se encuentran ubicados en zonas de amenaza natural, tal como se visualiza en la Figura IV.24.

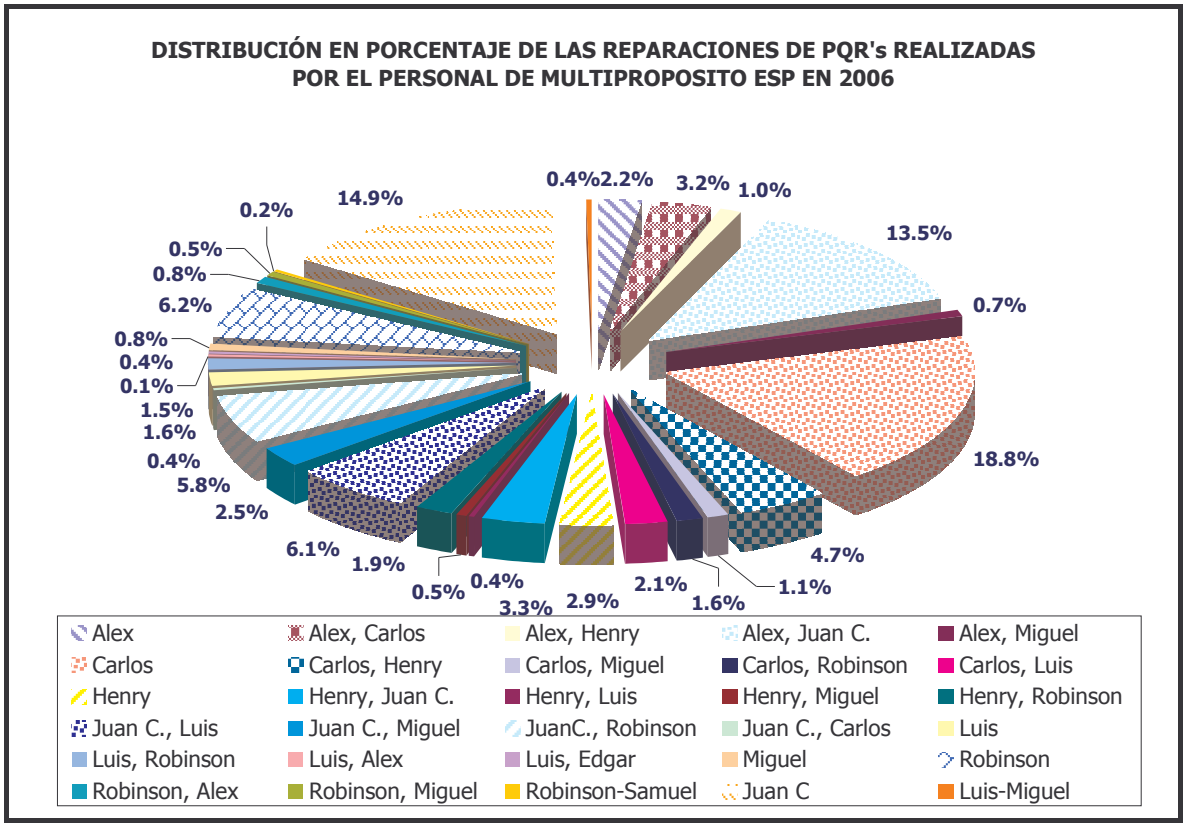


Figura IV.23. Conformación de cuadrillas de trabajo

La posibilidad de riesgo sísmico en la que se encuentra ubicada gran parte de la red de abastecimiento de agua potable del municipio, soporta un extra en cuanto a la problemática de gestión por parte de la empresa en el caso de que se llegara a presentar tal evento.

En la Figura IV.25 se visualiza el patrón de consumo diario de agua en el abastecimiento del municipio de Calarcá, de acuerdo con el modelo hidráulico del

sistema. La distribución del consumo es unimodal y se aprecia que las horas de mayor consumo corresponden a las horas centrales del día, con un factor de aproximadamente 1.4 respecto al consumo medio diario; por el contrario las horas de menor consumo corresponden a las del inicio de la madrugada con un factor de aproximadamente 0.6 respecto al consumo medio diario. También se aprecia cierta relación entre este patrón de consumos y la distribución de reclamos diarios vistos en la Figura IV.17.

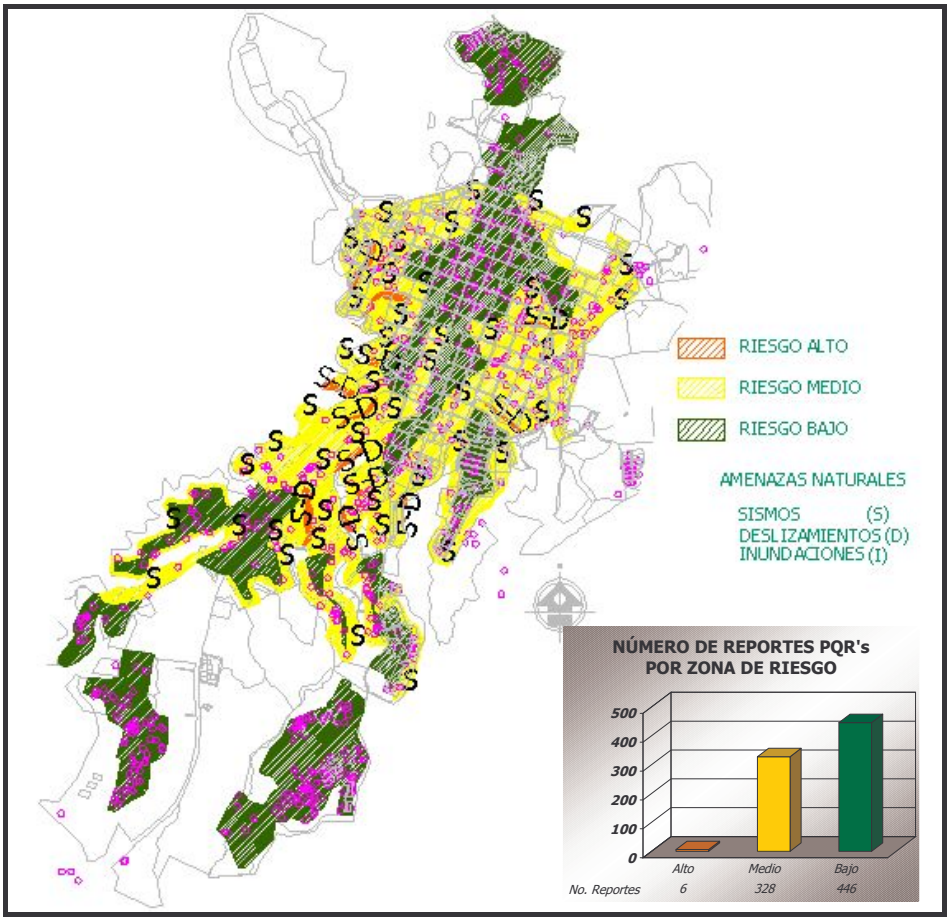


Figura IV.24. Ubicación de los PQR's en las zonas de riesgo por fenómenos naturales

En cuanto a los puntos de toma de presiones se tienen 12 sitios en la red tal como se visualiza en la Figura IV.26 y se describe en la Tabla IV.4. La nomenclatura de los nodos corresponde a la utilizada en el modelo hidráulico. Los datos con los que se cuenta corresponden al año 2007 para los días 18, 19, 28, y 29 de marzo; 27, 28, 30, y 31 de marzo; 24, 25, 27, y 28 de abril; 29, 30, y 31 de mayo; 25, 26, 27, y 28 de junio; 26, 27, 28, y 29 de julio; y 22, 23, 29, y 30 de agosto.

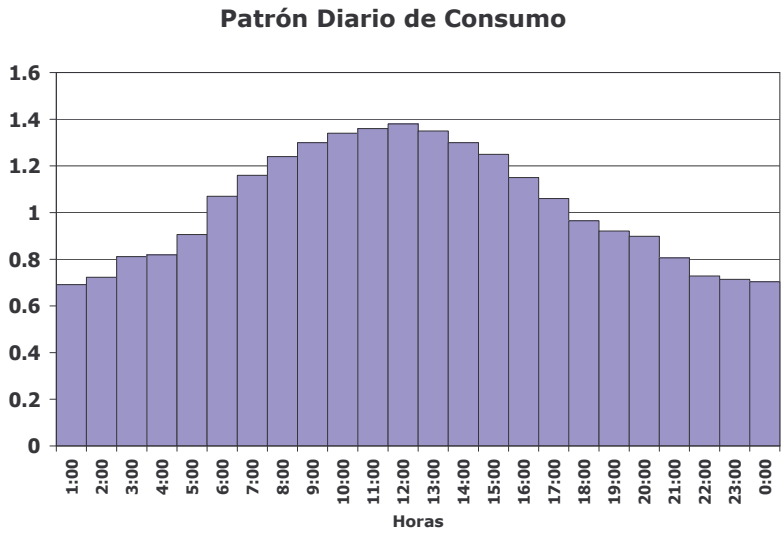


Figura IV.25. Patrón diario de consumo de agua en el municipio de Calarcá

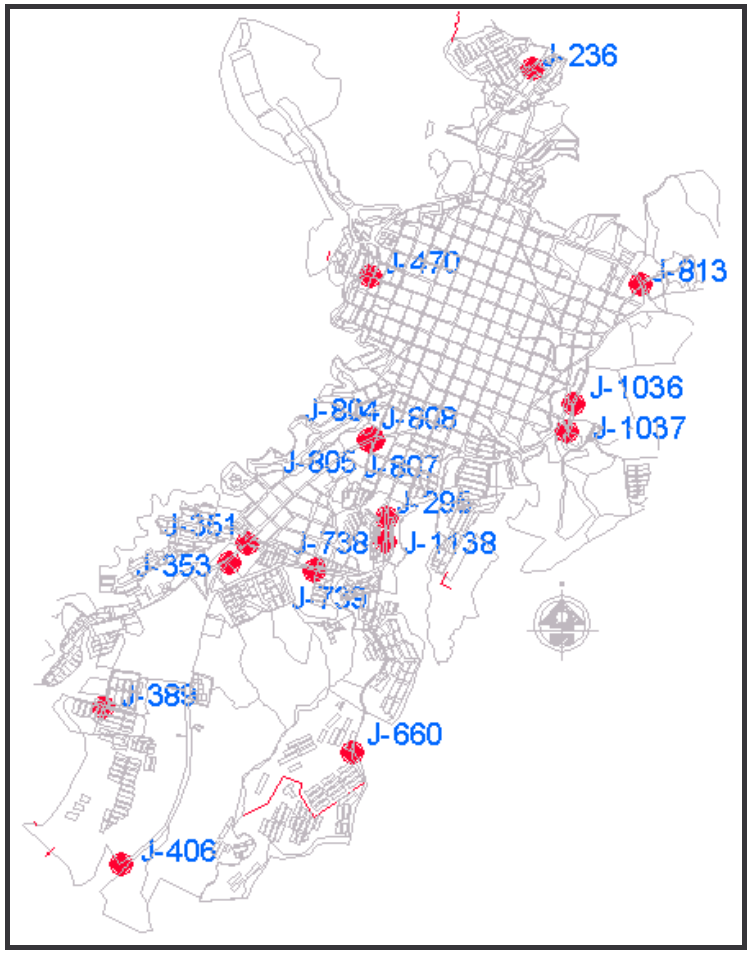


Figura IV.26. Ubicación de los puntos medidores de presión en la red

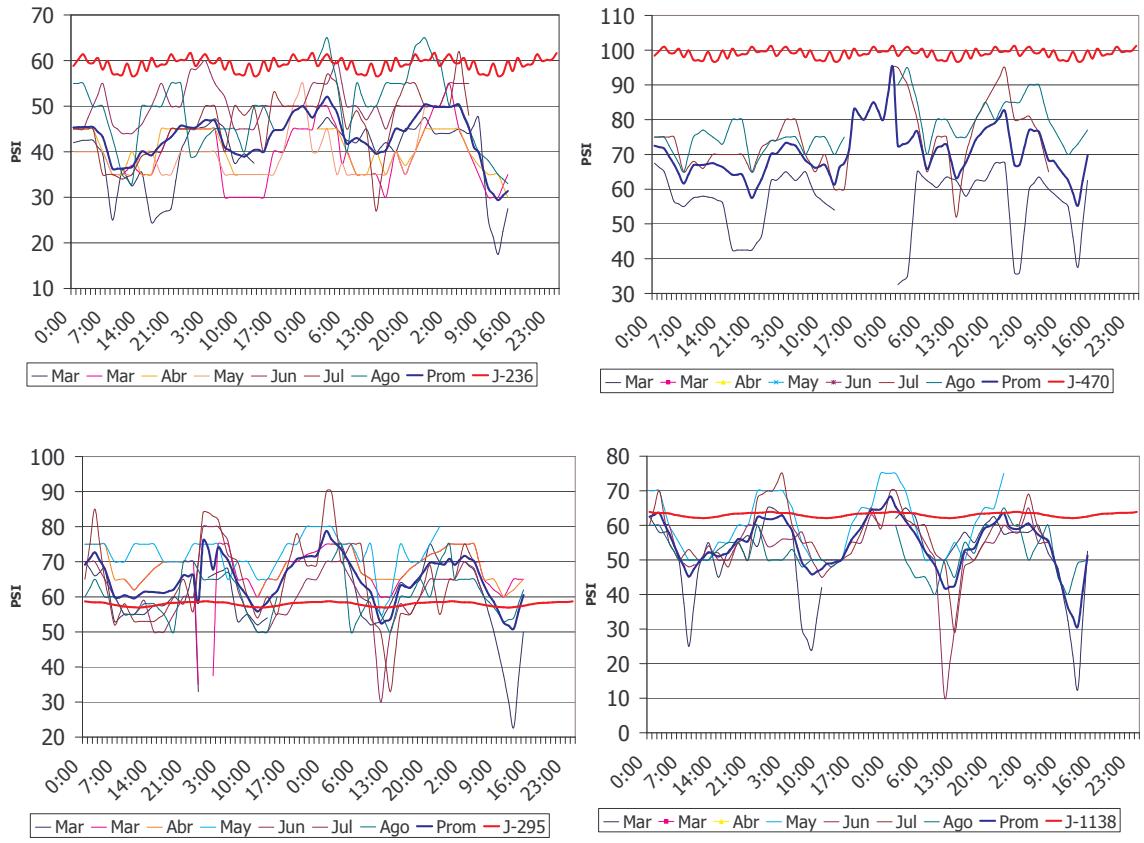
Se presenta un mayor control de la red en cuanto a la toma de presiones hacia la parte baja tal como se puede visualizar en la Figura IV.26. Aguas arriba y aguas abajo hace referencia al drenaje natural (topografía) del terreno, que es

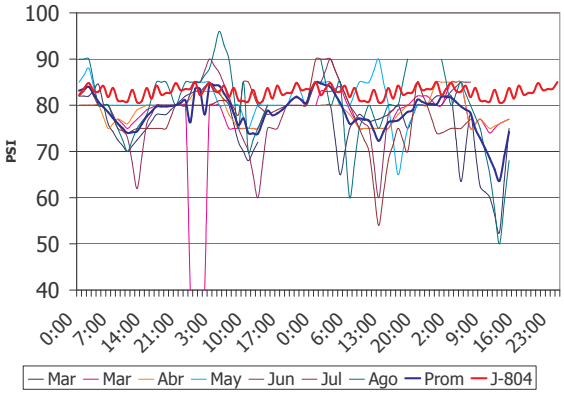
aproximadamente de norte a sur. En el nudo J-236 las tuberías que confluyen tienen un diámetro de 75 mm; en el J-470 son de 25 mm y 38 mm respectivamente; en el J-295 dos tuberías de 150 mm y 50 mm respectivamente, y un tramo aguas arriba de una válvula reductora de presión de 150 mm tarada en 50 m.c.a; en el J-1138 aguas abajo de la válvula reductora de presión tarada a 50 m.c.a y una tubería de 150 mm; los nodos J-804, J-805, J-807 y J-808 se encuentran aguas arriba y aguas abajo respectivamente de dos válvulas reductoras de presión taradas a 35 m.c.a., y tramos de tubería de 200 mm y 150 mm de diámetro. En el nodo J-351 confluyen dos tuberías, una de 250 mm y otra de 50 mm de diámetro, y dos tramos de 250 mm y 100 mm con válvulas reductoras de presión taradas a 35 m.c.a.

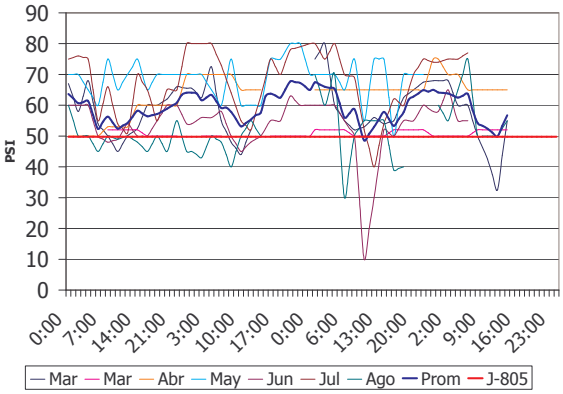
Punto	***Localización***	***Observaciones***	***Nodo***
1	Calle 49 Carrera 26		236
2	Calle 38 Carrera 31		470
3	Carrera 26 Calle 27	Aguas arriba	295
	Carrera 26 Calle 27	Aguas abajo	1138
4	Av. Colón Carrera 27 6”	Aguas arriba	808
	Av. Colón Carrera 27 6”	Aguas abajo	807
	Av. Colón Carrera 27 8”	Aguas arriba	804
	Av. Colón Carrera 27 8”	Aguas abajo	805
5	Entrada Veracruz		351
6	Av. Colón Cementerio	Aguas arriba	351
		Aguas abajo	353
7	Barrio Ecomar		389
8	Milciades Segura		406
9	Luis Carlos Galán	Aguas arriba	738
		Aguas abajo	739
10	Entrada Llanitos Gualará		660
11	Carrera 17 Calle 34	Aguas arriba	1036
		Aguas abajo	1037
12	Calle 41 Carrera 16		813

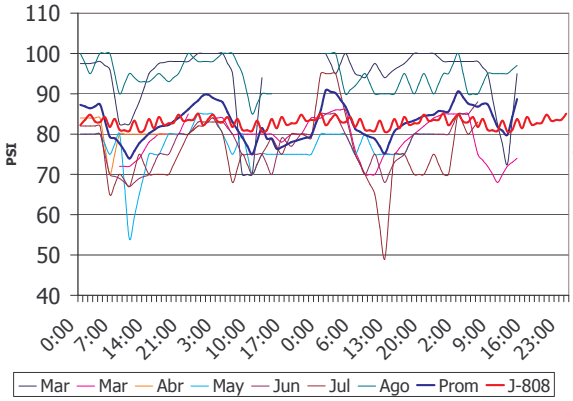
Tabla IV.4. Puntos de toma de presión en la red

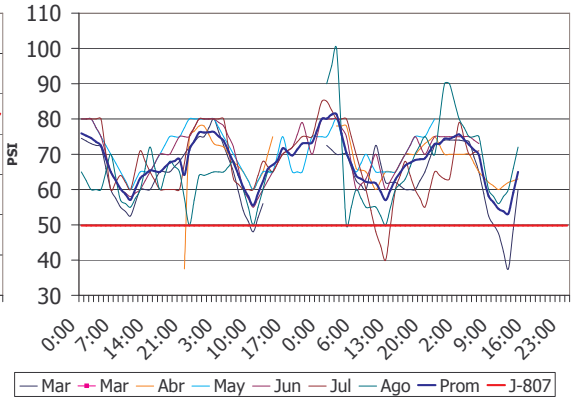
En el J-353 confluyen el tramo de tubería de 250 mm con válvula reductora de presión tarada en 35 m.c.a., y dos tuberías de 150 mm de diámetro. El J-389 tiene tres tuberías adyacentes de 75 mm de diámetro; el J-406 es un nodo terminal de tubería de 150 mm de diámetro; el J-738 tiene tres líneas adyacentes, dos de 150 mm y una de 100 mm de diámetro con válvula reductora de presión tarada a 35 m.c.a.; el J-739 tiene adyacentes dos tramos de tubería de 75 mm y 100 mm de diámetro, y una válvula en el tramo de 100 mm tarada a 35 m.c.a. El nodo J-660 tiene dos tramos adyacentes de 150 mm de diámetro; el nodo J-1036 tiene dos tramos adyacentes de diámetro 75 mm uno con válvula reductora de presión tarada a 40 m.c.a; y el J-1037 se ubica aguas abajo de la válvula anterior y con otro tramo adyacente de 75 mm. El nodo J-813 tiene dos tramos adyacentes de 200 mm de diámetro.

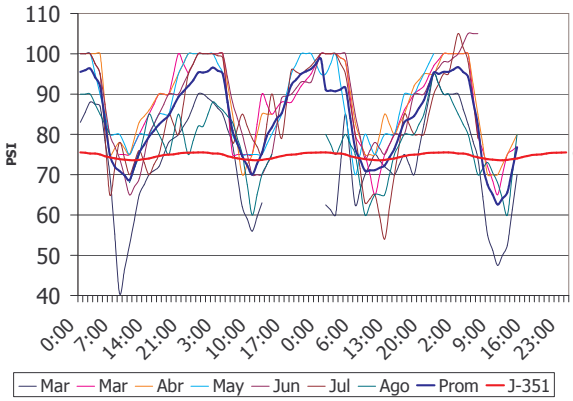


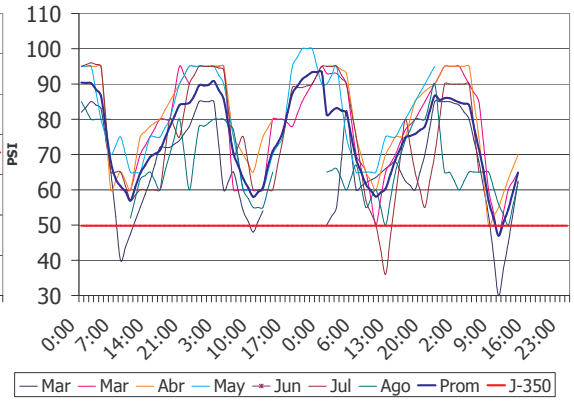

PSI
100
90
80
70
60
50
40
0:00 7:00 14:00 21:00 3:00 10:00 17:00 0:00 6:00 13:00 20:00 2:00 9:00 16:00 23:00
Mar Mar Abr May Jun Jul Ago Prom J-804

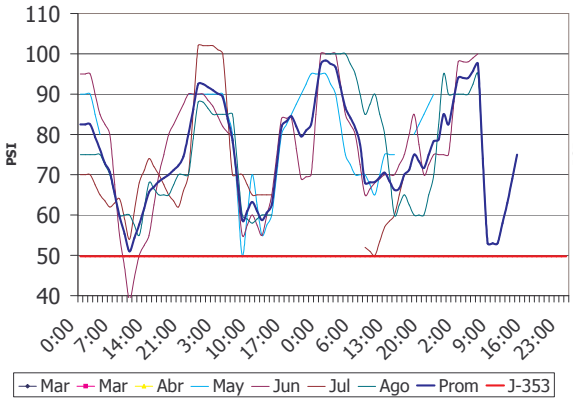

PSI
90
80
70
60
50
40
30
20
10
0
0:00 7:00 14:00 21:00 3:00 10:00 17:00 0:00 6:00 13:00 20:00 2:00 9:00 16:00 23:00
Mar Mar Abr May Jun Jul Ago Prom J-805

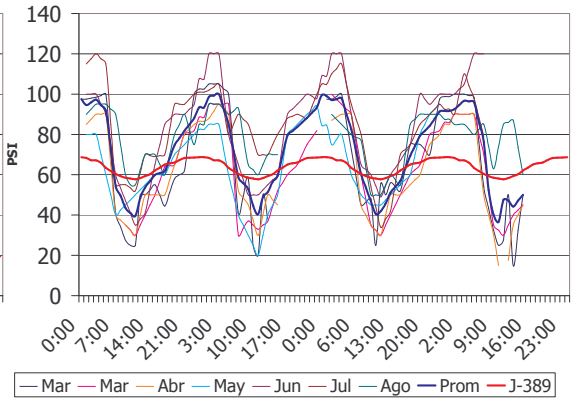

PSI
110
100
90
80
70
60
50
40
0:00 7:00 14:00 21:00 3:00 10:00 17:00 0:00 6:00 13:00 20:00 2:00 9:00 16:00 23:00
Mar Mar Abr May Jun Jul Ago Prom J-808


PSI
110
100
90
80
70
60
50
40
30
0:00 7:00 14:00 21:00 3:00 10:00 17:00 0:00 6:00 13:00 20:00 2:00 9:00 16:00 23:00
Mar Mar Abr May Jun Jul Ago Prom J-807


PSI
110
100
90
80
70
60
50
40
0:00 7:00 14:00 21:00 3:00 10:00 17:00 0:00 6:00 13:00 20:00 2:00 9:00 16:00 23:00
Mar Mar Abr May Jun Jul Ago Prom J-351


PSI
110
100
90
80
70
60
50
40
30
0:00 7:00 14:00 21:00 3:00 10:00 17:00 0:00 6:00 13:00 20:00 2:00 9:00 16:00 23:00
Mar Mar Abr May Jun Jul Ago Prom J-350


PSI
110
100
90
80
70
60
50
40
0:00 7:00 14:00 21:00 3:00 10:00 17:00 0:00 6:00 13:00 20:00 2:00 9:00 16:00 23:00
Mar Mar Abr May Jun Jul Ago Prom J-353


PSI
140
120
100
80
60
40
20
0
0:00 7:00 14:00 21:00 3:00 10:00 17:00 0:00 6:00 13:00 20:00 2:00 9:00 16:00 23:00
Mar Mar Abr May Jun Jul Ago Prom J-389

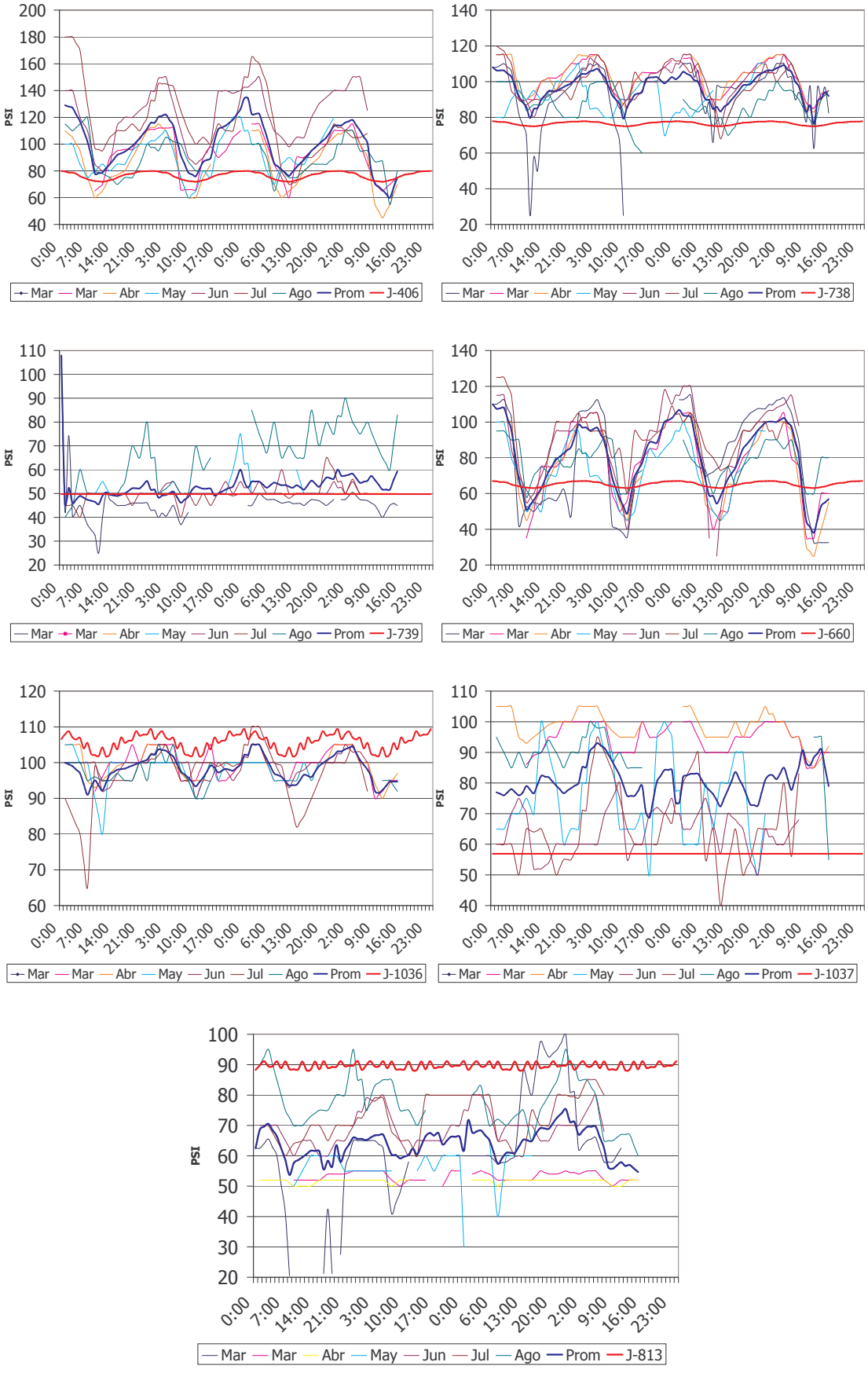


Figura IV.27. Gráficos de las presiones horarias en la red

En los gráficos anteriores (Figura IV.27) se visualizan los valores de las tomas de presión en libras por pulgadas cuadradas (PSI) para las fechas señaladas en cada uno de los nodos, así como los valores obtenidos del modelo hidráulico. De los valores medidos se sacó un promedio para compararlo con el valor que se toma en el modelo. Esta información es utilizada para tener un estimativo de la fiabilidad del modelo hidráulico. En las etiquetas de los gráficos se aprecian los meses de las fechas de medición con su correspondiente color.

Calculando la diferencia entre el valor de presión promedio medido en cada nodo y el valor horario dado por el modelo, se tienen los datos mostrados en la Tabla IV.5.

Nodo	J-236	J-470	J-295	J-1138	J-804
Diferencia m.c.a	11.15	20.09	-4.87	6.06	3.08
Nodo	J-805	J-808	J-807	J-351	J-350
Diferencia m.c.a	-6.78	-0.09	-12.65	-6.52	-17.62
Nodo	J-353	J-389	J-406	J-738	J-739
Diferencia m.c.a	-18.09	-6.01	-16.68	-14.25	-2.27
Nodo	J-660	J-1036	J-1037	J-813	
Diferencia m.c.a	-10.94	5.14	-16.68	18.09	

Tabla IV.5. Diferencia de presiones entre el promedio del modelo hidráulico y las medidas en la red en m.c.a.

Se aprecia una diferencia notable en algunos de los puntos de medición, siendo más crítica donde los valores de presión medidos en la red están por debajo de los valores estimados en el modelo hidráulico. Esto puede ser un indicativo del factor mencionado de la edad de los materiales, que ya han cumplido su vida útil y a la insuficiencia hidráulica de algunos tramos de la red.

Por último se tienen valores de índice de agua no contabilizada total para el año 2006, con una producción total en millones de metros cúbicos de 5.14 y facturados de 2.62 millones de metros cúbicos con un índice de 49.03% de agua

no contabilizada, para el año 2007 este valor es de 42.44%.

IV.4. Dificultades

Antes de definir la aplicación final que sustenta las herramientas de aprendizaje, fue importante desarrollar una serie de tareas para identificar las principales dificultades y decidir qué algoritmos utilizar. Los principales problemas identificados fueron:

- Dificultad en la consecución de la información
- Presentación de la información
- Selección de variables
- Pre-procesamiento
- Valoración de resultados

En cuanto a la consecución de la información, aunque está claro que sería de gran ayuda poder contar con el máximo de variables que nos aporten mayor consistencia a las predicciones o descripciones generadas, el tema de los abastecimientos de agua genera bastante sensibilidad entre los entes poseedores de la información, y no es tarea sencilla que se presten a facilitarla en el grado que nos es conveniente. Pero, gracias al apoyo de la empresa Multipropósito de Calarcá, hemos logrado que nos sean facilitados los datos con los que se desarrolla este trabajo. No cabe duda de que entre más información se tenga, se podría a la vez tanto mejor generalizar como explicitar los resultados obtenidos; y ante esto no podemos esconder el hecho de que sería más concluyente si se estudiase un mayor número de periodos de registros de PQR's, pero esta es la información con la que contamos y de la que un primer acercamiento nos permite plantearnos nuevos cuestionamientos acerca de la ruta a seguir. Por otra parte, se podría tener un mayor control monitorizado de la red, lo cual sin duda nos reportaría un gran beneficio para corroborar la bondad del modelo hidráulico.

La presentación de la información es igualmente un factor de problema para la realización de la tesis, más aún cuando las herramientas tecnológicas e informáticas disponibles en la actualidad permiten un manejo más sencillo de la misma. El no contar con una base de datos georeferenciada (SIG) o ficheros de texto con la información de origen, incrementan la posibilidad de error así como el tiempo necesario para su manejo.

En cuanto a la selección de variables, cuando se desarrolla una tarea de descubrimiento de conocimiento a partir de datos, es una tarea esencial la selección de variables a emplear; y es aquí donde radica la importancia del conocedor del dominio del que se está tratando, porque junto con el experto en técnicas de minería de datos, pueden definir qué paradigma es el apropiado a utilizar. Para nuestro caso, esta selección se realizó manualmente, descartando aquella información que no representaba relevancia alguna para el interés de los objetivos propuestos, tales como identificadores y nombres propios de la persona que comunica la incidencia. Para otra información, se numerizaron atributos categóricos para poderlos manejar como etiquetas numéricas, reduciendo el tamaño de la base de datos y para obtener una mejor representación de los resultados obtenidos.

Las tareas de preprocesamiento realizadas para el desarrollo de este trabajo se realizaron de la siguiente forma:

1. Toda la información de los archivos magnéticos de Peticiones, Quejas y Reclamos (PQR´s), a escala mensual para el año 2006 fue digitada en formato de base de datos. Esta información, para los 846 registros totales comprende, la fecha de notificación de la incidencia, la hora en que se notifica, el nombre de la persona que presenta la notificación, la localización, el tipo de red ya sea principal o domiciliaria sobre la que se presenta la incidencia, la descripción del incidente dada por el usuario, el diámetro y material del lugar, el concepto técnico dado por el funcionario de la empresa, el tipo de solución adoptada, la fecha de reparación, y el (los) funcionario(s) que llevan a cabo la reparación, tal como se visualiza en el ejemplo de reporte PQR de la Figura IV.12.

2. Al realizar la digitación de los datos se fueron encontrando y corrigiendo probables errores e inconsistencias en la información. Cabe destacar que de los datos de hora de presentación de la incidencia se dispone del 69.15% de la información, del material el 29.76% y del diámetro el 27.90% de la información.

3. Basándose en la información de ubicación del reporte dado en la ficha, se ubicó sobre el plano de la red cada uno de los puntos donde ocurre una incidencia.

4. Ubicados los puntos se identificó en el modelo hidráulico a qué tramos corresponden cada uno de ellos, para obtener los datos del modelo hidráulico utilizados para la aplicación de las técnicas de minería de datos.

5. Algunos valores de atributos categóricos se numerizaron para poder trabajar de forma práctica los modelos, reduciendo el tamaño de los datos tal como los tipos de daño, el funcionario, etc. Del modelo hidráulico se tomaron valores medios tanto para las pérdidas en cada tramo a lo largo de las 24 horas, como de la presión en cada punto PQR.

6. De la totalidad de la información disponible (184428 valores) únicamente se tiene como faltante un 4%, por tanto, la base de datos con la que se trabaja tiene un tamaño de 177429 valores luego del preprocesamiento de la información. El manejo de la información faltante que corresponde básicamente a materiales y diámetros en los reportes de PQR's, al estar dispersa por toda el área de estudio y corresponder a información que no fue posible obtener, se decidió no tenerla en cuenta para el desarrollo de la tesis. Por otra parte los algoritmos de árboles de regresión y clasificación tienen entre sus ventajas su robustez ante datos faltantes y espurios. No obstante en el primer prototipo que se presenta en el siguiente capítulo, los registros para los cuales no se contaba con toda la información fueron descartados para el modelado, con el inconveniente de la reducción en el tamaño de la base de datos a modelar.

Capítulo V

Resultados y discusión

V.1. Introducción

En este capítulo se resuelve la metodología propuesta para la consecución de los objetivos planteados de ejecución práctica de la tesis. Se presentan los resultados obtenidos después de aplicar esta metodología, y se exhibe una discusión a partir de estos mismos. La metodología seguida para obtener estos resultados es propuesta en el **Capítulo II**, pasos del *KDD*. Para finalizar el capítulo se hace una serie de recomendaciones a la vista de los resultados obtenidos.

V.2. Manejo de la información

En este apartado se siguen los pasos preliminares de una tarea de *KDD* establecidos en el Capítulo II, así como los descritos en la metodología *CRISP-DM* (Chapman *et al.*, 1999), en cuanto a la comprensión del negocio o dominio tratado, y la comprensión y preparación de los datos.

La información contenida en cada uno de los reportes para el año 2006 fue digitada en una base de datos para su tratamiento. En este primer paso se realizó una limpieza de información que se consideró no fiable o poco clara. Esta limpieza se realizó de forma visual, descartando toda aquella que presentaba incongruencias en cuanto a la información escrita, y se realizó el preprocesamiento descrito en el capítulo anterior. En total se tienen 846 registros y 218 campos o atributos.

Con base en la información de la localización del reclamo, se ubicaron sobre un plano de la red de abastecimiento cada uno de los puntos de *Peticiones, Quejas y Reclamos* (PQR's), con lo cual se tiene el posicionamiento geográfico utilizado en el análisis. Adicionalmente, con estos puntos ubicados se obtuvieron los niveles de riesgo y los tipos de amenaza geomorfológicos para cada punto, basándonos en la Figura IV.24.

La información de los diámetros se expresa toda en milímetros (los reportes

PQR's presentan esta información en pulgadas), para tener iguales criterios de comparación respecto a la información que se obtiene del modelo hidráulico de la red. Los conceptos técnicos descritos en los reportes fueron clasificados en 13 tipos diferentes denominados "*Tipo Daño" o "(Damage)"* de la siguiente forma:

1	Fuga tubo roto
2	Fuga unión
3	Fuga adaptador
4	Fuga pitorra
5	Fuga, fraude
6	Fuga llave de paso
7	Daño interno
8	Fuga medidor
9	No hay daño
10	Fuga, acometida en mal estado
11	Fuga por universal
12	Robo medidor
13	Fuga acometida

El tipo de daño 9 *no hay daño* corresponde a que cuando se realiza la visita técnica conforme al reporte de PQR no se encuentra ningún daño.

1	Alex	11	Henry	21	Luis, Robinson
2	Alex, Carlos	12	Henry, Juan C.	22	Luis, Alex
3	Alex, Henry	13	Henry, Luis	23	Luis, Edgar
4	Alex, Juan C.	14	Henry, Miguel	24	Miguel
5	Alex, Miguel	15	Henry, Robinson	25	Robinson
6	Carlos	16	Juan C., Luis	26	Robinson, Alex
7	Carlos, Henry	17	Juan C., Miguel	27	Robinson, Miguel
8	Carlos, Miguel	18	Juan C., Robinson	28	Robinson-Samuel
9	Carlos, Robinson	19	Juan C., Carlos	29	Juan C
10	Carlos, Luis	20	Luis	30	Luis – Miguel

De la misma forma, se han clasificado las cuadrillas de trabajo (obreros de campo) en 30 grupos. Estos grupos se conformaron tomando como base la

información suministrada en los formularios de PQR's. El objetivo de este agrupamiento es el de buscar la existencia de posibles relaciones entre los reportes de daños y las cuadrillas de trabajo.

Como información adicional, se obtuvieron registros de precipitación para el año 2006 de otras dos estaciones aparte de la estación La Bella mencionada en el Capítulo IV, que aunque se encuentran ubicadas al sur de dicha estación alejadas del casco urbano pero en términos municipales de Calarcá, y ante la escasa información pluviométrica con la que se cuenta, son utilizadas como ejercicio para buscar si existe alguna relación entre éstas con el funcionamiento de la red, de acuerdo al modelo utilizado. Estas estaciones se presentan a continuación.

Estación	*Latitud (N)*	*Longitud (W)*
El Jardín	4º28′	75º42′
Quebradanegra	4º27′	75º40′

Del modelo hidráulico de la red se obtuvo, para cada tramo en el que se ubican los puntos PQR's, el material de tubería del tramo, el diámetro, la longitud, la rugosidad, el caudal circulante, la pérdida en el tramo, las demandas inicial y final en el tramo al igual que la presión en cada extremo del tramo, con lo cual se hizo un estimativo promedio del valor de la presión en cada punto PQR.

V.3. Selección de datos

Como primer paso para la selección de datos relevantes, se realizó gráficamente la visualización de las relaciones entre los diferentes atributos con el fin de observar posibles relaciones entre estos. Como se ve en las figuras a continuación, no se aprecia una relación clara entre las diferentes variables, especialmente entre el Tipo de Daño (Damage) y el resto de variables; los caudales horarios presentan cierta correlación entre si debido a la variabilidad impuesta por la curva de modulación en la distribución horaria de los caudales en el sistema a lo largo del día. Se ha elegido de acuerdo con la información disponible el campo *tipo de daño* como la variable objetivo, ya que es una

referencia al momento de intentar plantear un modelo de gestión de la red. Si se logran relacionar los daños que se producen en la red, teniendo en cuenta variables reales del día a día de funcionamiento de la misma, se puede generar un programa de gestión que permita establecer entre otros, los correctivos a tomar para que no se vea afectado el cliente durante el trascurso de un evento de daño o fallo.

En la Figura V.1 se visualiza la dispersión de la nube de puntos correspondientes a los datos de la variable tipo de daño (damage) en su versión numerizada y los correspondientes a las variables diámetro, material descrito en los reportes de PQR's, y la variable binaria corrección del daño. No se observa correlación entre las diferentes variables, tendiendo más a notarse dispersión entre las variables *tipo de daño* y los diámetros y materiales; para la variable binaria de si o no se corrige el daño se aprecia que para una de las respuestas se cuenta con todos los tipos de daño mientras que para la otra no. También se ve en esta figura que no se cuenta con la totalidad de la información (faltante) para los diámetros y materiales.

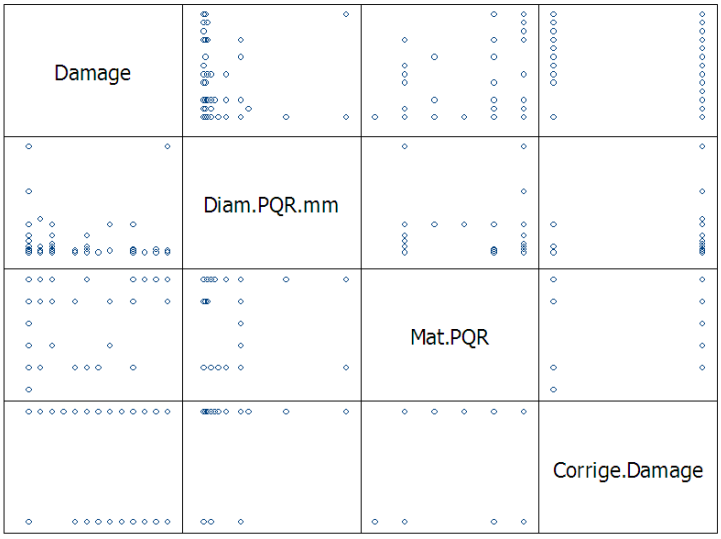


Figura V.1. Diagrama de puntos de las variables Tipo de daño, Diámetro en el PQR, Material en el PQR, Corrección de daño

En la Figura V.2 se visualiza la relación entre el campo tipo de daño (numerizada) y el campo diámetro reportado. Se ve más claramente la dispersión de diámetros para cada uno de los tipos de daño. Para los tipos de daño *fuga por*

pitorra y *no hay daño* no se tiene la información del diámetro. El tipo de daño *fuga medidor* sólo cuenta con dos registros para el diámetro en los reportes.

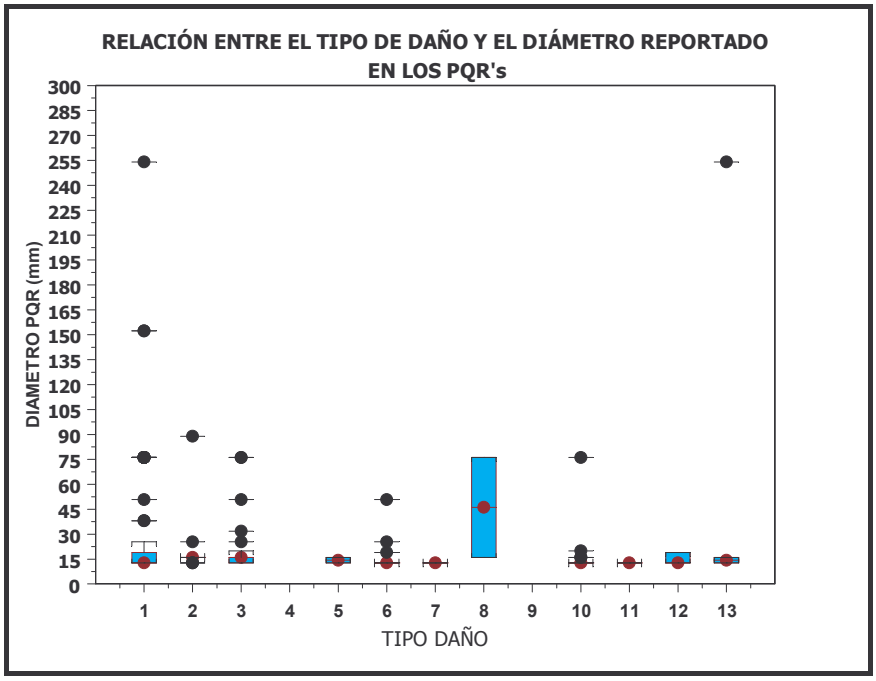


Figura V.2. Relación entre el tipo de daño y los diámetros de los reportes PQR's

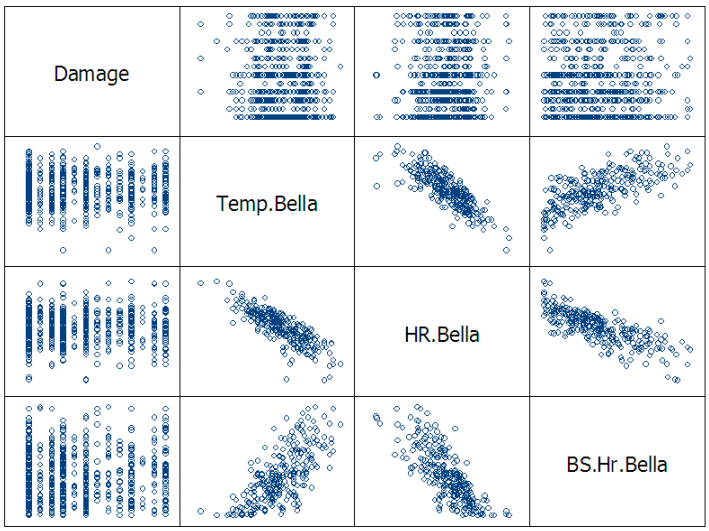


Figura V.3. Diagrama de puntos de variables Tipo de daño, Temperatura, Humedad relativa y Brillo solar en la estación La Bella

En la Figura V.3 se visualiza el diagrama cartesiano entre las variables tipo de daño (numerizada) y las variables climatológicas correspondientes a la temperatura, humedad relativa, y brillo solar en la estación La Bella. La nube de puntos de estas variables climatológicas muestra cierta correlación entre ellas, y una gran dispersión de los tipos de daño con respecto a estas variables

climatológicas.

Por tanto, se observa en la Figura V.4 la relación entre las variables tipo de daño (numerizada) y la temperatura en la estación La bella. Se ve cierto grado de dispersión en la distribución de datos para cada uno de los tipos de daños con distribuciones poco uniformes. El tipo de daño 5 *fuga por fraude* ocurre para temperaturas un poco más elevadas.

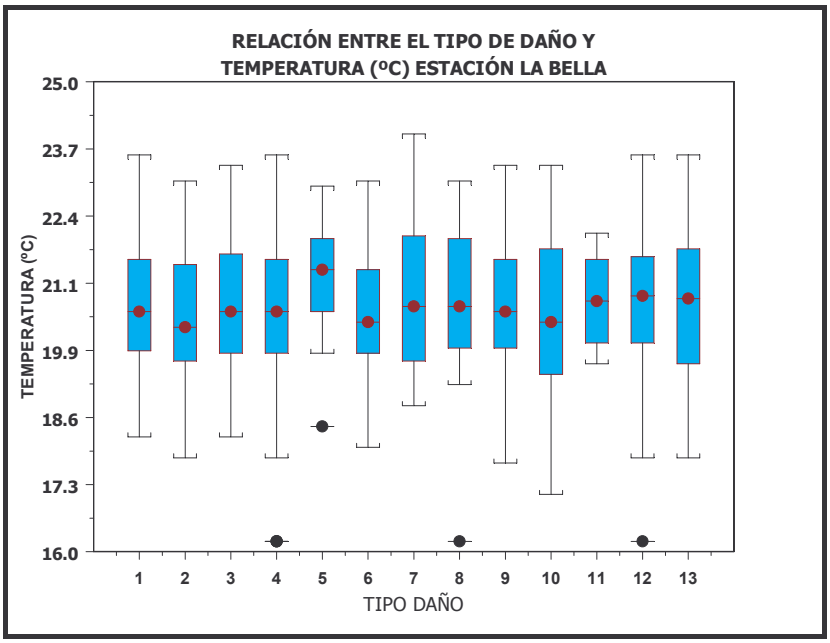


Figura V.4. Relación entre el tipo de daño y la temperatura en la estación la Bella

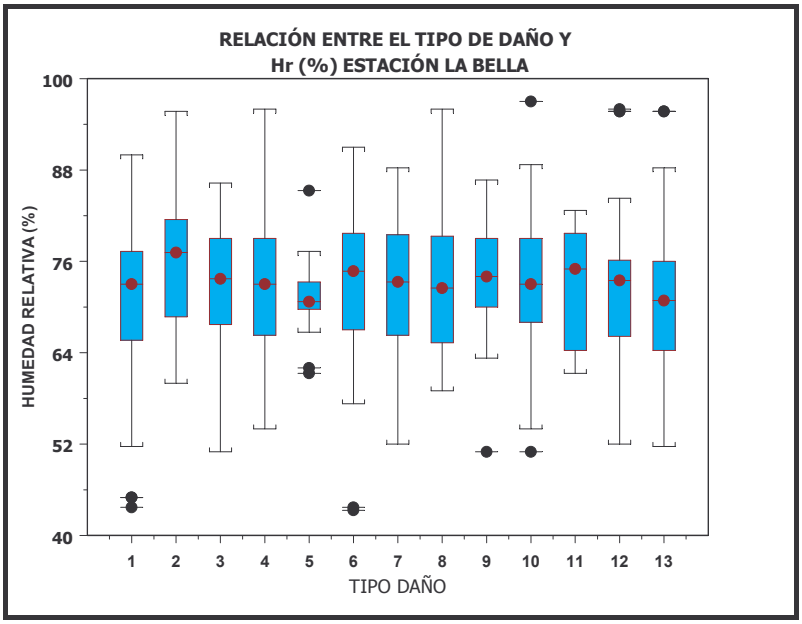


Figura V.5. Relación entre el tipo de daño y la humedad relativa en la estación La Bella

En la Figura V.5 se visualiza la relación entre la variable tipo de daño (numerizada) y el campo correspondiente a la humedad relativa en la estación La Bella. Se observa que la distribución de los datos está bastante dispersa para cada uno de los tipos de daño, así como la asimetría en las distribuciones. Para el caso del tipo de daño 5 *fuga por fraude* se tiene menor dispersión de la variable de humedad relativa y un valor de la mediana algo menor que para el resto de los daños reportados.

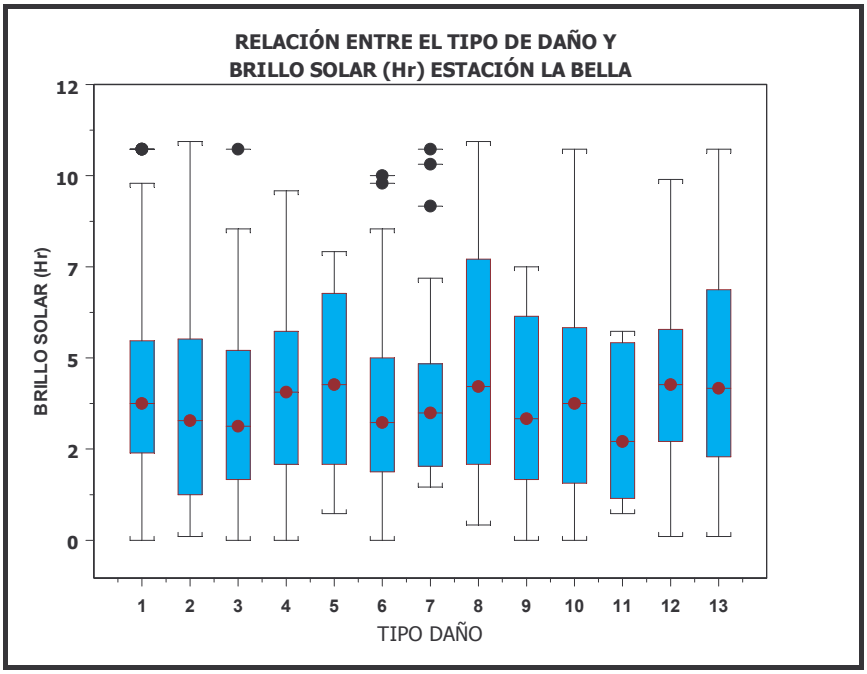


Figura V.6. Relación entre el tipo de daño y el brillo solar en la estación La Bella

Se observa en la Figura V.6 la relación entre la variable tipo de daño (numerizada) y el campo correspondiente al brillo solar en la estación La Bella. Se aprecia gran dispersión y asimetría en la distribución de los datos por tipo de daño, con excepción del tipo de daño 11 *fuga por universal* que presenta menor dispersión. Adicionalmente, en la Figura V.7 se visualiza la relación entre la variable tipo de daño (numerizada) y la variable correspondiente a la precipitación en la estación La Bella, apreciándose una gran dispersión y asimetría en el conjunto de datos. En general los datos se concentran en su mayoría en niveles bajos de precipitación siendo más dispersos los valores altos. La distribución presenta una asimetría positiva o sesgada a la derecha.

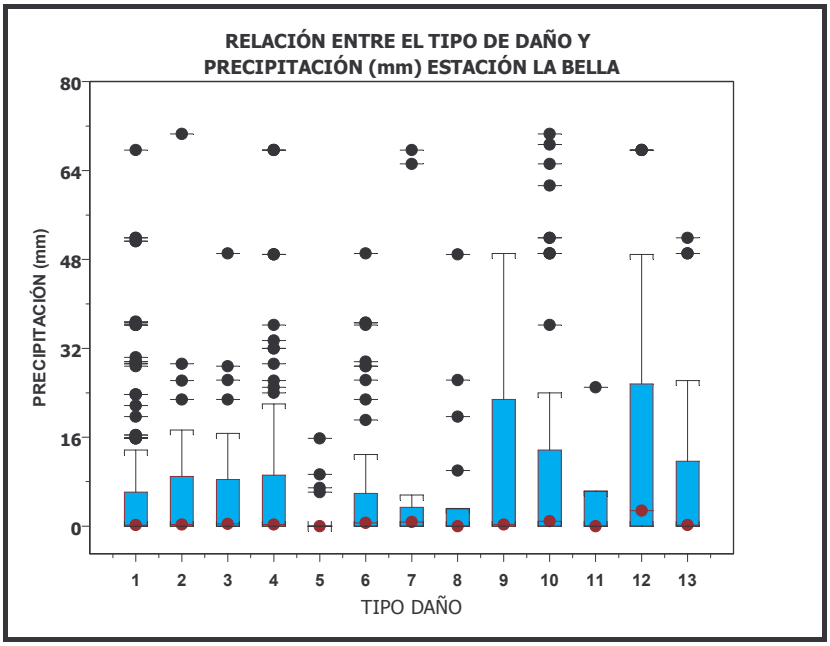


Figura V.7. Relación entre el tipo de daño y la precipitación en la estación La Bella

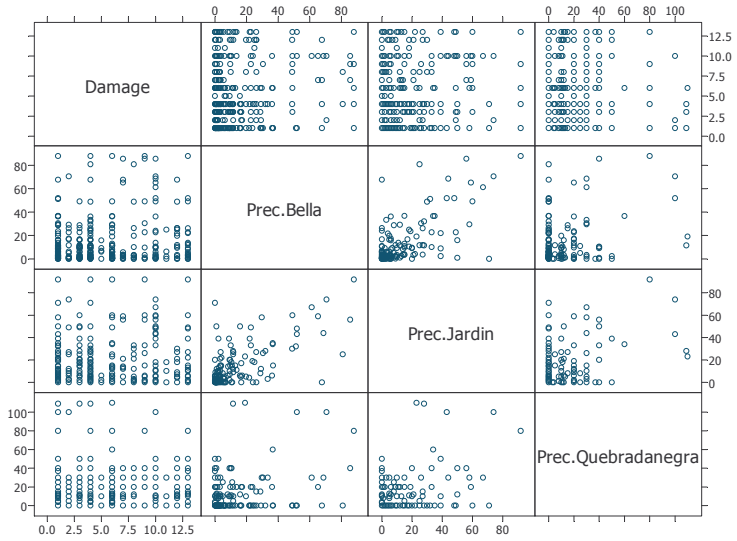


Figura V.8. Diagrama de puntos de variables Tipo de daño, Precipitación en las estaciones La Bella, Jardín, Quebradanegra

En la Figura V.8 se visualiza la dispersión de puntos en el diagrama cartesiano que representan la variable tipo de daño (numerizada), y los campos correspondientes a las precipitaciones diarias en las estaciones La Bella, Jardín y Quebradanegra. Se observa correlación entre las estaciones pluviométricas, especialmente entre las estaciones la Bella y Jardín, pero entre la variable tipo de daño y cada una de estas variables se observa gran dispersión.

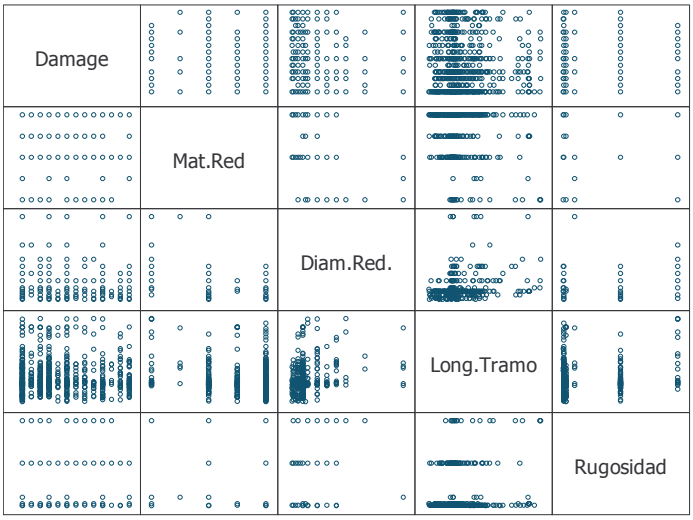


Figura V.9. Diagrama de puntos de variables Tipo de daño, Material, Diámetro, Longitud del tramo y Rugosidad en el modelo hidráulico

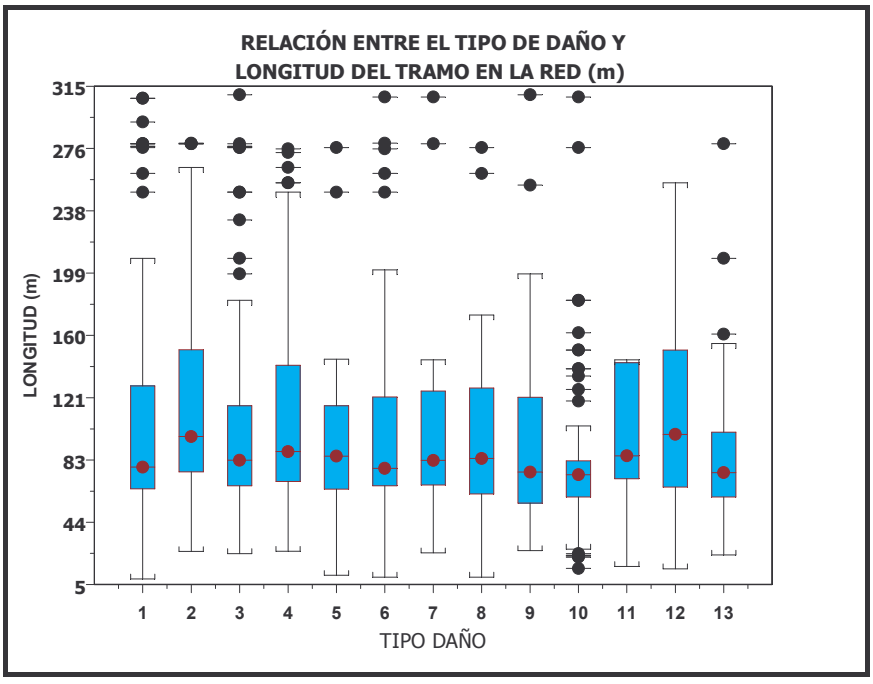


Figura V.10. Relación entre el tipo de daño y la longitud de los tramos en la red

Se observa en la Figura V.9 la nube de puntos del diagrama cartesiano correspondiente a la variable tipo de daño (numerizada), y los valores correspondientes a las variables tomadas del modelo hidráulico (material, diámetro, longitud y rugosidad). Entre estas variables se aprecia dispersión en los datos. En la Figura V.10 se visualiza la relación entre la versión numerizada del tipo de daño y la longitud correspondiente al tramo de la red en el cual se ubica el reporte de daño. Se observa dispersión y asimetría para la distribución del conjunto de datos, donde los valores atípicos corresponden a la topología de la

red. La mayoría de los datos de las longitudes de cada tramo en la red corresponden a valores menores de 100 metros, presentándose algunos valores altos (atípicos) para la longitud de tramo en una red de distribución.

En la Figura V.11 se visualiza la relación entre la variable numerizada del tipo de daño y la variable correspondiente a la rugosidad del material tomada del modelo hidráulico. Al igual que en las figuras anteriores se observa que la distribución de los datos presenta bastante dispersión. También se observa que la mayoría de los datos corresponden a materiales con rugosidades bajas, teniendo la distribución un sesgo a la derecha.

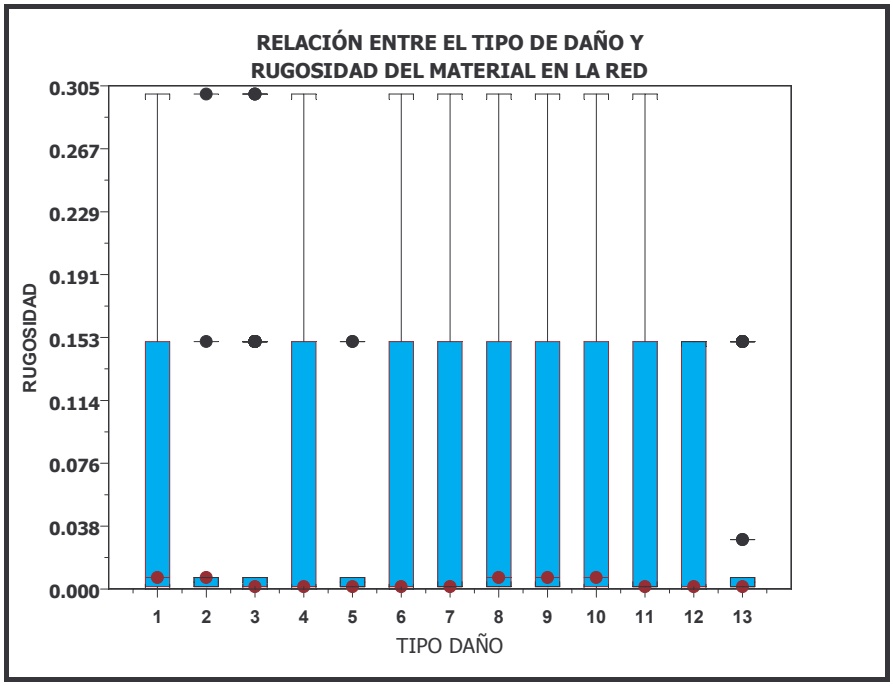


Figura V.11. Relación entre el tipo de daño y la rugosidad del material en la red

Se observa en la Figura V.12 la nube de puntos correspondiente al diagrama cartesiano de los datos de la variable tipo de daño (numerizada), y la variable correspondiente a los caudales horarios (9, 10, 11, y 12) tomados del modelo hidráulico. La tendencia sigue siendo la dispersión de los datos entre el tipo de daño y esta variable de caudales. Los caudales horarios presentan la correlación esperada del modelo hidráulico de acuerdo con la curva de modulación. El valor extremo que se observa en la gráfica corresponde al tramo de la conducción a la salida de la planta de tratamiento.

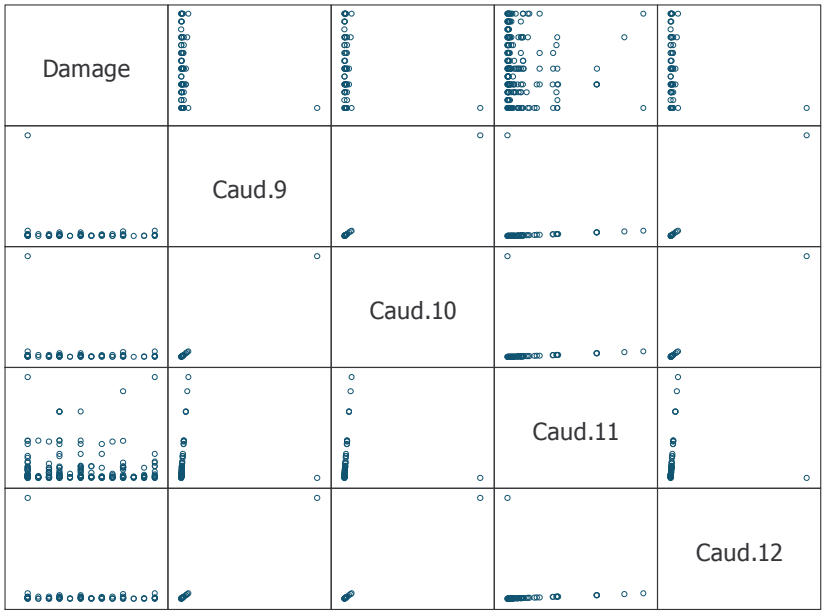


Figura V.12. Diagrama de puntos de variables Tipo de daño, Caudales horarios en el modelo hidráulico

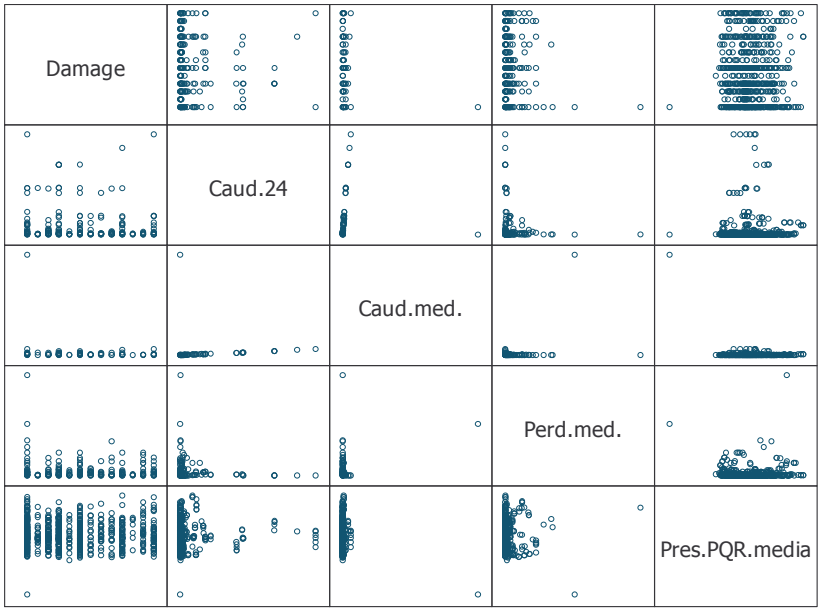


Figura V.13. Diagrama de puntos de variables Tipo de daño, Caudal y Pérdida media del modelo hidráulico, Presión media en el PQR.

En la Figura V.13 se visualiza la dispersión de la nube de puntos correspondiente al diagrama cartesiano de las variables tipo de daño (numerizada) y de las variables derivadas del modelo hidráulico (caudal promedio del diario horario, pérdida media diaria horaria y presión media diaria horaria en los puntos PQR). Se observa la correlación existente entre los valores medios de los caudales, pérdidas y presiones en cada uno de los puntos de PQR.

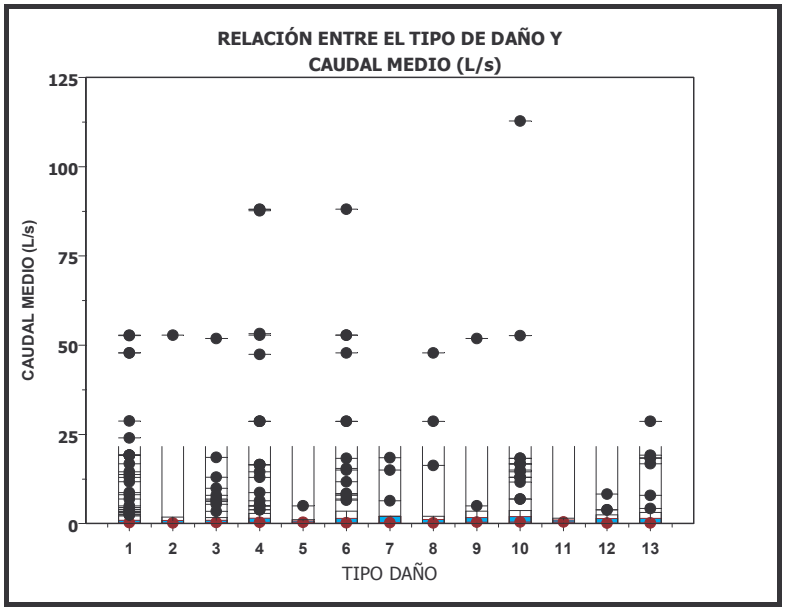


Figura V.14. Relación entre el tipo de daño y el caudal medio en la red

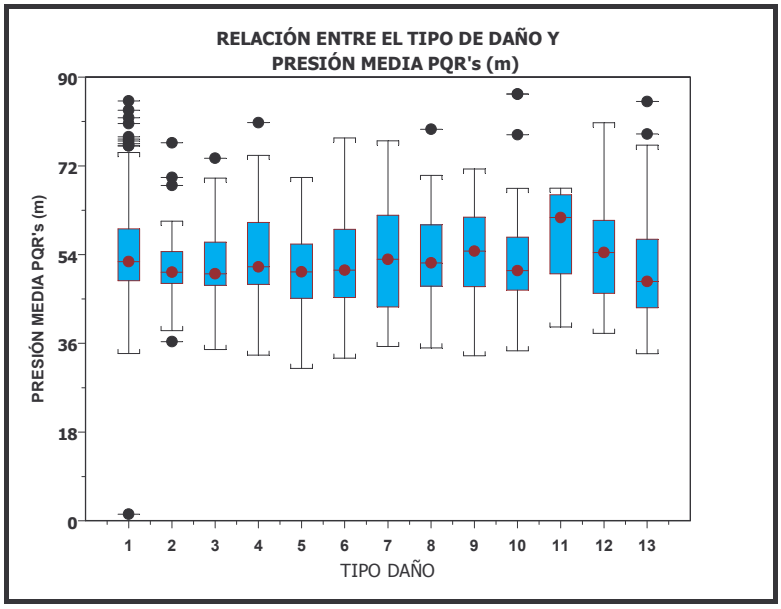


Figura V.15. Relación entre el tipo de daño y la presión media en los PQR's

En las Figuras V.14, V.15, y V.16 se visualizan las relaciones mostradas en la Figura V.13. En general se aprecia una gran dispersión para la distribución de los valores de caudales, presiones y pérdidas en la red con respecto al valor numerizado del tipo de daño. En la distribución de los caudales medios no se aprecia ningún tipo de tendencia con respecto al tipo de daño, aparte de la gran dispersión. Con respecto a la presión media se observa asimetría en el conjunto de datos para todos los tipos de daños, sin apreciarse valores altos de la mediana del conjunto de datos, aumentando su valor en el caso del tipo de daño 11 *fuga*

por universal. En cuando a la pérdida media horaria en los tramos de la red se aprecia igualmente dispersión exceptuando el tipo de daño 7 *daño interno* en el que se ve una distribución más homogénea.

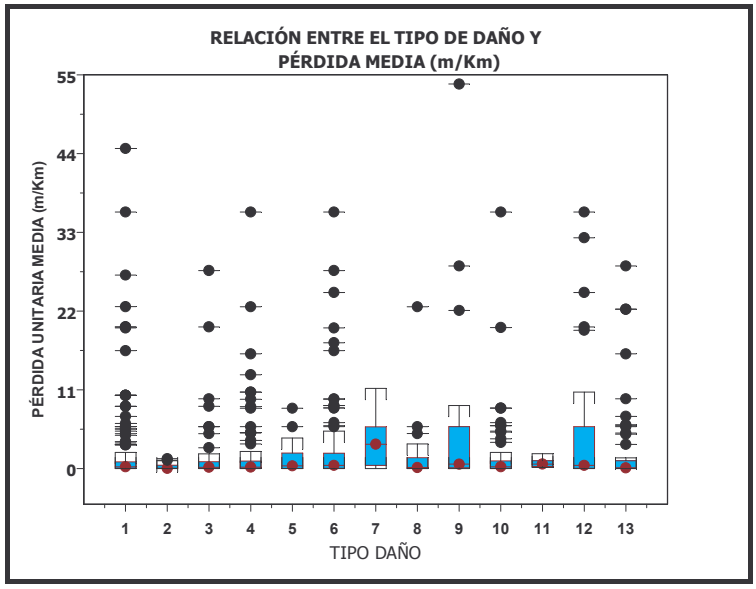


Figura V.16. Relación entre el tipo de daño y la perdida media en la red

La Figura V.17 corresponde a la visualización de la dispersión de la nube de puntos del diagrama cartesiano para la variable tipo de daño (numerizada), y los valores de los atributos correspondientes a las presiones horarias (9, 10, 11, 12) tomadas del modelo hidráulico en cada uno de los puntos PQR. Se aprecia una correlación fuerte entre las presiones horarias en la red, y dispersión del conjunto de datos para los tipos de daños con respecto a las presiones.

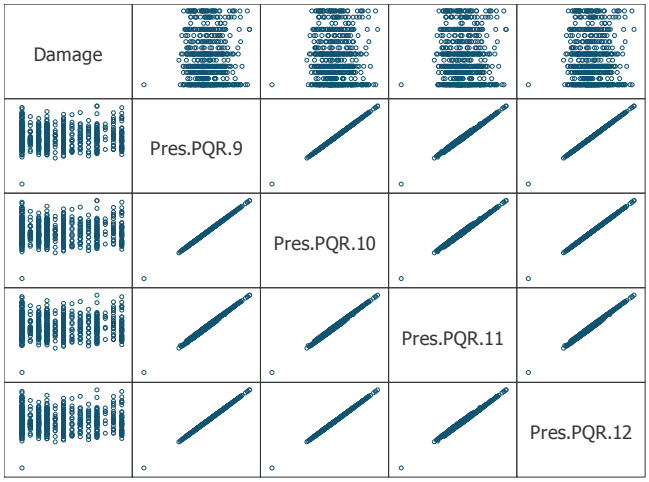


Figura V.17. Diagrama de puntos de variables Tipo de daño, Presiones horarias en el PQR del modelo hidráulico

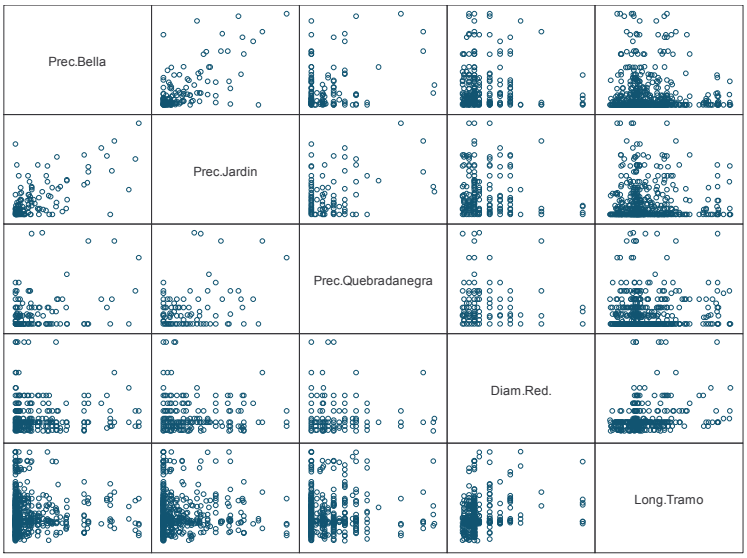


Figura V.18. Diagrama de puntos de variables utilizadas en los prototipos desarrollados

Por otra parte, en la Figura V.18 se visualiza la dispersión de la nube de puntos correspondiente al diagrama cartesiano de algunas de las variables consideradas en los prototipos, tales como las precipitaciones en las estaciones climatológicas, y los datos del modelo hidráulico de los diámetros y longitudes en la red correspondientes a la ubicación geográfica del punto de PQR. No se aprecian tendencias de correlación entre los diámetros y las longitudes en la red con las precipitaciones de las estaciones climatológicas.

En la Figura V.19 se observa la relación entre la variable correspondiente a los diámetros reportados en los registros de PQR's y la variable del material igualmente reportado en estos registros en el municipio de Calarcá, superponiendo la ubicación geográfica de PQR's; se aprecia que la mayor parte de los diámetros corresponde a un diámetro menor de 50 mm (2") distribuidos en toda el área, lo cual es indicativo de que la mayor parte de reportes de daños corresponden a domiciliarias; esto, en principio, implica mayor riesgo de pérdida de agua sin control en la red, debido a que la detección de este tipo de daños es más complicada que si se presenta en una tubería de mayor diámetro. En cuanto a la ubicación espacial no se puede concluir una tendencia hacía un lugar específico.

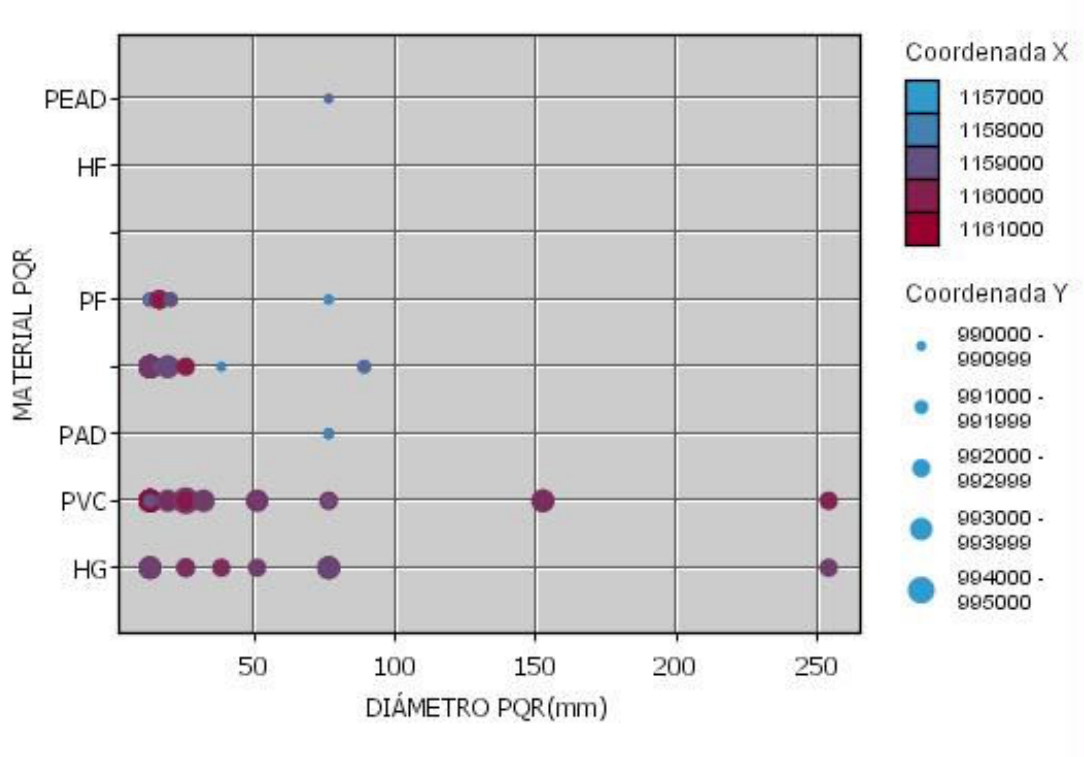


Figura V.19. Relación entre el diámetro y el material en los registros PQR's

La Figura V.20 muestra la relación entre los datos de las variables del tipo de daño (numerizado) y el material tomado de los reportes de peticiones quejas y reclamos PQR's, superponiendo su ubicación espacial. Es importante mencionar que el 71% de la información no reporta el tipo de material al que hacer referencia el daño. Se aprecia que el tipo de daño 1, *fuga por tubo roto,* se presenta en todos los tipos de materiales y abarca gran porción del municipio. Para el tipo de daño 4, *fuga por la pitorra,* no se tiene información del material.

Esta figura también permite observar que las tuberías de polietileno están ubicadas sin presentar una disgregación extensa en el municipio y tampoco se presenta un número elevado de número de daños con este material, debido a que probablemente corresponden a un tipo de material más reciente; los *fraudes* (tipo de daño 5) se detectan más en materiales de hierro galvanizado y plástico flexible. El tipo de daño 11, *fuga por universal,* se presenta más en tuberías de PVC. Las *fugas por las acometidas* (tipo de daño 13), seguramente debidas a problemas en la instalación, se presentan más en las tuberías plásticas.

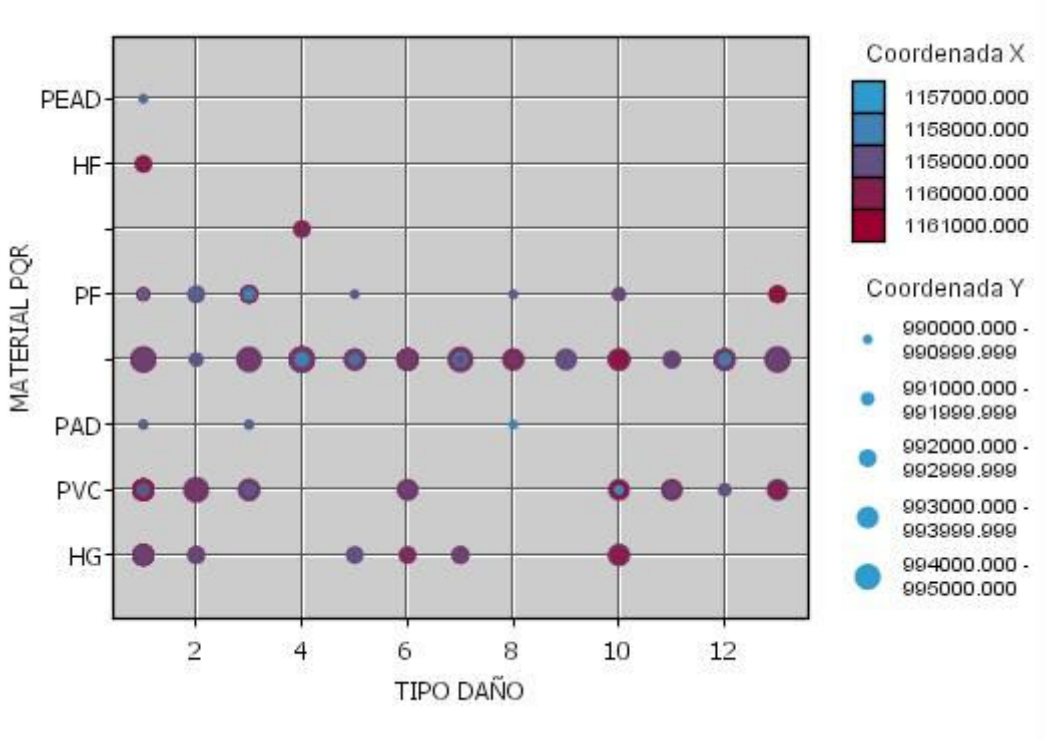


Figura V.20. Relación entre el tipo de daño y el material reportado en el PQR

De los reportes se tiene que en solo uno de ellos el material es hierro fundido, en 48 casos el material corresponde a hierro galvanizado, en 6 a polietileno de alta densidad, en 81 a plástico flexible, y en 98 de los casos el material es PVC. Para los reportes de tuberías en hierro fundido y plástico flexible, no se encuentra ninguna coincidencia con el material del modelo hidráulico de la red. Las tuberías de hierro galvanizado tienen un 58.33% de coincidencia, las de polietileno de alta densidad el 50%, y las de PVC el 66.33%. En cuanto a la ubicación geográfica, en general, están bastante disgregados los materiales en todo el municipio, aunque las tuberías plásticas se tienen más hacia el sur, y las de hierro galvanizado en el centro (mayor antigüedad).

En la Figura V.21 se visualiza la relación entre los datos de las variables correspondientes al tipo de daño reportado (numerizado) y el material de la red tomado del modelo hidráulico de acuerdo con la ubicación espacial de cada uno de los reportes de PQR's. Esto corresponde a realizar *clusters* de la localización del tipo de daño reportado con el material presente en el modelo. Los valores que aparecen sin material, son porque no están dentro del modelo hidráulico (corresponden a datos de PQR's para los cuales no se tiene el modelo).

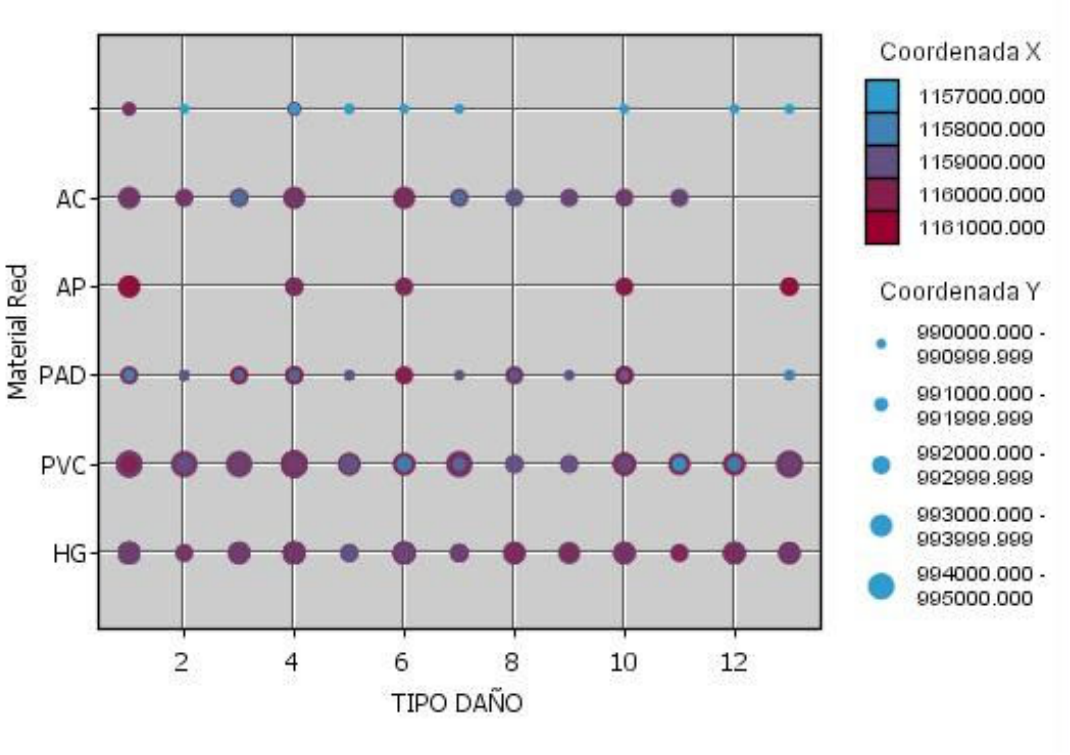


Figura V.21. Relación entre el tipo de daño y el material de la red

Como se puede observar, aunque para los materiales de tuberías de asbesto cemento (AC) y de hormigón armado (AP) no se tuvieron reportes dentro de los daños (ya que la mayor problemática está presente en las redes domiciliarias, que corresponden a diámetros más pequeños en otro tipo de materiales) si se tiene correspondencia para la red principal de acuerdo con la ubicación de cada uno de estos reportes. Esto corresponde a una de las hipótesis que se están planteando para la realización de este ejercicio práctico, y es el encontrar relaciones entre las diferentes variables obtenidas del modelo hidráulico y los daños reportados.

La Figura V.22 muestra la dispersión en el rango de presiones medias para los tipos de daños reportados (numerizados) superponiendo la ubicación geográfica, notándose que no se presenta una relación explícita entre las dos variables, así como tampoco se aprecia una conclusión con respecto a posiciones geográficas al respecto, es decir se encuentran disgregadas por todo el área el rango de presiones.

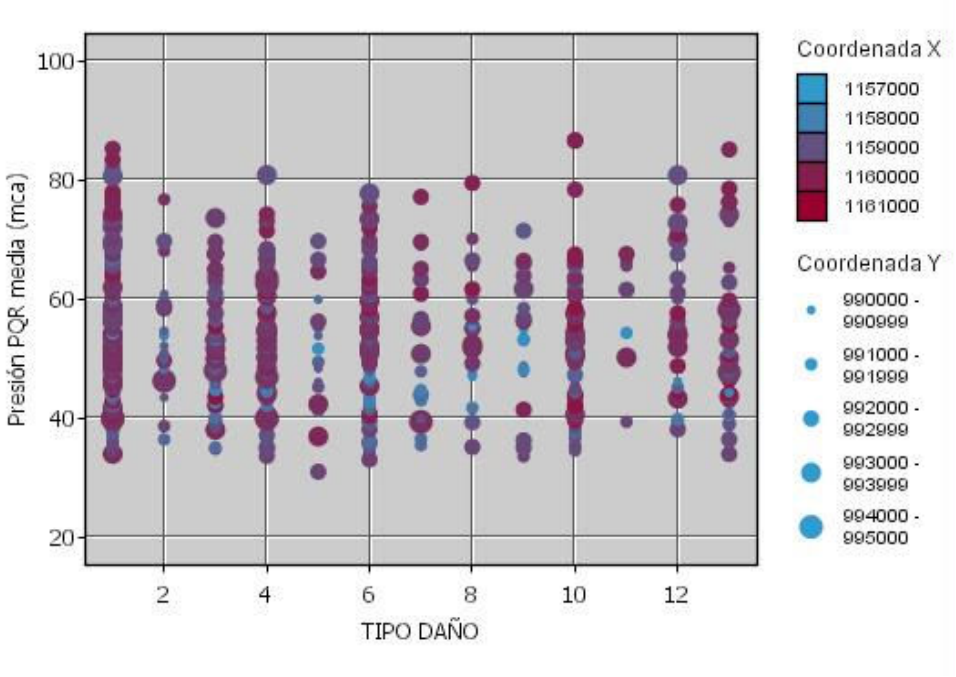


Figura V.22. Relación entre el Tipo de daño y la Presión media diaria en PQR's

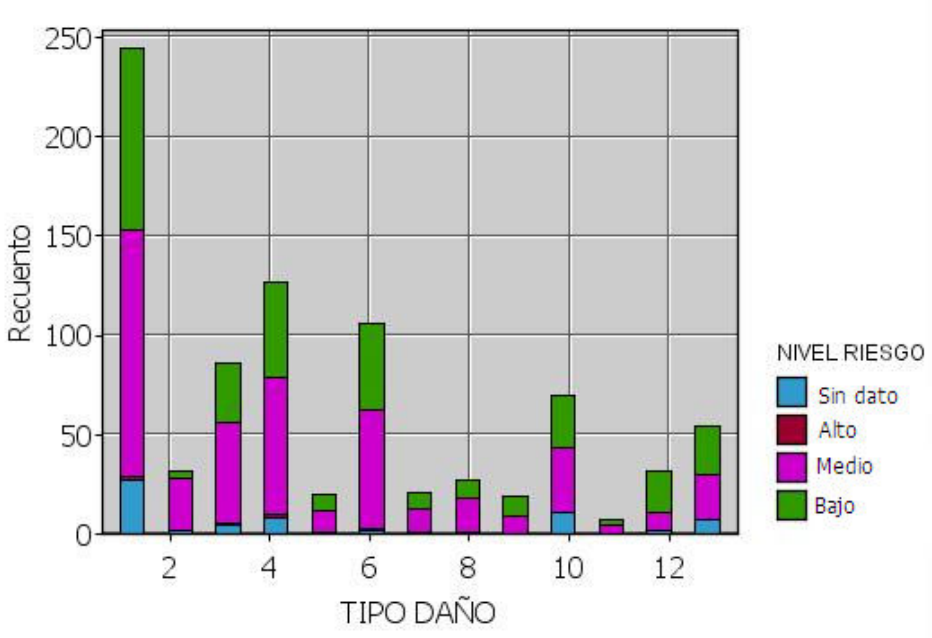


Figura V.23. Distribución Tipo de Daño según el nivel de riesgo

El diagrama de barras de la Figura V.23 representa un recuento del tipo de daño de acuerdo con el nivel de riesgo. Se aprecia que la mayoría de los puntos PQR's están ubicados en zonas de riesgo bajo, teniéndose una pequeña cantidad de puntos en zonas de riesgo alto, pudiendo presentar problemas serios debido a su alto grado de vulnerabilidad en cuando a sismos y deslizamientos.

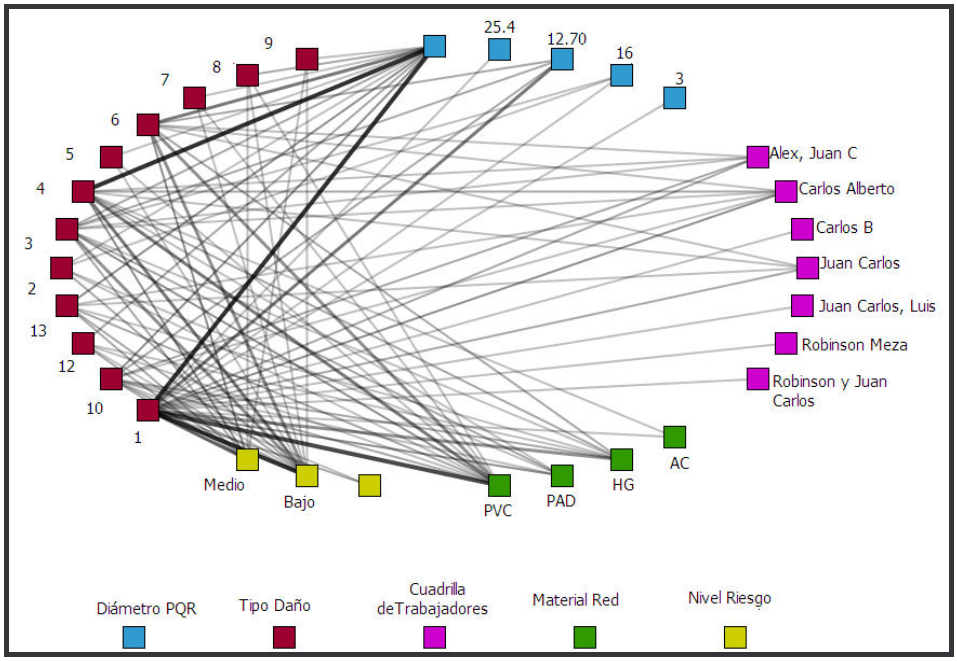


Figura V.24. Fuerza de relaciones entre diferentes variables

La malla direccional de la Figura V.24 muestra la fuerza de las relaciones desde el campo *tipo de daño* (numerizado) hacía los campos *diámetro, funcionario, material y nivel de riesgo*. La conexión se muestra haciendo uso de diferentes intensidades de líneas para indicar su fuerza. Se visualizan enlaces fuertes desde el tipo de daño 1 hacia el campo diámetro para el cual no se tiene dato y para 12.70 mm (1/2 pulgada); hacia el campo nivel de riesgo para los valores medio y bajo; y hacia el campo material para los valores PVC y HG. El campo tipo de daño 4 presenta enlaces fuertes hacia el campo diámetro para cuando no se cuenta con el valor; hacia el campo material para el valor PVC; y hacia el campo nivel de riesgo bajo. El campo tipo de daño 6 presenta enlaces fuertes hacia el campo diámetro cuando no se dispone del dato y hacia el campo nivel de riesgo bajo. Para los demás campos se aprecian enlaces medios como en el caso del tipo de daño 1 y el trabajador Carlos Alberto; o enlaces débiles como en el caso del tipo de daño 1 y el material AC.

V.4. Modelado

Tal como se define la minería de datos, la búsqueda de patrones o relaciones entre variables no es una tarea ni mucho menos trivial. En el caso

específico del que tratamos en esta tesis se tienen ciertas circunstancias añadidas que, sin pretender elevar como excusa, no es menos cierto que afectan la etapa de modelado de la información y validación de los resultados obtenidos. Básicamente, al no contar con mayor cantidad de información real de la red (mediciones), se parte de supuestos obtenidos del modelo hidráulico. No obstante, los objetivos que se plantean son cubiertos, al querer presentar las técnicas de *KDD* (Knowledge Discovery in Data Bases) y *Data Mining* como herramientas de utilidad en el diseño, operación y manejo de los sistemas de abastecimiento de agua, y generalizarlo a cualesquiera de ellos.

El trabajo de modelado de la información se realiza con la herramienta Clementine 9.0 SPSS (Anexo 1); después de hacer una revisión de diferentes herramientas en el mercado, tanto de libre distribución como de pago, se optó por esta ya que cuenta con una interfase visual bastante amigable, el formato para la introducción de la información es sencillo, así como una cantidad de algoritmos implementados que cubren las necesidades y objetivos propuestos; la visualización de los resultados es también idónea para estos fines, o pueden ser fácilmente exportados para ser utilizados en otros programas.

Adicionalmente, se tuvieron en cuenta los criterios para la selección de una herramienta de minería de datos (Klösgen y Zytkow, 2002), que se presentan en la Tabla V.1, los cuales son satisfechos por el programa Clementine 9.0.

Entrada (Input)	
Datos	Tipos de datos soportados
	Opciones de trasformación / preprocesamiento
Dominio de conocimiento	Tipos de dominio de conocimiento a explotar
Salida (Output)	
Algoritmos	Tipos de tareas de minería soportados
	Métodos disponibles para realizar la tarea
	Propiedades de los algoritmos (exactitud, robustez, escalabilidad,....)
	Sofisticación del algoritmo (bagging, boosting, ...)
Presentación	Visualización
	Reportes

Usuario	Tipos y rol de usuario Guía, documentación de pasos, repetición, automatización Intuición de interconexión, interacción
Tecnología	Plataformas Lenguaje de programación Integración con otros sistemas (especialmente sistemas de bases de datos) Estructura arquitectónica y de módulos
Soporte	Documentación Servicio y consultaría Viabilidad de proveedores y productos

Tabla V.1. Criterios para la selección de una herramienta de minería de datos

Además, la universidad Politécnica de Valencia cuenta con licencias de tipo concurrente para ser utilizadas dentro del campus de la herramienta Clementine 9.0. El trabajo se realizó en un computador personal AMD Duron 900 MHz, 64 KBytes de cache, 256 MBytes de memoria, y 20 GBytes de disco duro.

Como se muestra en las gráficas presentadas en los apartados anteriores, no se aprecia de antemano una relación clara entre las diferentes variables de que se dispone y el daño reportado, aspecto que se presenta sugerente para la utilización de técnicas capaces de encontrar correspondencias ya sea de tipo cualitativo en cuanto a clasificaciones, o de tipo predictivo entre las diferentes variables.

Los modelos que se han elegido para el tratamiento de la información son: los árboles de decisión, clasificación y regresión (*C&RT, CHAID, QUEST, C5.0*), las redes neuronales, y las redes de Kohonen, por su adaptabilidad al tipo de información presente y al dominio del problema planteado en los objetivos.

Del total de los registros de la base de datos sólo se tiene un 4% de información faltante. Esta información pérdida básicamente corresponde a valores no encontrados en los reportes tales como, ubicaciones, diámetros, materiales, etc. Ante la dificultad para obtener esta información faltante y al estar dispersa por toda el área del municipio, se ha descartado para su utilización en los

modelos propuestos. No obstante, una de las ventajas de la utilización de técnicas de árboles de regresión y clasificación, es su robustez ante información faltante y valores extremos.

En el siguiente apartado se presentan los modelos o prototipos entrenados cómo desarrollo de esta tesis. En general, se dividieron en tres prototipos para agrupar diferentes modelos entrenados en cada uno de ellos de acuerdo con características propias de cada prototipo. En todos los prototipos entrenados la variable respuesta fue tomada como *el tipo de daño* tanto en su versión categórica cómo la etiqueta numerizada, por considerarla importante en la toma de decisiones para la gestión de la red de abastecimiento de agua potable del municipio. La etiqueta de la versión numerizada es tomada por el algoritmo como un valor categórico y no como un rango numérico. Con esta numeración de variables entre otros se reduce el tamaño de la base de datos entrenada facilitando la ejecución del algoritmo, así como se adapta para poder utilizar modelos que requieran de este tipo de dato como entrada.

V.4.1. Soluciones del modelado

Cada uno de los siguientes prototipos fue planteado con el objetivo de ir desarrollando de forma ordenada modelos que permitieran ampliar el conocimiento de la información disponible, así como intentar mejorar el resultado al modelado propuesto. El discurso presentado sigue la forma en que se trabaja en Clementine 9.0 siguiendo rutas tal como se aprecia en el Anexo 1.

V.4.1.1. Prototipo 1

En este prototipo se descartaron los registros en los cuales se tienen valores faltantes, valores extremos, y valores espurios. Aunque en principio los algoritmos elegidos para el modelado de los datos son bastantes robustos ante este tipo de información, se quiere probar como se comportan los modelos entrenados sin contar con estos datos. Uno de los inconvenientes de descartar

esta información es la reducción del tamaño de la base de datos de entrada al modelo. La base de datos completa cuenta con un total de 218 campos y 846 registros, para un total de 184428 valores. Al efectuar estas tareas de preprocesamiento de la información, la cantidad de la información de los registros (223) resultantes se reduce a un total de 48614 valores. El entrenamiento de estos modelos se realizó con la totalidad de los datos, es decir no se partió la base de datos en prueba y comprobación. No obstante, algunos de los modelos entrenados fueron probados con la totalidad de la información (846 registros) para verificar su poder de clasificación.

Aunque, como se ha mencionado con anterioridad, las técnicas de minería de datos tienen la ventaja sobre las técnicas estadísticas tradicionales, de encontrar patrones y tendencias en los datos sin necesidad de tener preestablecida una relación causa efecto entre éstos, se ha elegido intentar la estimación de la relación de los daños reportados y el resto de atributos con los que se cuenta, debido a su utilidad práctica en la gestión del abastecimiento. Por otra parte, ya que los modelos de aprendizaje automático pueden ser construidos tanto para mejorar el conocimiento de los datos como para efectuar clasificaciones y predicciones, el planteamiento que se realiza en esta tesis pretende realizar un estudio de los dos modelos; comprensión y clasificación.

Modelo	*Algorit.*	*Clasificaciones %*		*Total Reg.*	*Neuronas en capas*			*T.Entrenam.*		
		Corr.	*Err.*		*Entr.*	*Ocul.*	*Sali.*	*Hr*	*min*	*Sg*
Árbol	C&RT	69.96	30.04	223				0	0	23
ANN	Múltiple	39.91	60.09	223	35	39-2	13	1	5	56
ANN	Poda	53.81	46.19	223	18	2	13	0	22	22
Árbol	C&RT	76.23	23.77	223				0	7	57
ANN	PODA	26.00	74.00	846	18	2	13	0	22	22
Árbol	C&RT	30.26	69.74	846				0	0	23
Árbol	C&RT	37.00	63.00	846				0	7	57
Árbol	C&RT	60.99	39.01	846				0	0	9
Árbol	C&RT	69.06	30.94	223				0	0	6

Modelo	*Algorit.*	*Clasificaciones %*		*Total Reg.*	*Neuronas en capas*			*T.Entrenam.*		
		Corr.	*Err.*		*Entr.*	*Ocul.*	*Sali.*	*Hr*	*min*	*Sg*
Árbol	C&RT	40.90	59.10	846				0	13	56
Árbol	C&RT	40.78	59.22	846				0	12	37
Árbol	CHAID	33.81	66.19	846				0	0	49
Árbol	C&RT	35.11	64.89	846				0	2	41
ANN	Dinámico	31.68	68.32	846	3	3-4	13	0	9	31
Árbol	C&RT	35.68	64.42	846				0	1	15

Tabla V.2. Resumen de cada uno de los diferentes modelos entrenados y tiempo de entrenamiento para en prototipo 1

Tal como se pude observar en la tabla anterior los tres primeros modelos entrenados corresponden a árboles de clasificación y regresión *C&RT*, y a redes neuronales (método de poda y múltiple), de los cuales se obtienen 69.96%, 53.81%, y 39.91% respectivamente de valores clasificados correctamente. De igual forma el tiempo de entrenamiento de cada uno de los modelos fue de 23 segundos para el árbol, 22 minutos y 22 segundos para la red con el método de poda, y de 1 hora 5 minutos 36 segundos para la red con el método múltiple.

Del modelo *C&RT* citado se filtró la información para tener en cuenta los datos que el algoritmo no consideró importantes en este modelo, con la cual se obtuvo un nuevo árbol que clasifica el 76.23% de los datos correctamente. No obstante dentro de la información tenida en cuenta en este modelo aparece el identificador del registro, que corresponde a una variable que no es relevante para clasificar los tipos de daño.

Posteriormente, los dos modelos de árboles *C&RT (69.96%)* y el árbol de red neuronal poda (53.81%) fueron aplicados a la totalidad de la base de datos, obteniéndose unos valores de porcentaje de clasificaciones correctas de 37%, 30.26%, y 26% respectivamente, lo cual muestra que con la totalidad de los datos se pierde eficacia en la estimación del tipo de daño con estos modelos entrenados.

En vista de que los árboles *C&RT* tienen mayor eficiencia y el tiempo y

recursos computacionales gastados son menores, se utilizaron las redes de Kohonen como clasificadores previos a la utilización de un modelo de *C&RT* con el objetivo de intentar mejorar las estimaciones obtenidas anteriormente. Este árbol presenta un porcentaje de clasificaciones correctas del 60.99% para el campo tipo de daño, para la totalidad de la información (846 registros). Además, de este conglomerado de Kohonen se descartaron, al igual que en el análisis anterior, los datos faltantes o espurios (al descartar estos datos se descarta toda la fila del registro) y se entrena de nuevo un árbol *C&RT* con el cual se obtiene un porcentaje de clasificaciones correctas de 69.06%, aunque solo quedan un total de 223 registros de la base de datos. La configuración de creación del modelo de Kohonen fue la siguiente:

- Tiempo: 15 minutos
- Semilla aleatoria: 123
- Ancho: 10
- Longitud: 7
- Decrecimiento de tasas de aprendizaje: exponencial
- Vecindad: 2
- Eta inicial. 0.3
- Ciclos: 20
- Vecindad: 1
- Eta inicial: 0.1
- Ciclos: 150

Por último se entrenó un árbol *C&RT* con la totalidad de los datos (846 registros) y se obtuvo un porcentaje de clasificaciones correctas de 40.9 para el tipo de daño; posteriormente se filtraron los datos que el algoritmo considera

importantes para reducir el conjunto de datos y entrenar un nuevo modelo C&RT sin estos datos con el cual se obtuvo un porcentaje de clasificaciones correctas de 54.61% para el campo tipo de daño. A continuación se entrenaron un par de redes de Kohonen sucesivas con la finalidad de agrupar la información e intentar mejorar las clasificaciones, para luego entrenar dos árboles (*CHAID* y *C&RT*) y una red neuronal. El porcentaje de las clasificaciones correctas 33.81% para el modelo *CHAID*, 35.11% para el modelo *C&RT* y 31.68% para la red neuronal por el método dinámico; una última red de Kohonen conectada a las dos anteriores se utilizó para intentar mejorar la predicción del árbol *C&RT* con lo cual se obtuvo porcentaje de clasificaciones correctas de 35.68%.

Con base en los resultados obtenidos a partir del porcentaje de clasificaciones correctas para el campo tipo de daño en cada uno de los modelos, claramente los árboles de clasificación y regresión presentan un mejor ajuste tanto para los modelos en los que se ha descartado información por faltantes, nulos, fuera de rango o perdidos (223 ocurrencias) y para la información completa (846 ocurrencias).

V.4.1.2. Prototipo 2

En este prototipo se utilizaron datos de entrenamiento y comprobación para validar los modelos. Se entrenaron varios modelos intentando mejorar lo obtenido en el prototipo anterior. Se utilizó la totalidad de los 846 registros, sin descartar información. En Clementine se pueden dividir los datos en dos muestras (entrenamiento y comprobación) o en tres (entrenamiento, comprobación y validación).

- Entrenamiento y comprobación: si se tienen dos muestras, los modelos se construyen con datos de entrenamiento y se comprueban con datos de prueba.

- Entrenamiento, comprobación y validación: en caso de utilizar tres muestras, los modelos se construyen con datos de entrenamiento, se refinan con datos de prueba y se comprueban con datos de validación.

Modelo	*Algoritmo*	*% Clasificaciones Correctas*		*Neuronas en capas*			*Tiempo Entrenamiento*		
		Entre.	*Comp.*	*Entr.*	*Ocul.*	*Sali.*	*Hr*	*min*	*Sg*
Árbol	C&RT	77.72	32.13				0	0	24
ANN	Rápido	10.72	12.85	214	10	4	0	21	23
ANN	Dinámico	27.64	31.33	214	25-6	4	13	5	2
Árbol	Múltiple	25.63	30.12	214	28-5	4	0	17	34
ANN	Poda	5.86	4.82	214	30	4	0	31	0
Árbol	RBFN	25.96	26.51	214	20	4	0	47	51
Árbol	Exhaustiva	12.73	12.85	200	28-20	4	0	49	15
Árbol	RBFN-Pers1.	26.97	18.07	214	20	4	0	49	15
Árbol	RBFN-Pers2.	29.98	20.48	214	40	4	0	40	42

Tabla V.3. Resumen de configuración y resultados de los modelos entrenados en el prototipo 2

Se realizó el entrenamiento de modelos de árboles de clasificación y regresión, así como diferentes entrenamientos de redes neuronales con diferentes métodos (anexo 1), dividiendo la información en datos de entrenamiento y datos de comprobación, 70% y 30%, respectivamente.

Se intentó mejorar el modelo con más alto porcentaje de clasificaciones correctas de los presentados en la Tabla V.3, que corresponde al modelo *C&RT* (árboles de regresión y clasificación); con este fin se filtró la información obtenida en este modelo reduciendo el tamaño del conjunto de datos y dejando solo la información que el algoritmo considera importante, seguidamente se entrenó una red neuronal por el método RBFN (124,40,4) con lo cual se obtuvo un porcentaje de clasificaciones correctas de 29.98% para los datos de entrenamiento y de 25.30% para los datos de comprobación. A partir del filtro también se entrenó una red de Kohonen para, a partir de los grupos generados, comprobar el comportamiento e intentar mejorar los modelos.

Luego de esta red de Kohonen los datos se dividieron en dos muestras (entrenamiento y comprobación) y en tres (entrenamiento, comprobación y validación). En el caso de entrenamiento y comprobación los modelos se

construyeron con datos de entrenamiento y se comprobaron con datos de prueba. Si se utilizan tres muestras, los modelos se construyen con datos de entrenamiento, se refinan con datos de prueba y se comprueban con datos de validación. El resumen de los resultados obtenidos y la configuración de los modelos se presentan en la Tabla V.4.

Modelo	*Algoritmo*	*% Clasificaciones Correctas*			*Neuronas en capas*			*T.Entrenam.*		
		Ent.	*Com.*	*Val.*	*Entr.*	*Ocul.*	*Sali.*	*Hr*	*min*	*Sg*
Árbol	C&RT	57.29	36.14					0	7	27
ANN	RBFN	31.49	24.50		124	40	4	1	26	16
ANN	Dinámico	21.61	24.50		124	44-8	4	0	31	49
Árbol	C&RT	59.05	31.91	26.74				0	8	52
ANN	RBFN	33.60	29.96	20.93	124	40	4	0	54	8

Tabla V.4. Resultados de los modelos después de Kohonen en el prototipo 2

Basándonos en los resultados que se presentan en la Tabla V.3, en la cual se aprecia que una de las mejores redes neuronales entrenadas corresponde a la *Red de función de base radial,* se entrenaron diferentes tipologías de red para intentar mejorar su estimación, tal como se muestra en la Tabla V.5.

Modelo	*Algoritmo*	*% Clasificaciones Correctas*		*Neuronas en capas*			*Tiempo Entrenamiento*		
		Entre.	*Comp.*	*Entr.*	*Ocul.*	*Sali.*	*Hr*	*min*	*Sg*
ANN	RBFNpers.	39.03	24.90	210	80	4	0	59	55
ANN	RBFNpers.	48.91	20.48	210	150	4	0	52	4
ANN	RBFNpers.	55.11	22.09	210	200	4	0	34	14
ANN	RBFNpers.	57.12	22.49	210	200	4	0	50	6
Árbol	C&RT	61.31	36.55				0	18	45

Tabla V.5. Diferentes configuraciones de modelos RBFN en el prototipo 2

V.4.1.3. Prototipo 3

A partir de las aproximaciones presentadas en los prototipos anteriores se

entrenaron varios modelos basados en el conocimiento acerca del comportamiento previo aprendido de estos prototipos. Para estos modelos se tuvieron en cuenta la totalidad de los registros (836) de la base de datos. La información fue partida en conjuntos de entrenamiento y comprobación, o de entrenamiento, prueba y validación para verificar la bondad de los modelos obtenidos.

	Entrenamiento	*Comprobación*
C&RT	52.60%	30.12%
C5.0	42.55%	31.33%
RQEST	30.22%	32.13%

Tabla V.6. Porcentajes de clasificaciones correctas para el prototipo 3

En la Tabla V.6 se observan los porcentajes de clasificaciones correctas, tanto para los datos de entrenamiento como de comprobación para el campo *tipo daño,* con diferentes modelos de árboles de clasificación que son los que mejores resultados han presentado en los prototipos anteriores. Aunque el modelo de reglas generado a partir del algoritmo *QUEST* presenta un porcentaje algo mejor para los datos de comprobación, para los datos de entrenamiento este valor es el más bajo de los tres modelos.

Ya que en los prototipos anteriores han presentado mejor comportamiento, a partir del modelo de Árboles de Clasificación y Regresión *C&RT* se generó un nodo filtro para descartar los atributos que el algoritmo no considera importantes y se entrenó a continuación una red neuronal con el modelo RBFN, del cual se obtiene un porcentaje de datos clasificados correctamente para los datos de entrenamiento y para los datos de comprobación del 29.82% y del 26.1%, respectivamente.

A continuación se entrenó una red de Kohonen como intento de mejora de la estimación de la clasificación con lo cual se obtuvieron los siguientes resultados:

	Entrenamiento	*Comprobación*
C&RT	55.44%	29.32%
RBFN	31.32%	26.10%

Realizando partición de entrenamiento, comprobación y validación para los modelos entrenados a partir de la red de Kohonen se obtuvieron los siguientes porcentajes de clasificaciones correctas:

	Entrenamiento	*Comprobación*	*Validación*
C&RT	51.89%	25.29%	24.82%
RBFN	26.64%	24.51%	18.60%

Por otra parte, entrenando diferentes métodos de modelos de redes neuronales con la totalidad de la información se obtuvieron los siguientes resultados para cada una de las dos particiones:

	Entrenamiento	*Comprobación*
PODA	14.57%	16.06%
RBFN(1)	29.15%	32.13%
PODA EXHAUSTIVA	23.79%	30.12%
RBFN(2)	28.48%	29.72%
RBFN(3)	36.01%	29.72%
MULTIPLE	27.81%	31.73%
DINAMICO	16.25%	21.69%
RAPIDO	27.30%	31.73%
RBFN(4)	26.97%	24.90%

(1) Configuración 210-20-4, 9000 ciclos, 9min42seg
(2) Configuración 210-20-4, 18000 ciclos, 19min23seg
(3) Configuración 210-200-4, precisión estimada 29.143, 7hr47min22seg
(4) Configuración 210-20-4, precisión estimada 27.322, 66hr11min55seg

Tabla V.7. Clasificaciones correctas para diferentes métodos de ANN en el prototipo 3

Se observa que las redes neuronales mejoran en un pequeño porcentaje las estimaciones de clasificaciones para los datos de comprobación, pero para los datos de entrenamiento sigue siendo mejor el resultado obtenido con los modelos de árboles.

La precisión estimada se basa en la diferencia entre los valores pronosticados y los valores reales de los datos de entrenamiento, que se calcula de acuerdo a la siguiente formula:

$$\left(1-\left|\mathrm{Re}\,al-pronosticado\right|/\left(Rangodelcampodesalida\right)\right)*100,0\,,$$

donde *Real* corresponde al valor real del campo de salida, *pronosticado* corresponde al valor estimado por la red y *Rango del campo de salida* corresponde al rango de valores del campo de salida (el mayor valor del campo menos el menor valor). Esta exactitud se calcula en cada registro y la exactitud global corresponde a la media de valores para todos los registros de los datos de entrenamiento.

Una última prueba realizada como intento de mejorar los resultados obtenidos de clasificación, consistió en, a partir de la información original, entrenar un árbol de regresión y clasificación *C&RT* con el cual se obtuvieron porcentajes de clasificaciones correctas del 57.12% y del 28.11% tanto para los datos de entrenamiento como los de comprobación, respectivamente. A partir de éste árbol se generó un nodo filtro para reducir el conjunto de datos y descartar los campos que el algoritmo no considera importantes y se entrena un árbol C5.0 con el cual se obtiene un 74.59% de los datos clasificados correctamente; también se entrenó un árbol C5.0 utilizando aumento[4] con el cual se obtiene el 87% de los datos clasificados correctamente; a partir de este se generó de nuevo un nodo filtro para descartar reducir el conjunto de datos y descartar los campos que el algoritmo no considera importantes, a partir del cual se obtienen los siguientes resultados de clasificaciones correctas en porcentaje:

	Entrenamiento	*Comprobación*
C5.0	86.60%	32.53%
C5.0	76.72%	25.30%

[4] El algoritmo C5.0 cuenta con un método especial para mejorar su precisión denominado aumento. Este método genera varios modelos en una secuencia. El primero de ellos se crea con el procedimiento habitual. A continuación se crea otro modelo que se basa especialmente en los registros clasificados erróneamente por el primer modelo. Seguidamente se crea otro modelo que se basa en los registros clasificados erróneamente por el modelo anterior, y así sucesivamente. Por último, para clasificar los casos, se les aplica todo el conjunto de modelos de acuerdo con un procedimiento de votación ponderada para combinar los distintos pronósticos en un pronóstico global. La opción número de ensayos permite controlar el número de modelos que deben utilizarse para el modelo aumentado.

	Entrenamiento	*Comprobación*
RBFN(1)	28.14%	26.91%
C&RT	52.60%	30.12%
RBFN(2)	35.68%	28.51%
RBFN(3)	36.68%	30.92%

(1) Configuración 89-40-4, 165hr36min32seg
(2) Configuración 89-200-4, 90000 ciclos, 6hr2min13seg
(3) Configuración 89-200-4, 22hr31min12seg

V.4.2. Discusión

Basándonos en los resultados de lo descrito en el apartado V.2 del manejo de la información, se puede establecer que no existe o no se aprecia de antemano una relación explícita entre cada una de las variables objeto del estudio, lo cual nos permite inferir, que de acuerdo con lo expuesto en el marco teórico de esta tesis, así como en la revisión bibliográfica realizada, que estamos ante un problema en el cual se presenta interesante hacer uso del denominado descubrimiento de conocimiento en bases de datos.

Como se ha mencionado, gran parte del éxito en la aplicación de una técnica de minería de datos recae en la calidad de la información utilizada; de acuerdo con lo anterior la información que se ha utilizado en el desarrollo de esta tesis tiene la importancia de corresponder a un sistema de abastecimiento de agua real, aunque presenta el inconveniente de estar fuera de nuestro alcance el verificar la condición de los datos reportados en cuanto a su exactitud, veracidad, y completa que se presenta. Igualmente creemos que el estudio presentaría resultados con mayor aprovechamiento si se contara con mayor cantidad de mediciones de monitorización sobre la red. No obstante, el potencial del manejo de información que presentan las diferentes técnicas de minería de datos resalta para ser tenidas en cuenta como herramientas de apoyo a la gestión en los sistemas de abastecimiento de agua.

V.4.2.1. Prototipo 1

Con este prototipo se obtuvo una primera aproximación del conocimiento de la base de datos a partir de los algoritmos seleccionados, y por tanto un tanteo del comportamiento de estos algoritmos con base en las tareas requeridas. Como se ha mencionado anteriormente se eligió como variable objetivo el campo *Tipo de daño* por representar un aspecto importante en la gestión de los sistemas de abastecimiento de agua.

De acuerdo con los resultados obtenidos en la Tabla V.2 ampliamos la discusión de este primer prototipo basándonos en los árboles de regresión y clasificación, ya que, aparte de presentar mejores resultados, tienen la ventaja de ser interpretables al no ser cajas negras del tipo de redes neuronales; además, adicionalmente los tiempos de entrenamiento son menores para los árboles que para las redes. Algunos modelos en este prototipo se han entrenado descartando los registros con información faltante y espuria así como los valores atípicos *(outliers)*, que a su vez fueron utilizados para modelar la totalidad de la información; esto tiene el inconveniente de ser un modelo demasiado optimista, ya que se tiende a sobreestimar la información y la estimación del error es sesgada a la baja. No obstante, la información que proporcionan los diferentes modelos entrenados acerca de la relación entre los diferentes atributos con el tipo de daño reportado resulta de gran utilidad como soporte a la toma de decisiones en la gestión del abastecimiento y como un primer acercamiento hacia la información.

El primer árbol entrenado *C&RT* (69.96% de clasificaciones correctas) tiene en cuenta pocos registros debido a la selección de descarte de valores perdidos, faltantes, nulos o fuera de rango. La profundidad del árbol es de 6; tanto el modelo obtenido, como el campo objetivo, las variables de entrada y la configuración de entrenamiento, se aprecian en el Anexo 2A. Este modelo fue entrenado con los valores medios de caudales, pérdidas y presiones descartando los valores horarios. El campo que aporta menos impureza corresponde al *Material PQR* (se aprecia aquí una de las grandes bondades de este tipo de algoritmos ante

información faltante), dividiendo el árbol por una parte por los registros donde falta el valor del material, y por otra por el resto de materiales.

El resumen de cada una de las ramas para cada uno de los modelos presentados en los anexos, corresponde a la moda o valor más frecuente de cada uno de los registros de la rama. Adicionalmente, al final de cada una de las ramas se muestra la información acerca del número de registros a los que se aplica la regla (ocurrencias), así como la proporción de registros para los que la regla es verdadera (confianza).

En la Figura V.25 se visualiza la distribución de materiales respecto a cada unos de los tipos de daños (numerizados) como dato de entrada al modelo.

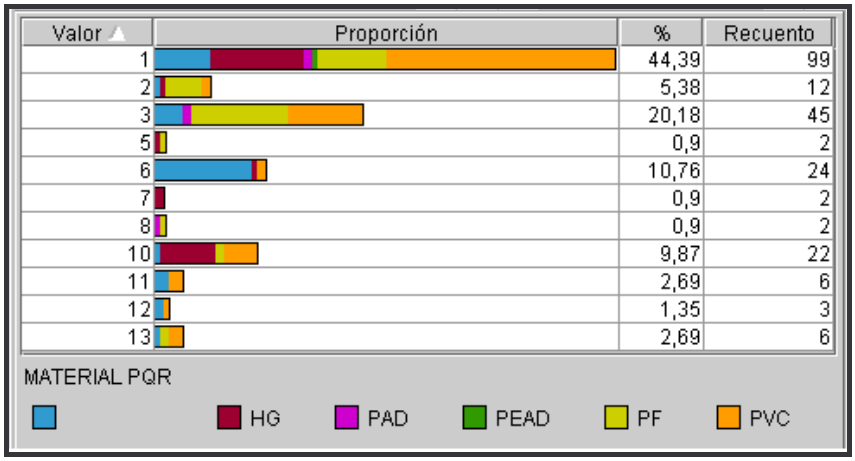


Figura V.25. Distribución de los materiales reportados en los PQR's como entrada al modelo C&RT presentado en el Anexo 2A

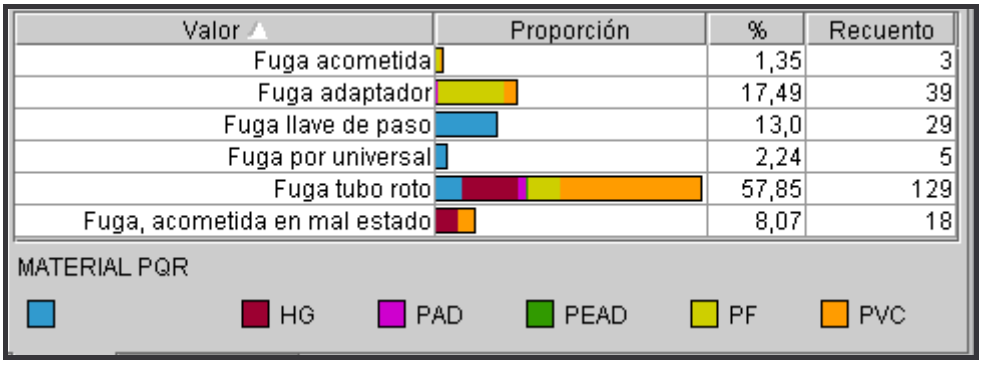


Figura V.26. Resultado de la distribución de los materiales reportados con el modelo C&RT presentado en el Anexo 2A

En la Figura V.26 se observa la distribución de los materiales luego de ejecutado el modelo, se ve que el modelo no es capaz de clasificar casi la mitad

de los daños. El tipo de daño 6 *fuga llave de paso* es clasificado en mayor proporción en los registros para los cuales no se cuenta con información de entrada.

En las dos figuras siguientes (V.27 y V.28) se visualiza la distribución de las cuadrillas de trabajo con respecto a la variable objetivo, tanto como dato de entrada al modelo, a la vez que como resultado del modelado con el algoritmo de C&RT correspondiente al modelo presentado en el Anexo 2A.

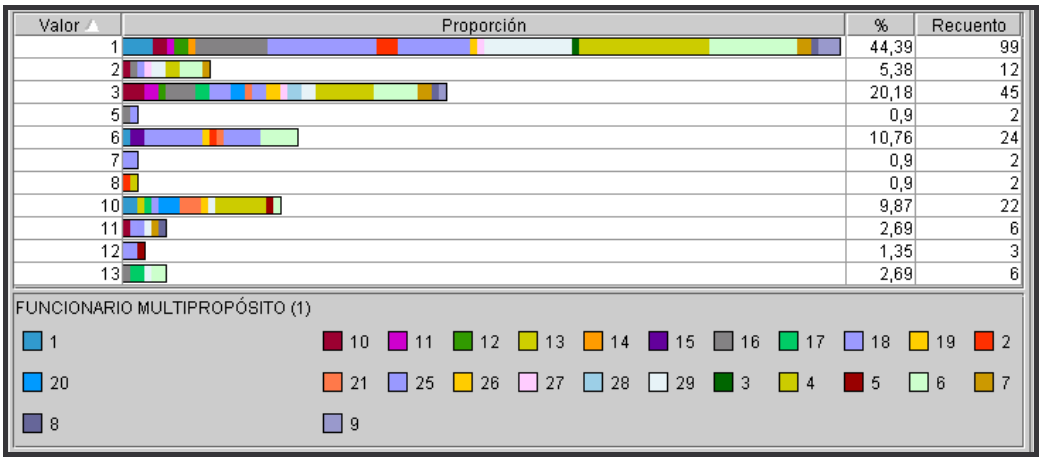


Figura V.27. Distribución de las cuadrillas de trabajo como dato de entrada al modelo C&RT presentado en el Anexo 2A

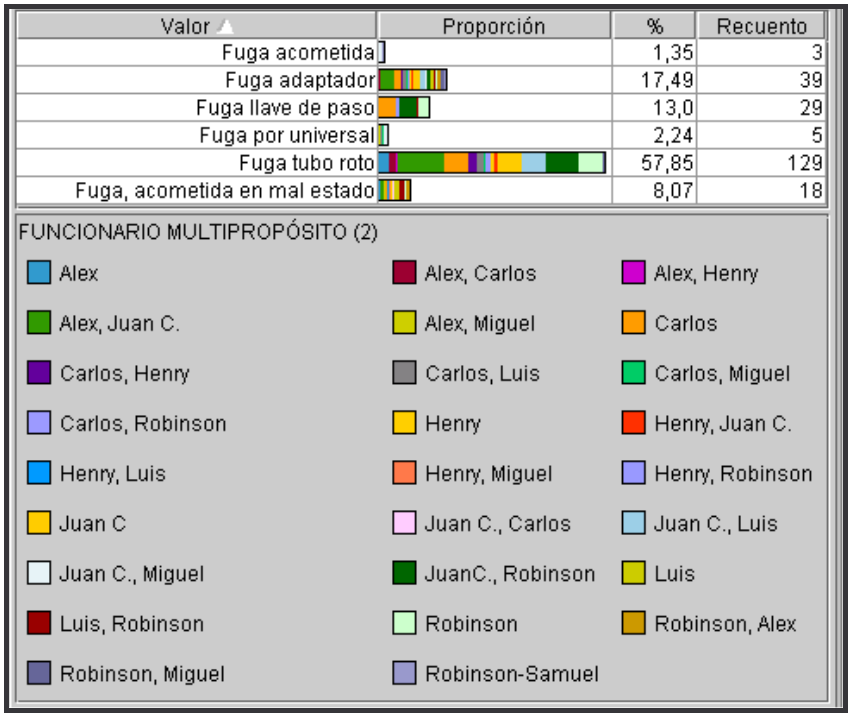


Figura V.28. Distribución de las cuadrillas de trabajo resultantes del modelo C&RT presentado en el Anexo 2A

Los registros que han sido descartados por tener datos faltantes o fuera de límite conllevan a que no se tengan en cuenta los tipos de daño 4 *fuga por pitorra,* y 9 que corresponde al caso en el que al realizar la empresa la verificación del reporte se encuentra con que no hay daño.

El árbol *C&RT* que presenta la cantidad de clasificaciones correctas más altas para el *tipo de daño* (76.23), se generó a partir de incluir campos que habían sido descartados en el modelo anterior. Para este modelo el campo que aporta menos impureza es el *Diámetro PQR* (de nuevo se aprecia el poder del algoritmo ante falta de información) tal como se aprecia en el Anexo 2B. No obstante la mejora en las clasificaciones del campo *Tipo de daño* se aprecia en este modelo que se hace uso de variables que no tienen una correspondencia con el tipo de daño, tal como los identificadores de registros; pero por otra parte se toman los valores horarios de las variables de caudales, presiones y pérdidas, descartados en el modelo anterior, que mejoran el resultado obtenido. En el Anexo 2B se presenta tanto el modelo obtenido como el resumen de configuración para el árbol entrenado.

Al probar estos dos modelos anteriores con la totalidad de la información (846 registros) se disminuye el poder clasificatorio de cada uno de ellos, tal como se puede apreciar en la Tabla V.2. En esta tabla se observa que se obtienen porcentajes de valores bien clasificados para el campo *Tipo de daño* de 30.26% y 37% con cada uno de los modelos.

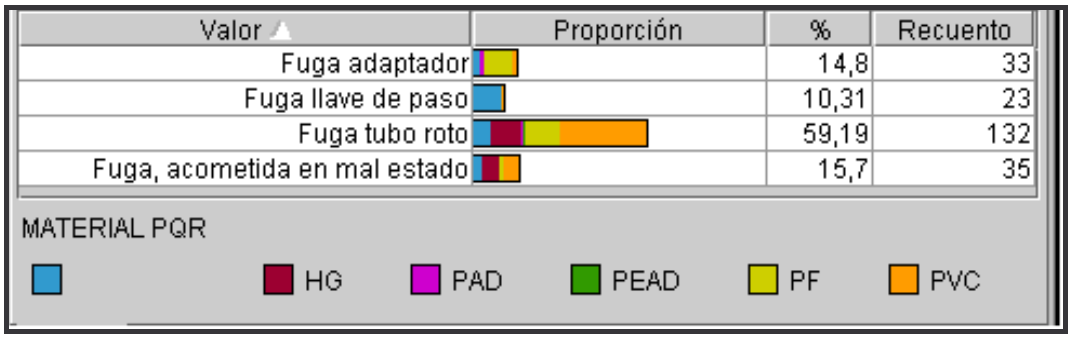


Figura V.29. Distribución de materiales reportados de acuerdo al modelo C&RT presentado en el Anexo 2C

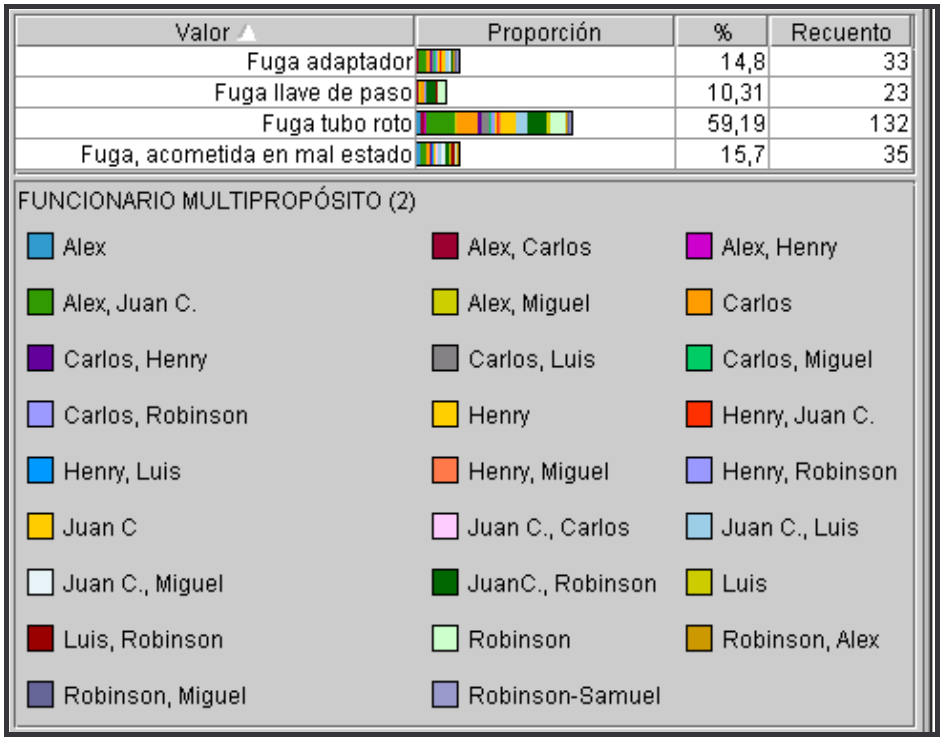


Figura V.30. Distribución de cuadrillas resultantes de acuerdo al modelo C&RT presentado en el Anexo 2C

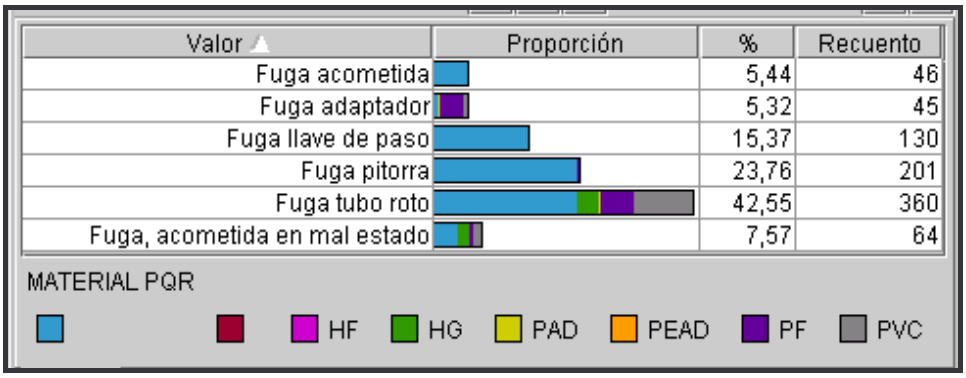


Figura V.31. Distribución de materiales reportados correspondiente al modelo C&RT resultante del Anexo 2D

Los modelos entrenados a partir de las redes de Kohonen presentan mejores resultados para cuando se hace la selección de datos perdidos o fuera de rango correspondiente al modelo *C&RT* con un porcentaje de clasificaciones correctas de 69.06%, disminuyendo cuando se realiza con la totalidad de la información (modelo *C&RT* con unas clasificaciones correctas de 60.99%). Estos modelos exhiben el inconveniente de no tener la claridad de los anteriores ya que presentan sus resultados basándose en los grupos de la red de Kohonen (ver Anexos 2C y 2D). La matriz de Kohonen generada es de 10 x 7, con una capa de entrada de 240 neuronas y una capa de salida de 70 neuronas. En las Figuras

V.29, V.30, V.31 y V.32 se pueden visualizar las distribuciones de algunas de las variables más discriminantes en estos modelos resultantes respecto a la variable tipo de daño.

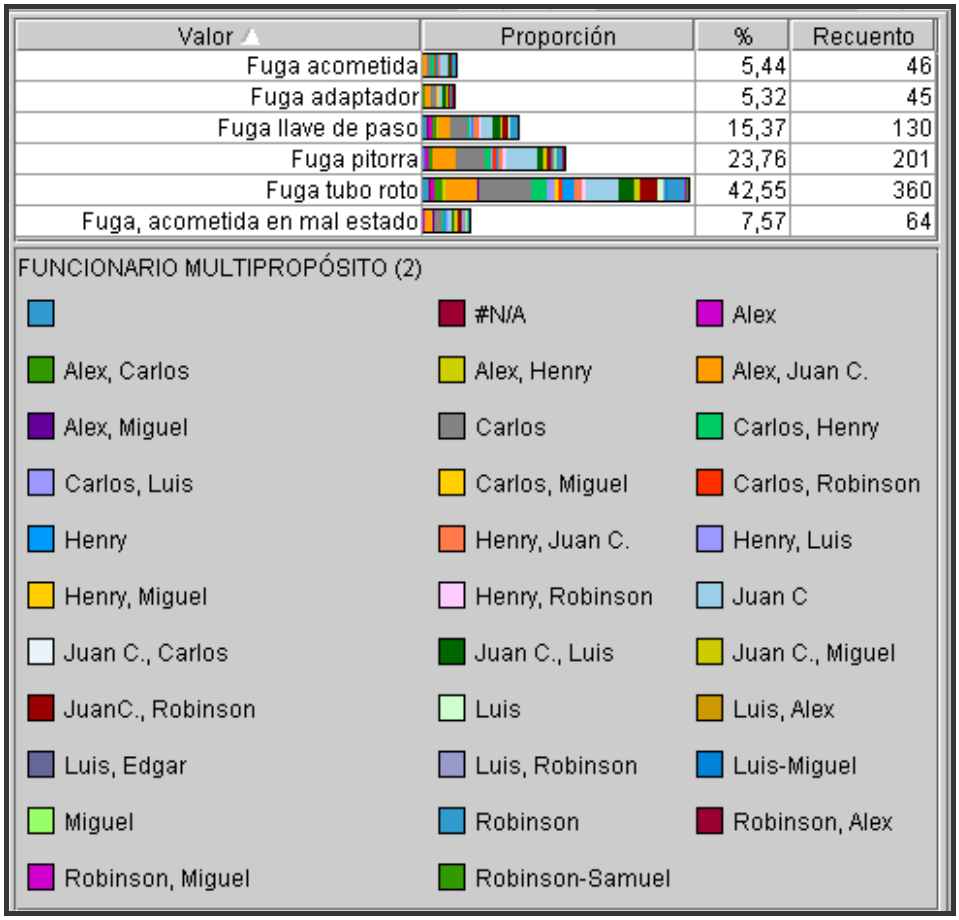


Figura V.32. Distribución de las cuadrillas de trabajo correspondiente al modelo C&RT resultante del Anexo 2D

El modelo *C&RT* que corresponde a un porcentaje de clasificaciones correctas de 40.90% se entrena haciendo uso de la totalidad de los registros; la división del árbol inicia por el campo *material PQR*, partiendo una rama por los registros donde falta el dato y la otra por los demás materiales. Igualmente uno de los campos por los que se divide el árbol corresponde al *ID* de las líneas de registros, que como ya se ha mencionado no tienen una relación directa con el campo objetivo. Este modelo permite apreciar el poder de simplificación del algoritmo en la elaboración del árbol, ya que solo 8 campos quedan incluidos finalmente. A continuación se entrenaron modelos de árbol con la totalidad de los datos descartando los campos *ID* y *Descripción,* que no son indicativos de tener

importancia en la clasificación. Se entrenó un nuevo modelo *C&RT* con el que se obtuvo un porcentaje de clasificaciones correctas de 49.53% obteniéndose la primera partición del árbol por el atributo *material PQR*, dejando una rama con los registros que no tienen dato para este campo y la siguiente con los materiales, tal como se aprecia en el Anexo 2E donde se tiene tanto el modelo obtenido como el resumen de configuración. Igualmente se entrenó un árbol de decisión mediante la Detección Automática de Interacciones mediante Chi-Cuadrado *CHAID,* tal como se puede apreciar en el Anexo 2F, que igualmente realiza la primera división del árbol a partir del campo *material PQR*; la profundidad del árbol es de 7.

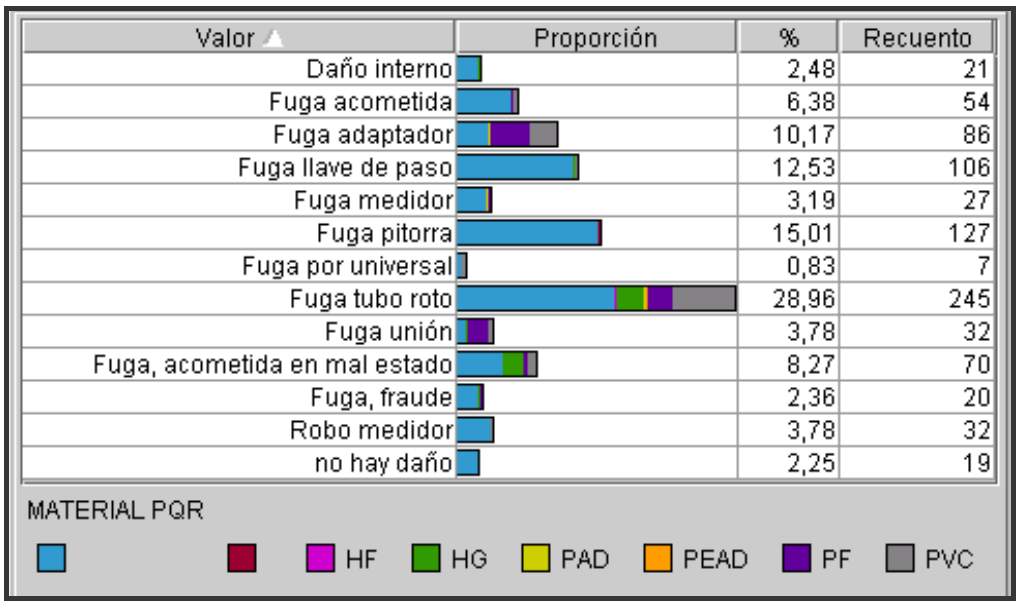


Figura V.33. Distribución de la variable material reportado entrante al modelo presentado en el Anexo 2E

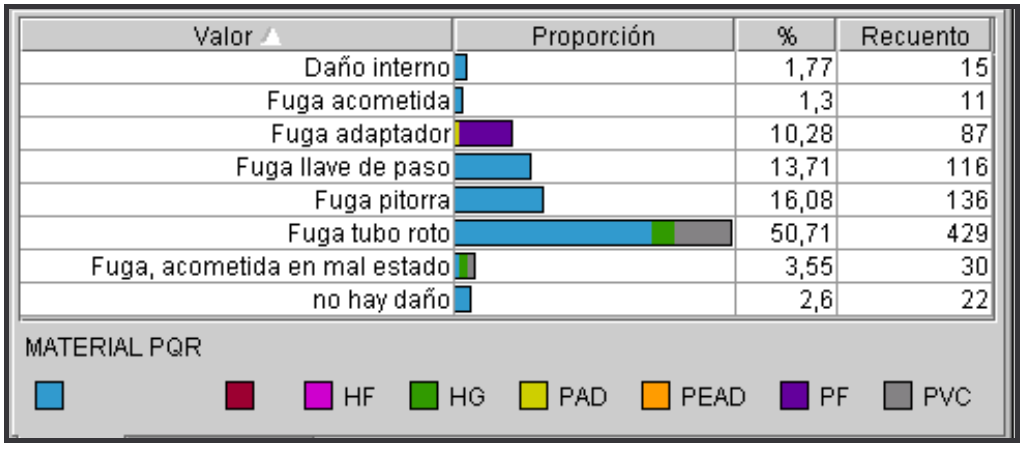


Figura V.34. Distribución de la variable material reportado resultante del modelo presentado en el Anexo 2E

En las Figuras V.33 y V.34 se visualiza como en el modelo resultante no se clasifican algunos de los tipos de daño dados como dato de entrada. Igualmente se puede apreciar una de las fortalezas de los algoritmos de árboles al clasificar la variable teniendo datos faltantes.

En las Figuras V.35, V.36, V.37, y V.38 se pueden ver las distribuciones de las variables categóricas más representativas tanto para los datos de entrada al modelo, como para los datos de resultados presentado en el Anexo 2E.

Valor	Proporción	%	Recuento
Daño interno		2,48	21
Fuga acometida		6,38	54
Fuga adaptador		10,17	86
Fuga llave de paso		12,53	106
Fuga medidor		3,19	27
Fuga pitorra		15,01	127
Fuga por universal		0,83	7
Fuga tubo roto		28,96	245
Fuga unión		3,78	32
Fuga, acometida en mal estado		8,27	70
Fuga, fraude		2,36	20
Robo medidor		3,78	32
no hay daño		2,25	19

SE CORRIGIÓ DAÑO

No Si

Figura V.35. Distribución de la variable se corrigió daño entrante al modelo presentado en el Anexo 2E

Valor	Proporción	%	Recuento
Daño interno		1,77	15
Fuga acometida		1,3	11
Fuga adaptador		10,28	87
Fuga llave de paso		13,71	116
Fuga pitorra		16,08	136
Fuga tubo roto		50,71	429
Fuga, acometida en mal estado		3,55	30
no hay daño		2,6	22

SE CORRIGIÓ DAÑO

No Si

Figura V.36. Distribución de la variable se corrigió daño resultante del modelo presentado en el Anexo 2E

La variable *"se corrigió daño"* corresponde a información derivada de los formularios de los reportes de PQR's a partir de la solución dada al reporte. En algunos casos se presentó que por motivos ajenos a la empresa no se solucionara el daño, ya sea porque el dueño de la vivienda no se encontraba en el momento de la reparación, o porque le correspondiera a éste hacerse cargo de arreglar el desperfecto provocado, y por tanto corresponde al valor *NO* de la variable. En general se aprecia variedad en la distribución de las cuadrillas en cuanto a los tipos de daños, es decir no se tiene una tendencia de la influencia de las mismas hacia un tipo específico de daño.

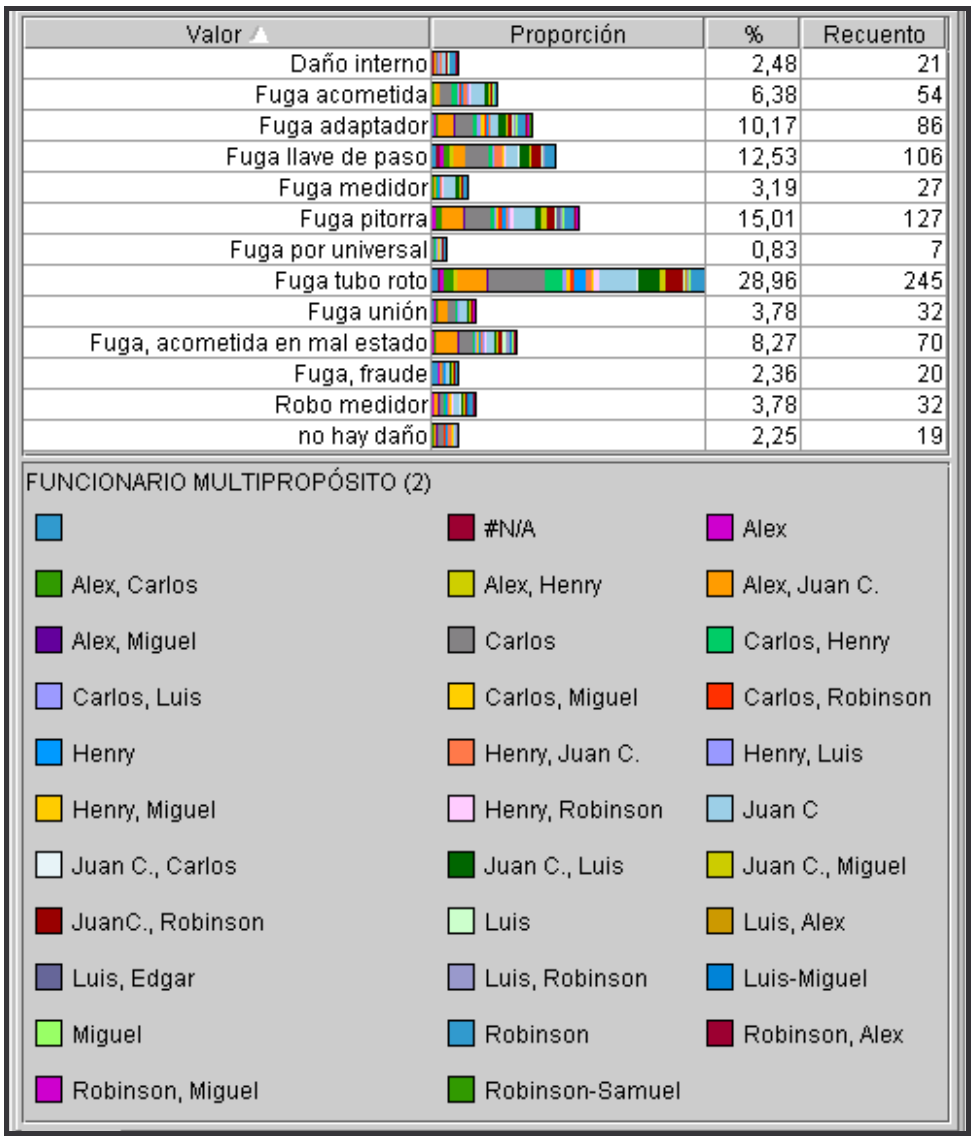
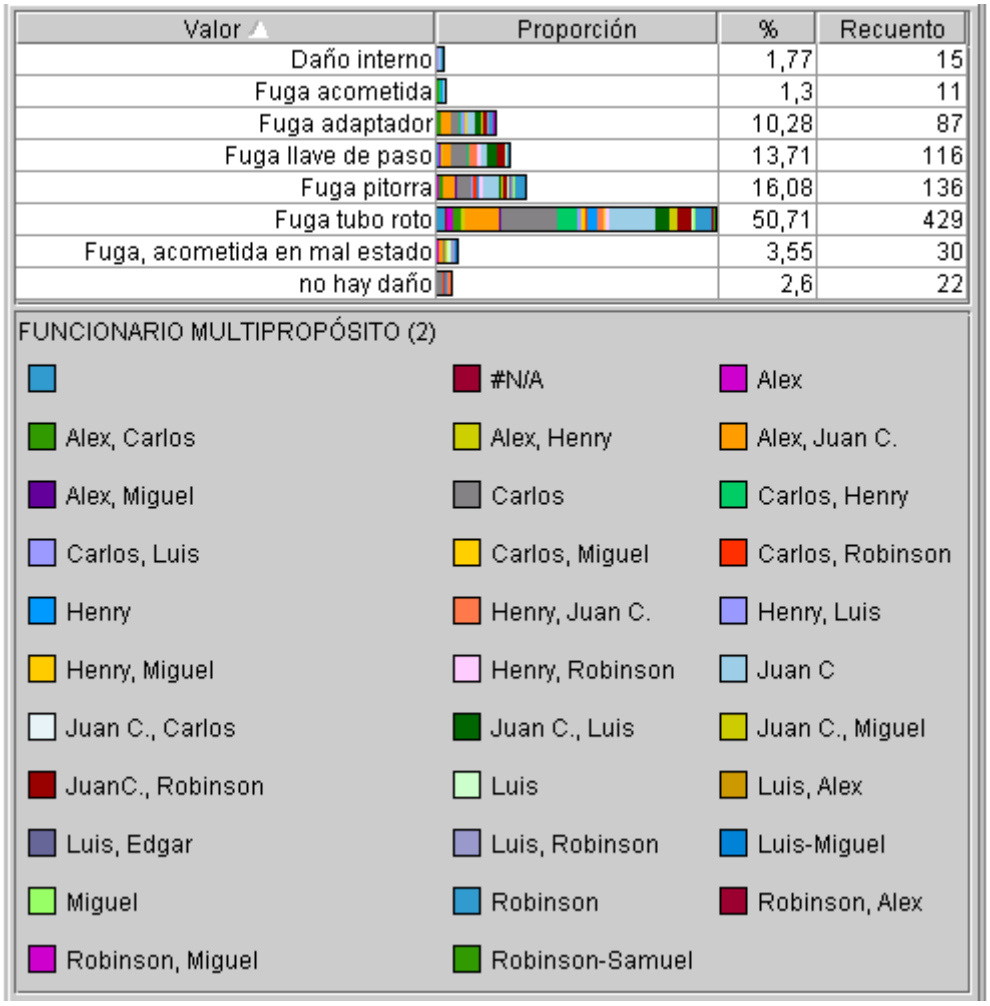


Figura V.37. Distribución de la variable cuadrilla de trabajadores entrante al modelo presentado en el Anexo 2E



Valor	Proporción	%	Recuento
Daño interno		1,77	15
Fuga acometida		1,3	11
Fuga adaptador		10,28	87
Fuga llave de paso		13,71	116
Fuga pitorra		16,08	136
Fuga tubo roto		50,71	429
Fuga, acometida en mal estado		3,55	30
no hay daño		2,6	22

Figura V.38. Distribución de la variable cuadrilla de trabajadores resultante del modelo presentado en el Anexo 2E

V.4.2.2. Prototipo 2

En los diferentes modelos entrenados en este apartado se efectúa la validez de los diferentes modelos, dividiendo la información ya sea en datos de *entrenamiento-comprobación* ó *entrenamiento-comprobación-validación*, es decir, la estimación del error de la clasificación en cada uno de los modelos es realizada a partir de la selección aleatoria de un conjunto de datos (70%) para entrenar el modelo y otro conjunto (30%) para comprobarlo. En caso de utilizar tres conjuntos se utiliza 60% de entrenamiento, 30% de comprobación y 10% de validación.

Los algoritmos basados en reglas presentan mejor comportamiento en cuanto a los resultados obtenidos para las clasificaciones, que los modelos

basados en algoritmos de redes neuronales artificiales, de acuerdo con los resultados presentados en el numeral V.4.1.2. No obstante, se consigue una mejoría al entrenar diferentes tipologías de modelos de redes neuronales con el algoritmo *RBFN* en los datos de entrenamiento, consiguiendo clasificaciones similares a los modelos *C&RT*, aunque tanto para los modelos de redes como los árboles las clasificaciones para los datos de comprobación no son tan satisfactorios, obteniéndose mejores resultados para estos últimos.

Aparte de esta mejora en la predicción de los modelos de redes neuronales, se destaca que los tiempos de entrenamiento son menores para los modelos de clasificación y regresión que en las redes neuronales; por tanto parecen más atractivos además de por la ya citada mayor explicités en la interpretación de resultados que ofrecen los árboles.

V.4.2.3. Prototipo 3

En los modelos entrenados en este apartado los campos *tipo daño* y *funcionario multipropósito* fueron numerizados para disminuir el tamaño del conjunto de datos y tener expresiones más cortas para los resultados obtenidos. Además, se refinan los modelos de acuerdo con lo visto en los dos prototipos anteriores.

Modelo	*Tiempo de entrenamiento*			*Clasificaciones correctas (%)*			*Neuronas en capas*		
	Hr	*min*	*Seg*	*Entre.*	*Comp.*	*Valid.*	*Entr.*	*Ocul.*	*Sali.*
C&RT	0	29	4	52.6	30.12				
RBFN	14	8	44	29.82	26.10		129	40	4
C&RT	0	24	46	55.44	29.32				
RBFN	24	32	13	31.32	26.10		122	40	4
C&RT	0	25	55	51.89	25.29	24.42			
RBFN	20	10	39	26.64	24.51	18.60	122	40	4
C5.0	0	0	11	42.55	31.33				
RQUEST	0	0	28	30.32	32.13				

Modelo	*Tiempo de entrenamiento*			*Clasificaciones correctas (%)*			*Neuronas en capas*		
	Hr	*min*	*Seg*	*Entre.*	*Comp.*	*Valid.*	*Entr.*	*Ocul.*	*Sali.*
PODA	0	11	39	19.77	18.88		179	20-15-10	4
RBFN	0	9	42	29.15	32.13		210	20	4
Exhaustiva	0	17	16	23.79	30.12		210	30-20	4
RBFN	0	19	23	28.48	29.72		210	20	4
RBFN	7	47	22	36.01	29.72		210	200	4
Múltiple	0	12	18	27.81	31.73		210	2	4
Dinámico	0	8	32	27.81	31.73		210	3-4	4
Rápido	0	12	4	27.81	31.73		210	20-15-10	4
RBFN	66	11	55	26.97	24.90		210	20	4
C&RT	0	35	38	57.12	28.11				
C5.0	0	0	27	74.59					
C5.0	0	2	32	87.00					
C5.0	0	0	15	76.72	25.30				
C5.0-R	0	1	27	86.60	32.53				
RBFN	165	36	32	28.14	26.91		89	40	4
C&RT	0	12	55	52.60	30.12				
RBFN	6	2	13	35.68	28.51		89	200	4
RBFN	22	31	12	36.68	30.92		89	200	4

Tabla V.8. Tiempo de entrenamiento, porcentaje de clasificaciones y número de neuronas para los diferentes modelos entrenados en el prototipo 3

En la Tabla V.8 se presentan los resultados de cada uno de los modelos entrenados en este apartado: el tiempo de entrenamiento; el porcentaje de clasificaciones correctas para los datos de entrenamiento y comprobación, así como, para los de validación en caso de que se haya realizado esta tercera partición; y el número de neuronas en cada una de las capas de redes neuronales entrenadas.

De nuevo, de acuerdo con los resultados obtenidos de los diferentes modelos, se aprecia que, en general, los modelos de árboles funcionan mejor que los modelos de caja negra de redes neuronales, no solo en cuanto al porcentaje

de clasificaciones correctas obtenidas por los diferentes modelos sino también por los tiempos de entrenamiento que, como se observa en la Tabla V.8, para el mejor modelo de árbol es de 1 minuto y medio, mientras que la mejor red neuronal se ha entrenado más de 20 horas obteniéndose casi un 50% de clasificaciones correctas para los datos de entrenamiento en el árbol y casi un 2% en los datos de comprobación.

A diferencia de los prototipos anteriores 1 y 2, en este apartado el modelo entrenado de árbol que mejor resultados presenta es el C5.0 con aumento, con un 86.6% de los datos clasificados correctamente para los datos de entrenamiento, aunque la mejoría en los datos de comprobación no es tan significativa, debido a que este método tiende a sobreestimar los datos en el entrenamiento.

Por ejemplo, de los modelos de árboles presentados en la tabla anterior, se aprecia en el Anexo 2G que el modelo C5.0 que presenta un porcentaje de clasificaciones del 76.72% para los datos de entrenamiento tiene una profundidad de árbol de 12 y cuenta con aproximadamente 20 campos en el árbol; pero su tamaño en líneas es mayor que el árbol *C&RT* (Anexo 2H) cuyo porcentaje de clasificaciones correctas es de 52.60% para los datos de entrenamiento pero, en cuanto a las clasificaciones correctas para los datos de comprobación, es del 25.30% para el primero (C5.0) y del 30.12% para los datos *C&RT*. Para este árbol se tiene una profundidad de 13, y el modelo esta conformado por 16 campos. Aquí se ve que el modelo C5.0 intenta sobreestimar tanto los valores, que se pierde generalidad; por tanto su poder clasificatorio para los datos de comprobación pierde eficacia.

Como se aprecia en la tabla anterior hay un conjunto de reglas generado a partir del modelo C5.0 que mejora los resultados tanto en el entrenamiento como la prueba (86.60%, 32.53%). Para este caso, se hizo uso del método del aumento propio del algoritmo, para la mejora de la precisión. Básicamente, con el modelo se generan varios modelos en una secuencia. El primero se genera de forma habitual, en el siguiente se tienen en cuenta fundamentalmente los registros mal clasificados en el modelo anterior, y luego un tercero que tienen en cuenta los mal

clasificados en el segundo, y así sucesivamente. Al final se hace un voto ponderado para clasificar los registros. Para nuestro caso, al aumentar o disminuir la cantidad de aumentos no se obtienen mejoras en la precisión para los datos de prueba. El inconveniente que presenta este modelo es la extensión del mismo, y como se menciona con anterioridad, no se presentan mejoras significativas con respecto a árboles más cortos.

En la Figura V.39 se visualiza la distribución de los conjuntos de entrenamiento y comprobación con respecto a los tipos de daño (numerizados) correspondientes al modelo presentado en el Anexo 2G.

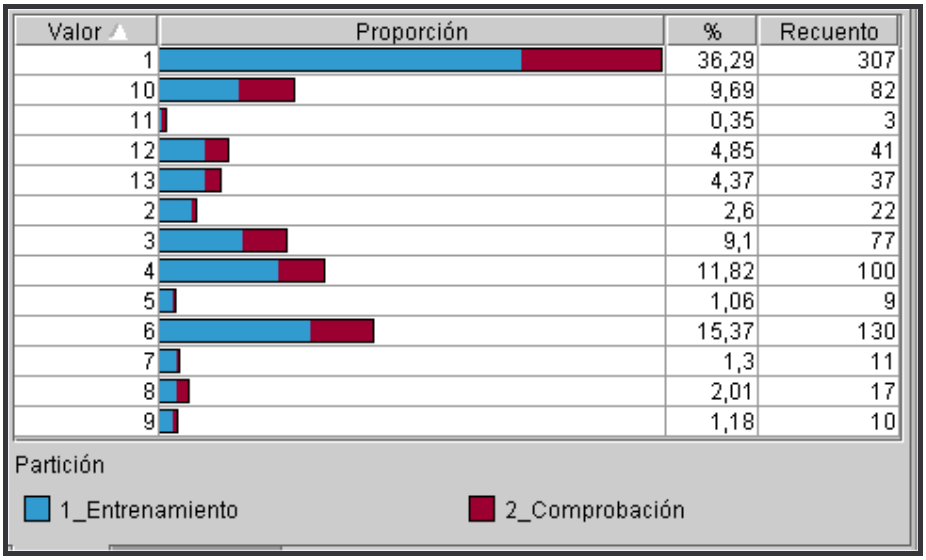


Figura V.39. Distribución de los conjuntos de entrenamiento y comprobación para el modelo mostrado en el Anexo 2G

A continuación se visualizan las distribuciones resultantes de las variables más significantes del modelo mostrado en el Anexo 2G (Figuras V.40, V.41, V.42, V.43, y V.44). Estas distribuciones se realizaron con respecto a la variable respuesta de *tipo daño* numerizada.

Se puede apreciar que en este modelo se presentan clasificaciones para todos los tipos de daños reportados en los formularios de PQR's. De nuevo se manifiesta una de las ventajas de la utilización de algoritmos basados en árboles ya que se clasifican las variables así no se tenga la información completa, como ocurre con los materiales reportados.

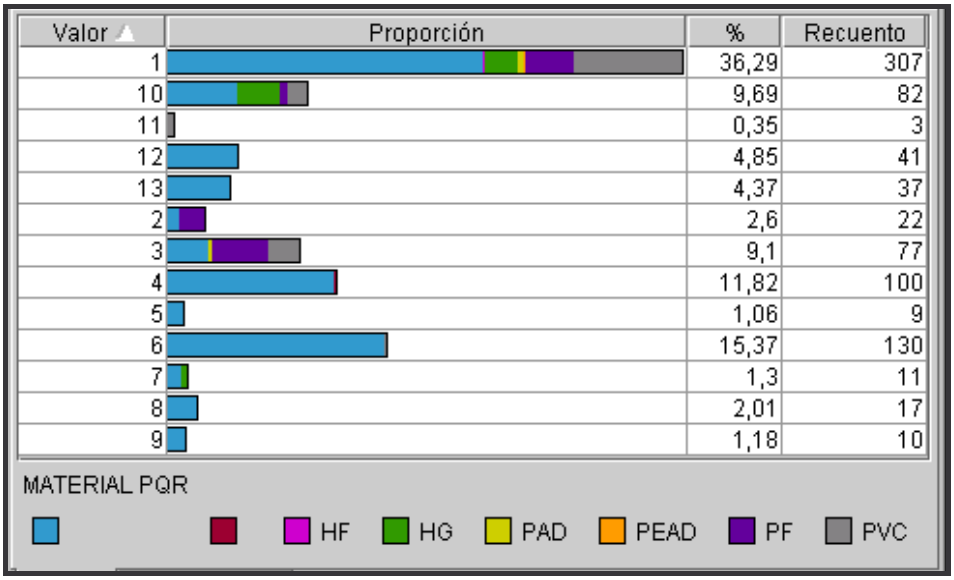


Figura V.40. Distribución de la variable material reportado resultante del modelo presentado en el Anexo 2G

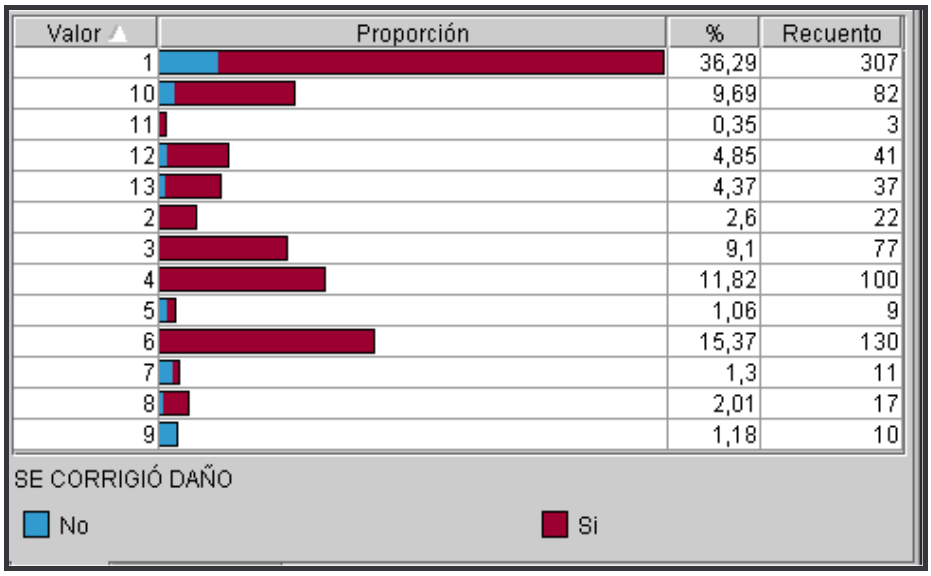
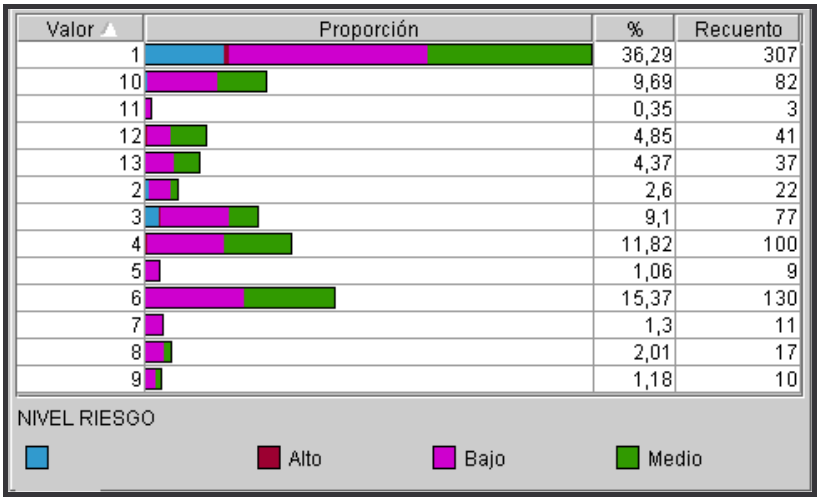


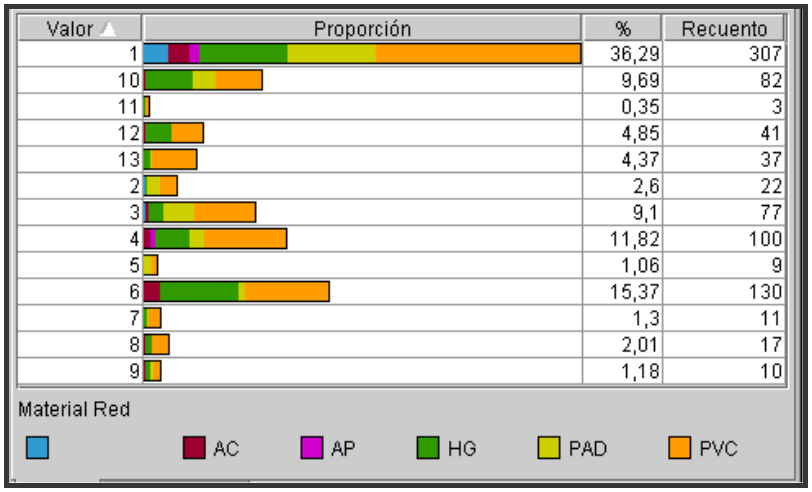
Figura V.41. Distribución de la variable se corrigió daño resultante del modelo presentado en el Anexo 2G

En cuanto a la variable *"se corrigió daño"*, se aprecia que para la mayoría de los reportes la respuesta fue afirmativa. En cuanto al nivel de riesgo la mayor parte de la información resultante se ubica en zonas con algún nivel, consideración importante a tener en cuenta para la gestión del sistema de abastecimiento de agua. La distribución de materiales en la red muestra que la mayor parte de los tipos de daños se encuentran ubicados en zonas más cercanas a materiales plásticos.



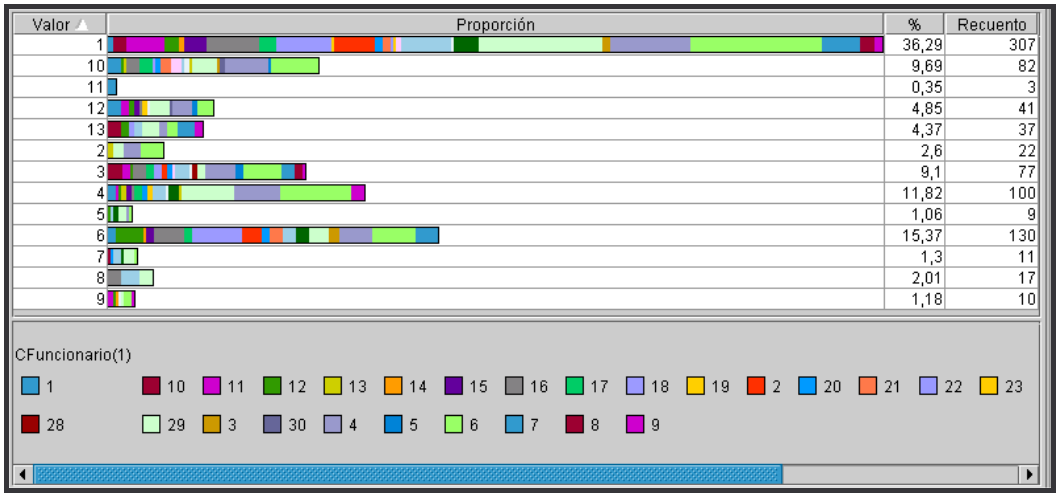
Valor	Proporción	%	Recuento
1		36,29	307
10		9,69	82
11		0,35	3
12		4,85	41
13		4,37	37
2		2,6	22
3		9,1	77
4		11,82	100
5		1,06	9
6		15,37	130
7		1,3	11
8		2,01	17
9		1,18	10

Figura V.42. Distribución de la variable nivel de riesgo resultante del modelo presentado en el Anexo 2G



Valor	Proporción	%	Recuento
1		36,29	307
10		9,69	82
11		0,35	3
12		4,85	41
13		4,37	37
2		2,6	22
3		9,1	77
4		11,82	100
5		1,06	9
6		15,37	130
7		1,3	11
8		2,01	17
9		1,18	10

Figura V.43. Distribución de la variable material en la red resultante del modelo presentado en el Anexo 2G



Valor	Proporción	%	Recuento
1		36,29	307
10		9,69	82
11		0,35	3
12		4,85	41
13		4,37	37
2		2,6	22
3		9,1	77
4		11,82	100
5		1,06	9
6		15,37	130
7		1,3	11
8		2,01	17
9		1,18	10

Figura V.44. Distribución de la variable cuadrilla de trabajadores resultante del modelo presentado en el Anexo 2G

El modelo *C&RT* del Anexo 2H muestra que, aunque para los datos de prueba el poder clasificatorio no es del todo bueno, la rama correspondiente a los materiales presentes en los reportes presenta unos valores de confianza bastante aceptables para el modelo. La clasificación muestra cómo para diámetros superiores a una pulgada (25.4mm), el tipo de daño que se presenta es el de las fugas por tubo roto, mientras que para diámetros inferiores, se aprecia una mayor dependencia con respecto a las cuadrillas de funcionarios, que en principio no parece guardar lógica, pero que podría formar parte de un estudio acerca de la incidencia que tienen las tuberías reparadas por cada cuadrilla, si se repiten fallos.

La precipitación es otro factor que genera ganancia en información para el algoritmo en diámetros menores de media pulgada, aunque a este respecto se debe resaltar que los datos con los que contamos no son exactamente del casco municipal; así que este es otro factor a tener en cuenta como recomendación a mejorar, pero no deja de ser interesante la asociación entre las fugas por roturas de tuberías y fugas por un accesorio debidas a la lluvia. Para los diámetros superiores a media pulgada, el atributo de mayor ganancia es la longitud del tramo, para la cual igualmente se tienen pérdidas por roturas en tuberías y por uniones. La pérdida hidráulica del modelo, corresponde a otro de los atributos que producen ganancia de información.

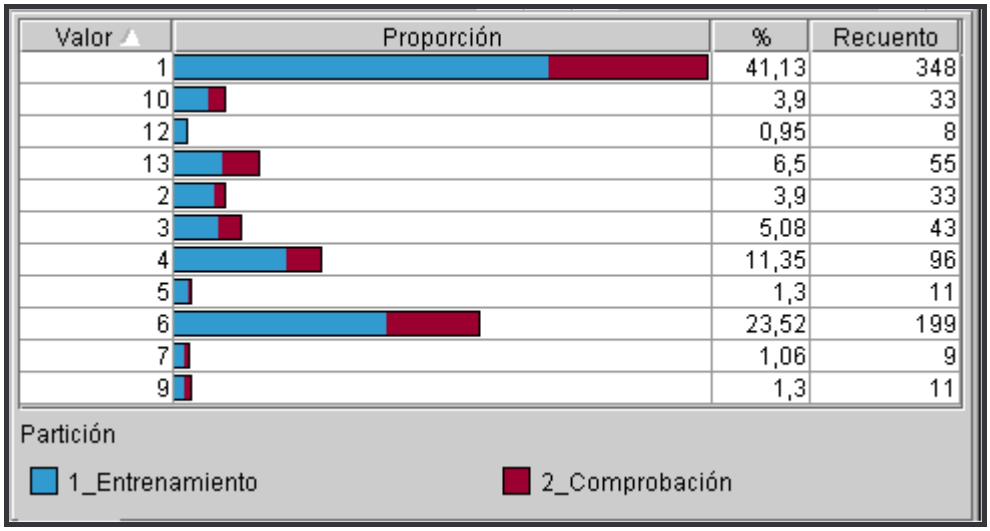


Figura V.45. Distribución de los conjuntos de entrenamiento y comprobación para el modelo mostrado en el Anexo 2H

En la Figura V.45 se visualiza la distribución de los conjuntos de

entrenamiento y comprobación con respecto a los tipos de daño (numerizados) correspondientes al modelo presentado en el Anexo 2H.

A continuación se visualizan las distribuciones resultantes de las variables más significantes del modelo mostrado en el Anexo 2H (Figuras V.46, V.47, V.48, y V.49). Estas distribuciones se realizaron con respecto a la variable respuesta de *tipo daño* numerizada.

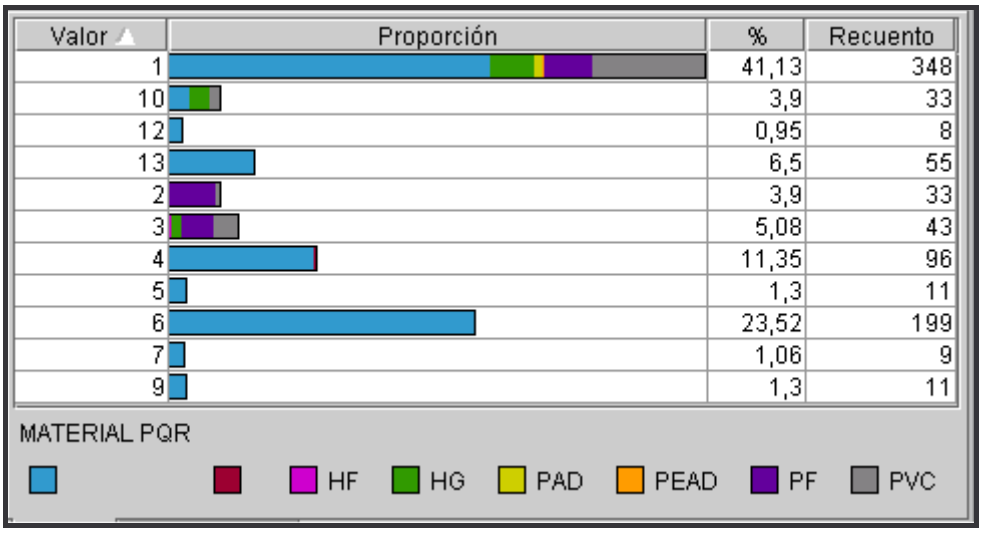


Figura V.46. Distribución de la variable material reportado resultante del modelo presentado en el Anexo 2H

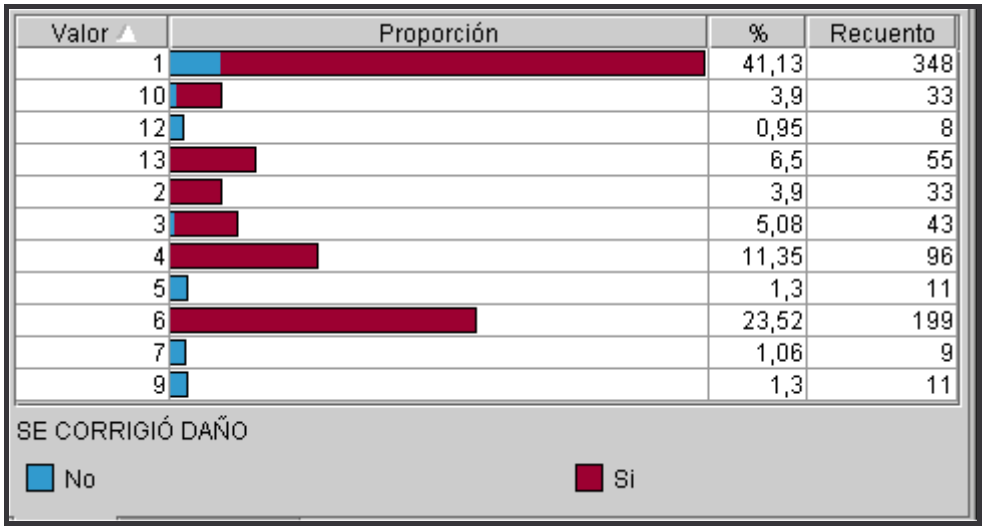


Figura V.47. Distribución de la variable se corrigió daño resultante del modelo presentado en el Anexo 2H

En general se puede visualizar que el algoritmo de clasificación y regresión C&RT clasifica mayor tipos de daño 1, 6 y 13 que el modelo de árbol C5.0.

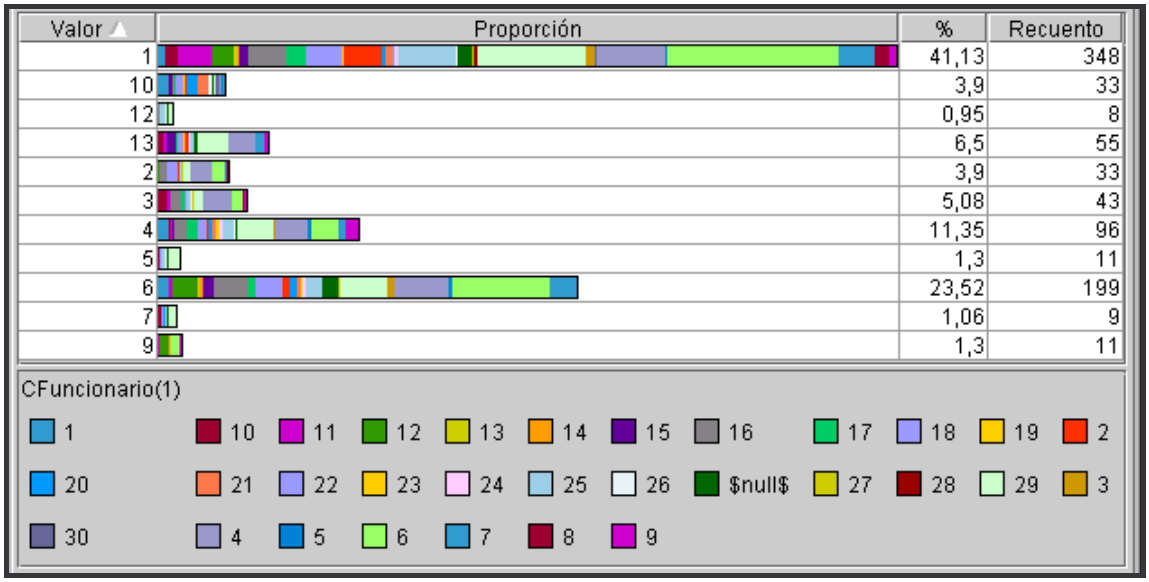


Figura V.48. Distribución de la variable cuadrilla de trabajadores resultante del modelo presentado en el Anexo 2H

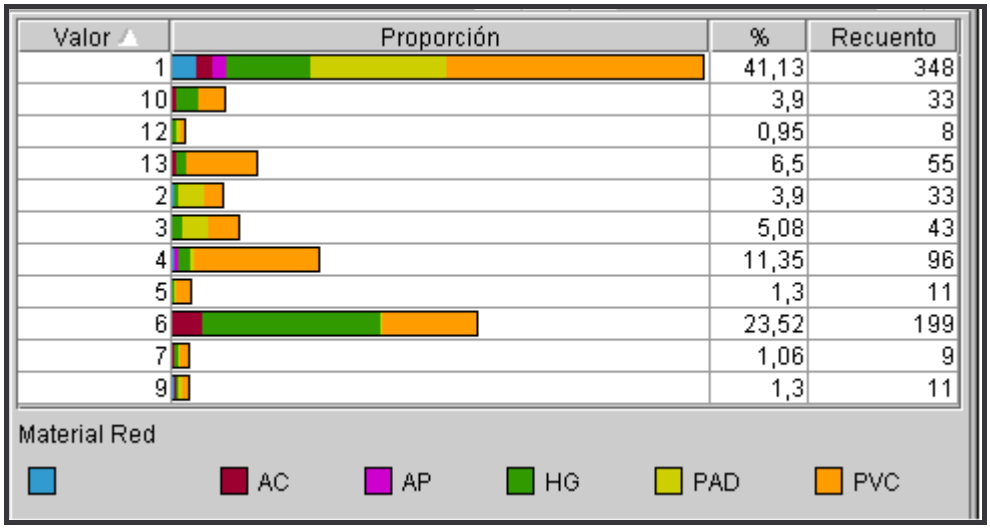


Figura V.49. Distribución de la variable material en la red resultante del modelo presentado en el Anexo 2H

V.5. Aplicación a la gestión del abastecimiento

Los prototipos descritos anteriormente han generado una serie de modelos, que pueden ser implementados en sistemas expertos que dan lugar al establecimiento de reglas, que al ser consultadas en el sistema provocan que una decisión sea tomada. Entre las características de gestión para el problema aquí abordado se tienen las siguientes:

1. En el modelo presentado en el Anexo 2A correspondiente al prototipo 1 se puede apreciar cómo el modelo le indica al gestor que en general se tienen dos grupos clasificatorios de tuberías de acuerdo con el material reportado en el formulario de PQR. Es interesante que una de las ramas del árbol o regla corresponda únicamente a cuando no se dispone del dato, mientras que la otra toma el resto de materiales. Este es uno de los poderes clasificatorios que presentan estos algoritmos simples de reglas de decisión y clasificación, y que los hace supremamente interesantes para aplicar en caso reales por su sencillez de implementación e interpretación. Este modelo aunque resumido debido al descarte de la información, presenta una primera aproximación a la problemática planteada, lo cual es importante a la hora de tomar decisiones respecto a reemplazos o rehabilitaciones, sin necesidad de ser un experto hidráulico.

2. El modelo obtenido en el prototipo 1 y presentado en el Anexo 2B le indica al gestor la importancia de la ubicación geográfica en la clasificación del daño. Aunque este modelo tiene en cuenta información que no es relevante para la gestión de la red tal como los identificadores de los registros, no deja de ser llamativo que se obtengan mejores valores en cuanto a clasificaciones correctas que las obtenidas con el modelo anterior. Aunque este hecho puede ser provocado por dejar el campo descripción como variable de entrada al modelo (esta variable corresponde a la versión del usuario acerca del daño presentado), el algoritmo relaciona este dato con el del concepto técnico del cual se derivó la variable tipo daño. Este modelo por otra parte muestra la importancia de los valores horarios de caudales, pérdidas y presiones descartados en el modelo presentado en el Anexo 2A.

3. Los Anexos 2C y 2D (prototipo 1) pueden presentar cierta dificultad de interpretación para quien gestione la red de abastecimiento (Redes de Kohonen - cajas negras); pero si se cuenta con un sistema experto que descifre los resultados, estos modelos de menor tamaño en cuanto a líneas de resultados igualmente son de utilidad para efectuar un control y manejo del abastecimiento. Aquí, se representan grupos que tienen características similares y, por tanto, pueden ser conglomerados en características similares con respuestas igualmente

parecidas.

4. Tanto en el Anexo 2E como en el Anexo 2F (prototipo 1) se presentan árboles o reglas de clasificación y decisión; en las cuales, la partición principal se realiza de nuevo por el material reportado en el formulario de daños para la red de abastecimiento. Desde el punto de vista del modelo hidráulico se distingue en este modelo que, aunque son variables tenidas en cuenta para la construcción del mismo, tanto las demandas en los nudos como las rugosidades en las tuberías, aparecen con poca significancia en el caso de la primera y no aparecen en el caso de las rugosidades en la formulación final del árbol. Como ya se ha mencionado con anterioridad, estas variables generan incertidumbre en el modelo. Por tanto, las variables que el modelo encuentra como importantes para efectuar las clasificaciones del tipo de daño pueden ser fácilmente establecidas, ya sea por mediciones o por reconocimiento de campo.

5. En los modelos obtenidos en el prototipo 3 y presentados en los Anexos 2G y 2H se aprecia la habilidad del algoritmo de los árboles de clasificación ante datos faltantes, ya que la partición principal se hace a partir del campo *material PQR,* tomando a un lado como aporte de significancia valores del campo que no presentan dato. Aunque, no obstante, el ideal es contar con toda la información con el fin de obtener resultados de mayor precisión, ante datos faltantes el algoritmo presenta resultados con su valor estadístico de confianza. Desde el punto de vista hidráulico, se aprecia la poca importancia que presentan para el algoritmo los campos de rugosidad y demandas. En cambio, el campo del agrupamiento de cuadrillas de trabajo presenta cierta importancia en la partición, lo cual es interesante si, por ejemplo, se quiere plantear un sistema de control de gestión de las reparaciones efectuadas sobre la red. Las presiones en la red son tenidas en cuenta en algunas de las divisiones del árbol; por tanto, podría pensarse en un programa de control de presiones en estos puntos, y corroborar su influencia en los daños ocasionados. Por otra parte, el árbol es sugerente en cuanto a gestionar los diámetros y materiales de las tuberías en la red, de acuerdo con los daños reportados en cada uno de ellos.

6. El sistema de reglas generado a partir del algoritmo C5.0 con aumento, corresponde a un sistema extenso de clasificaciones que van mejorando su precisión, pero cuyo resultado final corresponde al mismo árbol del anexo 2G.

V.6. Recomendaciones

A la vista de los resultados obtenidos, que aunque son interesantes por lo novedoso de la metodología de la minería de datos que permite deducciones no obvias, y porque claramente pueden servir de apoyo en los sistemas de abastecimiento de agua, no desconocemos que estos podrían ser sustancialmente más atractivos para su aplicabilidad práctica. También resulta claramente sugerente que ésta metodología propuesta puede ser abordada en cualquier tipo de abastecimiento, lo que permite generalizar el esquema aquí planteado a infinidad de problemáticas dentro de la operación y gestión de los sistemas de abastecimiento de agua.

En cuanto a lo primero anteriormente dicho, estamos convencidos de que los resultados obtenidos a partir del análisis realizado podrían ser mejorados si la información básica a partir de la cual se realiza el estudio se toma y recolecta de forma estructurada para los fines del estudio; pero como un primer intento de abordar la temática, el estudio en un esquema real de abastecimiento nos ha proporcionado, un enriquecimiento en cuanto a la teoría del descubrimiento de información a partir de los datos, así como una confirmación de que el objetivo planteado como una suposición o creencia de que las herramientas nos podrían ser útiles, está sustentado.

Como ya se ha mencionado, la información que manejan los entes encargados de la distribución del agua pareciera tener un cierto nivel de oscurantismo y privacidad, lo cual hace bastante difícil su consecución, cuando no es que se tiene poca información del funcionamiento real de la red, como el caso bajo el cual trabajamos esta tesis, donde no se cuenta con mediciones reales o es escasa esta información.

Por tanto, se plantea como inevitable la necesidad de establecer una metodología para la toma y almacenamiento de los datos, necesarios para la utilización de técnicas de descubrimiento de conocimiento en bases de datos de sistemas de abastecimiento de agua. En este punto se establece la fuerte relación sinérgica entre las tecnologías de la información y la comunicación (TIC) y la minería de datos.

Esta relación debe abarcar todos los componentes del sistema que, a partir de un gran almacén de datos, permita ejecutar un análisis de todos los elementos del mismo por medio de las técnicas de minería de datos. Estos elementos deben comprender temas de producción, relaciones con los clientes, costos, gestión, calidad del agua, etc. Los datos para formar el almacén deben abarcar sistemas de adquisición de datos en tiempo real, medición y gestión del sistema de calidad del agua, sistemas de información geográfica (GIS) de las redes de tuberías, sistema de gestión de los cargos o derechos sobre el agua, sistema de información financiera, sistema de recursos humanos, y el modelo de la red.

Con esta información y la utilización de las técnicas mencionadas podemos ser capaces entre otros, como lo mencionan Guo y Ding (2009) de predecir:

- aquellos factores que tienen influencia sobre el periodo de servicio de las redes de tuberías a través del análisis de la base de datos,

- las tuberías que puedan ser un "cuello de botella", que puedan causar serios problemas en caso de que se produzca un daño en ellas a través de un análisis de la topología de la red,

- aquellas tuberías que se puedan mantener en el tiempo en el proyecto a través del análisis de la relación del registro histórico del mantenimiento de las redes de tuberías y los datos de la red de tubería,

- aquellas tuberías que tengan más posibilidad de sufrir daño después de un año o año y medio a través del análisis de las señales de los sensores instalados en las redes de tuberías, realizar un esquema combinado de

optimización minería de datos espacial y los datos diarios del abastecimiento.

Adicionalmente a esta parte técnico-operativa de la red, la metodología propuesta de la utilización de las herramientas de minería de datos, también sirve de apoyo en lo que es la imagen de la empresa, ya que se pueden mejorar y optimizar procedimientos, que permitan tanto una mejor gestión técnico-económica del sistema (principalmente con la reducción de fugas), como una mejor relación cara al cliente.

Capítulo VI

Conclusiones y desarrollos futuros

VI.1. Conclusiones generales

1. Se ha presentado en esta tesis un desarrollo exhaustivo de la aplicación de técnicas avanzadas en el manejo y tratamiento de datos conocido en la bibliografía como *descubrimiento de conocimiento en bases de datos,* y uno de sus pasos constitutivos, la *minería de datos,* como aplicación práctica en la gestión de sistemas de abastecimiento de agua. Adicionalmente, se ha presentado una revisión detallada del estado del arte en la última década, que coincide con un mayor auge de sistemas informáticos más eficientes y capaces de desarrollar estas tareas de forma rápida y eficaz.

2. Basándonos en esta revisión bibliográfica, se puede afirmar que el desarrollo aplicado a casos reales en sistemas de abastecimiento de agua de las técnicas aquí expuestas, no ha sido ampliamente desarrollado; situación la cual se debe principalmente, en nuestra opinión, a la dificultad para la consecución de la información y el secretismo de la misma por parte de los entes encargado del manejo y gestión de estos sistemas, a pesar de estar tratándose de un recurso natural público. Por tanto, con el desarrollo de esta tesis se quiere divulgar este conocimiento, con el fin de presentar los beneficios y posibilidades de actuación de la metodología aquí presentada hacía la gestión de los abastecimientos de agua.

3. Contar con la disponibilidad de información suficiente, y con la calidad necesaria para desarrollar cualquier técnica de descubrimiento de información a partir de datos, tal como se ha mencionado en varios capítulos de esta tesis, corresponde a la parte crucial para obtener resultados con la exactitud y calidad necesarias y requeridas. Por consiguiente debe haber compromiso por parte de todos los entes involucrados al plantearse una tarea de KDD, tanto por los expertos en el dominio como por los expertos en la tarea. Estamos convencidos de que estas herramientas, pueden proporcionar utilidad en el momento de resolver problemáticas de gestión y operación de la red, y que solo con el concurso de

quienes disponen de la información, se pueden llevar a cabo los desarrollos para la implementación sistemática de metodologías basadas en el análisis de la información.

4. Como se ha visto en parte del desarrollo del estado del arte, así como en el ejercicio práctica que se ha presentado en esta tesis, la aplicación de las herramientas aquí planteadas permite ya sea el desarrollo de modelos predictivos o de clasificaciones que satisfacen la gestión óptima del sistema; entre otros el relativo al establecimiento de políticas de rehabilitación de las redes de acuerdo al funcionamiento del sistema. Para el caso del abastecimiento aquí estudiado, se tiene un índice de agua no facturado demasiado alto, correspondiendo al 49.03% para el año 2006, según datos de la propia empresa encargada de la operación y gestión del sistema. Teniendo en cuenta que parte del mayor desembolso presupuestario en la gestión de los sistemas de abastecimiento de agua es debido al tema de mantenimiento de las redes, cualquier intento que se ejecute con el fin de optimizar este costo no admite discusión en cuanto al beneficio que se produce para el aprovechamiento del recurso natural hídrico y, por tanto, en el manejo ambiental del sistema, así como en un uso sostenible de las redes de tuberías del abastecimiento.

5. Un aspecto que se ultima después del trabajo realizado es que los beneficios económicos que se pueden obtener al hacer uso de las técnicas de minería de datos como apoyo a la gestión de los abastecimientos de agua, van más allá al mero establecimiento de modelos o patrones de gestión. Estas herramientas involucran cuantiosos ahorros, entre otros, por ejemplo, en el tiempo de inspecciones de la red o en el establecimiento de medidas anticipadas que permiten la previsión de los fondos y recursos necesarios para su ejecución, o en el establecimiento de reglas y modelos que reemplacen la heurística utilizada para ciertos procedimientos. Adicionalmente, a la vez en que se implementan medidas que permitan la previsión de posibles fallos del sistema, se está fortaleciendo la relación con los clientes al minimizar las posibles quejas por deficiencias en el suministro, cuidando la imagen de la empresa. Igualmente,

recalcamos el hecho de que esta metodología es aplicable a cualquier sistema de abastecimiento.

6. La metodología planteada permite tomar decisiones tanto en tiempo real como a medio y largo plazo, porque el modelo permite ser alimentado con nueva información de reportes, así como de mediciones en tiempo real sobre la red. Estas decisiones básicamente están enfocadas hacía la oportunidad que tiene el gestor de tener un control sobre la red; y poder prever con base en la información disponible, las medidas o acciones a acometer para evitar la interrupción del servicio de abastecimiento de agua potable.

VI.2. Conclusiones específicas

1. Para el caso práctico aquí expuesto el comportamiento de los árboles de decisión y clasificación presenta mejores resultados que los obtenidos al aplicar técnicas de redes neuronales, lo cual creemos es debido a la sencillez del algoritmo de los árboles pero a su vez a su gran poder clasificatorio. Al mezclar tanto variables numéricas como categóricas, la codificación de estas variables categóricas en el caso de las redes neuronales tiene un mal comportamiento comparado con los árboles de clasificación. Otro aspecto que resalta es el consumo de tiempo para el entrenamiento y generación de los modelos en cada uno de los algoritmos, obteniéndose resultados muy superiores con la generación de reglas.

2. El error de entrenamiento decrece cuando se incrementa el número de neuronas ocultas, pero el error en los datos de prueba se incrementa. Esto muestra que las redes neuronales tienden a sobreajustar los valores cuando se utilizan un mayor número de neuronas en la capa oculta.

3. El hacer uso de recursos como las redes de Kohonen para eliminar atributos no significativos, no aporta mejoras sustanciales en el modelado. En cuanto a los diferentes modelos de redes neuronales entrenados, el que mejor comportamiento presenta corresponde a las redes de función de base radial. De la

misma forma los algoritmos *CHAID* y *QUEST* implementados en Clementine 9.0 para la generación de árboles para realizar análisis de segmentación y clasificación, no aportan mejoras en el modelado con respecto a los algoritmos *C&RT* y *C5.0.*

4. Los modelos entrenados en el prototipo 1 demuestran que descartar los valores horarios de los caudales en los tramos, pérdidas y presiones; y tomar en vez de éstos el valor promedio de cada uno empeora los resultados obtenidos.

5. Los resultados presentados a partir del modelado en el prototipo 3 dividen los árboles a partir del material reportado, notándose la carencia de información con respecto a este atributo. Aunque, en promedio, el poder clasificatorio del algoritmo no es demasiado alto para el conjunto de prueba, no deja de ser indicativo de la eficacia de la metodología propuesta, ya que la división formada por los materiales de los que se cuenta reporte, tiene unos índices de correlación bastante aceptables en el modelo.

6. De acuerdo con lo anteriormente dicho, y para el caso del ejemplo aquí estudiado, se puede apreciar que para diámetros superiores a 1 pulgada (25.4 mm), se reportan en su mayoría rotura de tuberías, ante lo cual se debe plantear un estudio más profundo acerca de las posibles causas de este daño. Aquí de nuevo se insiste, en la necesidad de contar con una mejor organización y toma de la información, ya que se hace necesario conocer el tipo de suelo, el tipo de tráfico rodado, y la edad de la tubería, entre otros.

7. La variable brillo solar no parece que pudiera influir en los daños que se presentan en la red, pero tal como se puede ver en los resultados obtenidos en el Anexo 2G el algoritmo encuentra relación entre esta variable y el campo objetivo propuesto.

8. Uno de los pasos de tratamiento de la información como se menciona en la tesis es el de la selección de variables. Para nuestro caso, esta selección básicamente se realizó de forma manual, ya que la información con la que se cuenta es inherente al funcionamiento de la red. No obstante se destaca el hecho

de que los algoritmos utilizados hacen uso de la información que consideran más significativa en su ejecución; ya que los métodos de partición recursiva como los árboles utilizan una función de evaluación para seleccionar el atributo que presenta la mejor capacidad para discriminarse entre las clases.

VI.3. Conclusiones de recomendaciones de gestión

1. Se nota la presencia de un mayor tipo de daños correspondientes a las fugas por tubo roto en toda la gama de materiales. Este hecho es sugerente para la realización de una investigación acerca de las posibles causas a que esté provocando el daño. Adicionalmente, se presentaron mayor cantidad de número de daños en las conexiones domiciliarias, más difíciles de controlar por parte del gestor de la red, lo cual es indicativo de implementar políticas para el control de este tipo de fugas.

2. El poder y eficacia de la metodología propuesta radica principalmente en que el sistema de la red de abastecimiento se *auto conoce* a partir de su propia información. Adicionalmente, se puede actualizar constantemente con más información que permite un mejor conocimiento de su comportamiento. Las incertidumbres de la información de entrada a los modelos, y la subjetividad en parámetros de las formulaciones son soslayadas, ya que quedan inmersos dentro del propio comportamiento del sistema, y dejan de tener la importancia del momento de su aplicación.

3. De acuerdo con los diferentes modelos obtenidos en cada uno de los prototipos planteados, se puede apreciar que, en general las herramientas de aprendizaje automático y, en particular los algoritmos de árboles y reglas de clasificación y decisión, corresponden a modelos de sumo interés como herramientas de apoyo en la gestión de los sistemas de abastecimiento de agua potable. El poder trabajar con la información que se va generando en el propio sistema del abastecimiento, nos permite tener el modelo real del mismo, evitando el tener que trabajar con suposiciones o criterios empíricos de especialistas, así

como, con las incertidumbres planteadas en el modelado.

4. Por otra parte, la metodología, que puede ser aplicada en todo el sistema de abastecimiento, permite una gestión y control total de forma ordenada, ya que al contar con toda la información almacenada, y por medio de las herramientas propuestas, se puede dar el manejo adecuado a dicha información y obtener respuestas a cuestionamientos que podamos plantearnos. Es interesante, la sinergia provocada entre diferentes avances científicos y tecnológicos, que nos permiten plantearnos la utilización sistemática del manejo y tratamiento de datos; y el hecho de que se pueda contar con información en tiempo real, garantiza una mejor actuación en el sentido de inmediatez de gestión sobre el sistema.

VI.4. Desarrollos futuros

A pesar del gran avance del tema propuesto en esta tesis aplicado a varias ramas de la tecnología, información y ciencia en las dos últimas décadas, como se ha expuesto, el desarrollo de su aplicabilidad al tema de los abastecimientos de agua es escaso; por tanto se presenta un abanico de posibilidades de explotación hacia progresos en el tema del agua.

Tal como se menciona en la introducción de esta tesis, con la presentación de un modelo transparente basado en datos reales observados, se produce una mayor objetividad en la toma de decisiones acerca de la gestión de los sistemas de abastecimiento de agua. Pero para que este objetivo sea plenamente conseguido, se debe promulgar por una mayor implicación y colaboración entre los organismos encargados de la gestión y manejo de las redes de abastecimiento, y la comunidad del descubrimiento de conocimiento en bases de datos. Esto, sin duda, aportará incalculables beneficios a la gestión del recurso natural por medio de la distribución del agua.

La necesidad de gestión del recurso hídrico en los tiempos en que hay zonas del planeta en que el agua empieza a escasear, es un tema de la agenda de reuniones científicas de todo el mundo. Por tanto la motivación para la utilización

de las técnicas expuestas en este trabajo se sustenta por sí sola. En esta gestión deben tenerse en cuentas aspectos relacionados desde una estimación más real del consumo de agua en las redes, como el evitar las pérdidas y fugas del líquido desde la producción hasta la entrada del contador, por parte del ente encargado de su distribución, así como la gestión de concienciación hacia el usuario acerca del uso racional del recurso.

Específicamente, el itinerario a seguir dentro de lo que consideramos como posibles líneas de investigación futuras debe comprender los siguientes pasos:

- En primera instancia realizar una divulgación amplia del tema ante los entes y organizaciones encargadas del manejo de los abastecimientos del agua para intentar conseguir una mayor implicación que permita abordar el tema desde un punto de vista más real.

- Adecuar la forma en que se recolecta la información de un sistema de abastecimiento de agua, generar la inquietud acerca de que cualquier dato tomado durante las diferentes fases del sistema es de gran utilidad. Implementar la utilización de sistemas de información geográfica para facilitar el manejo y presentación de la información. Para el caso específico de la base de datos utilizada en esta tesis se debe mejorar la monitorización en tiempo real de la red, así como la implementación de un registro histórico de la red en el cual se puedan conocer datos referentes a tipos de tubos, edades de tuberías, fallos en cada tubería, etc.

- Como se concluye del trabajo realizado la mayor parte de los daños se localiza sobre las conexiones domiciliarias en la red. Esta tesis se ha planteado como una primera aproximación hacia la utilización de herramientas para el manejo de información en redes de abastecimiento de agua, por tanto tal como se indica en el párrafo anterior se debe mejorar la información de campo, para así poder realizar unas estimaciones más certeras de lo que está afectando el sistema, en nuestro caso los daños que ocurren en la red.

- Ampliar el abanico de posibilidades de aplicación de las diferentes

técnicas aquí expuestas a todos los procesos involucrados de toma, conducción, purificación y distribución de los sistemas de abastecimiento de agua.

- Profundizar en el afinamiento de las técnicas aquí expuestas, para dotarlas de mayor generalidad al momento de requerir su aplicación en cualquier sistema de abastecimiento de agua.

- Realizar una integración de todos los componentes constitutivos del sistema de abastecimiento de agua, desde los físicos y técnicos, hasta el componente humano; y por medio de las herramientas de manejo de los datos lograr manejar y gestionar el sistema de forma eficiente, con base en la propia información producida por el mismo. Como se ha presentado en este documento, se tienen diversidad de técnicas y se han expuesto diferentes problemáticas, pero desde un punto de vista ingenieril no se ha planteado una integración única del sistema para prevenir, detectar y sugerir un manejo ordenado y óptimo del abastecimiento. Es decir, establecer programas de rehabilitación, reducción de pérdidas y fugas de agua, control del funcionamiento hidráulico del sistema, gestión económica del sistema, relaciones con los clientes, entre otros.

- Ampliar el estudio de diferentes técnicas o herramientas para el manejo de la información en el caso de contar con información adecuada. Esto incluye todo el proceso de CRISP-DM y buscar nuevas alternativas de algoritmos que se adapten a esta información para mejorar los resultados.

Capítulo VII

Bibliografía

Acuerdo No. 015. (2000). *"Plan Básico de Ordenamiento Territorial del municipio de Calarcá para el periodo comprendido entre los años 2000 y 2009"*, Diciembre 31 de 2000.

Adjei, O.; Li Chen; Heng-Da Cheng; Cooley, D.H.; Cheng, R.J.; Twombly, X. (2001). *"A fuzzy search method for rough sets in data mining"*, IFSA World Congress and 20th NAFIPS International Conference, 2001. Joint 9th, Volume: 2, Pages 980-985.

Aesche, B.G.; Simpson, A.R. (1994). *"Optimization of pipe networks using an evolution strategy"*. Department of Civil and Environmental Engineering, University of Adelaide, Research Report No. R120. 51pp.

Afsarmanesh, H.; Camarinha-Matos, L.M.; Martinelli, F.J. (1997). *"Federated Knowledge Integration and Machine Learning in Water Distribution Networks"*, In Proceedings of the IFIP/IEEE/OE International Conference on Integrated and Sustainable Industrial Production ISIP'97, Chapman and Hall publication, Lisbon, Portugal, May 1997, 121-140.

Afshar, M.H.; Rajabpour, R. (2009). *"Application of local and global particle swarm optimization algorithms to optimal design and operation of irrigation pumping systems"*, Irrigation and Drainage, 58, 321-331.

Ali, O.G.; Chen, Y.T. (1999). *"Design quality and robustness with neural networks"*, IEEE Transactions on Neural Networks., Volume 10, Number 6, Pages 1518-1527.

Alonso, C.D.; Pérez-García, R.; Izquierdo, J.; Montalvo, I. (2009). *"Factores de fiabilidad y eficiencia en la toma de decisiones para la rehabilitación de tuberías"*, IX SEREA - Seminario Iberoamericano sobre Planificación, Proyecto y Operación de Sistemas de Abastecimiento de Agua, ST-10, 1-15. Valencia, Spain.

Ampazis, N.; Perantonis, S.J.; Taylor, J.G. (2001). *"A dynamical model for the analysis and acceleration of learning in feedforward networks"*, Neural Networks, Volume 14, Pages 1075-1088.

An, A.; Shan, N.; Chan, C.; Cercone, N.; Ziarko W. (1995). *"Discovering Rules from Data for Water Demand Prediction"*, Proceedings of IJCAI'95 Workshop on Machine Learning in Engineering, Montreal, Canada. 15pp.

An, A.; Shan, N.; Chan, C.; Cercone, N.; Ziarko, W. (1997). *"Applying knowledge discovery to predict water-supply consumption"*, IEEE Intelligent Systems & Their Applications, Volume 12, Number 4, pages 72-78.

Apte, C.; Weiss, S. (1997). *"Data mining with decision trees and decision rules"*, Future Generations Computer Systems, Volume 13, Number 2-3, Pages 197-210.

Babovic, V.; Keijzer, M.; Rodríguez, D.; Harrington J. (2001a). *"An Evolutionary*

Approach to Knowledge Induction: Genetic Programming in Hydraulic Engineering", in Proceedings of the World Water & Environmental Resources Congress, May 21-24, 2001, Orlando, Fl. USA http://www.d2k.dk, 10pp.

Babovic, V.; Drécourt, J.; Keijzer, M.; Hansen, P. (2001b). *"Modelling of Water Supply Assets: A Data Mining Approach"*, D2K Technical Report 1000-1, 2000, February 6, 2001, 40pp.

Babovic, V.; Drécourt, J.; Keijzer, M.; Hansen, P. (2002). *"A data mining approach to modelling of water supply assets",* Urban Water, Vol. 4, 401-414.

Baecher, G.B. (2006). *"Mitigating Water Supply System Vulnerabilities"*, K.V. Frolov and G.B. Baecher (eds.), Protection of Civilian Infrastructure from Acts of Terrorism, 149–157. Springer.

Bargiela, A.; Pedrycz, W. (2001). *"Classification and clustering of granular data"*, IFSA World Congress and 20th NAFIPS International Conference, 2001. Joint 9th, Volume: 3, 2001. 1696-1701.

Bautista, R.; Millan, M.; Díaz, J.F. (1999). *"An efficient implementation to calculate relative core and reducts"*, Fuzzy Information Processing Society, NAFIPS. 18th International Conference of the North American, Pages 791-794.

Berry, M.J.A.; Gordon, L. (2000). *"Mastering Data Mining"*, Wiley, Second Edition, 494pp.

Berthold, M.; Hand, D.J. (Eds.) (2003), *"Intelligent Data Analysis – An Introduction"*, 2nd rev. and extended ed., Springer- Verlag, 514pp.

Bessler, F.T.; Savic, D.A.; Walters, G.A. (2003). *"Water reservoir control with data mining"*, Journal of water resources planning and management, Vol. 1, 26-34.

Bhattacharya, B.; Lobbreccht, A.H.; Solomatine, D.P. (2003). *"Neural Networks and Reinforcement Learning In Control of Water Systems"*, Journal of Water Resources Planning and Management, November - December 2003, 458-465.

Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J.; (1984). *"Classification and regression trees"*, Monterey, CA: Wadsworth and Brooks-Cole.

Buchheit, R.B.; Garrett, J.H. JR; Lee, S.R.; Brahme R. (2000). *"A Knowledge Discovery Framework for City Civil Infrastructure: A Case Study of the Intelligent Workplace"*, Engineering with Computers (16), 264-274.

Burn, S.; DeSilva, D.; Eiswirth, M.; Hunaidi, O.; Speers, A.; Thornton, J. (1999). *"Pipe Leakage - Future Challenges and Solutions"*, Proceedings of Pipes' 99, Wagga Wagga Australia, 12th - 14th October, 1999. 18pp.

Camariha-Matos, L.M.; Martinelli, F.J. (1999), *"Application of Machine Learning in water distribution networks assisted by domain experts"*, Journal of

Intelligent and Robotic Systems, Vol. 26, 325-352.

Carpenter, G.A. & Grossberg, S. (2003), *"Adaptive Resonance Theory"*, In Michael A. Arbib (Ed.). The Handbook of Brain Theory and Neural Networks, Second Edition (pp. 87-90). Cambridge, MA: MIT Press

Castro, J.L.; Castro-Schez, J.J.; Zurita, J.M. (2001). *"Use of a fuzzy machine learning technique in the knowledge acquisition process"*, Fuzzy sets and systems, Volume 123, Pages 307-320.

Catalina, G.; A. Recuperado el 20 de Julio de 2009, *"Introducción a las redes neuronales"*, de http://www.gui.uva.es/login/login/13/redesn.html

CENICAFÉ. (2008). *"Anuario Metereológico Cafetero 2006"*, Federación Nacional de Cafeteros de Colombia. Centro Nacional de Investigaciones de Café. Chinchiná (Colombia). 564p. Recuperado de www.cenicafe.org el 24 de Julio de 2008.

Cestnik, B.; Kononenko, I.; Bratko, I. (1987). *"ASSISTANT 86: A knowledge-elicitation tool for sophisticated users"*. Proceedings of the Second European Working Session on Learning (pp. 31-45). Bled, Yugoslavia: Sigma Press.

Chan, C.C. (1991). *"Incremental learning of production rules from examples under uncertainty: a rough set approach"*, International Journal of software engineering and knowledge engineering, Volume 1, Number 4, Pages 439-461.

Chan, CC. (1998). *"A rough set approach to attribute generalization in data mining"*, Information Sciences., Volume 107, Number 1-4, Pages 169-176.

Chapman, P.; Clinton, J.; Khabaza, T.; Reinartz T.; Wirth, R. (1999). *"The CRISP-DM Process Model"*.

Chauchat, JH.; Rakotomalala, R. (2001). *"Sampling Strategy for Building Decision Trees from Very Large Databases Comprising Many Continuous Attributes"*, Chap. 9, in Instance Selection and Construction - A Data Mining Perspective, H. Motoda & H. Liu ed., Kluwer Academic Publishers, Pages 179-188.

Cios, K.; Pedrycz, W.; Swiniarski, R. (1998). *"Data Mining: Methods for Knowledge Discovery"*, Kluwer Academic Publishers, Third Printing 2000, 495pp.

Cisty, M. (2000). *"Advantages of using genetic algorithms over deterministic methods in optimal design of the water networks rehabilitation"*, Proceedings of ALGORITMY 2000 Conference on Scientific Computing, 293-300.

Clementine® 9.0 Algoritms Guide. (2004). Integral Solutions Limited. EE.UU.

Clerc, M. (1999). *"The swarm and the queen: towards a deterministic and adaptive particle swarm optimization"*, Proc. 1999 *ICEC*, Washington, D.C.,

1951-1957.

Cook, DJ.; Holder, LB. (2000). *"Graph-based data mining"*, IEEE Intelligent Systems & Their Applications, Volume 15, Number 2, 22 pp.

Coulibaly, P.; Anctil, F.; Aravena, R.; Bobee, B. (2001). *"Artificial neural network modelling of water table depth fluctuations"*, Water Resources Research, Vol. 37, No. 4, 885-896.

Christodoulou, S.; Agathokleous, A.; Deligianni, A.; Aslani, P. (2009). *"Risk based asset management of water piping networks using neurofuzzy Systems"*, Computers, Environment and Urban Systems March 2009, Vol. 33, No. 2, 138-149.

Dandy, G.C.; Engelhardt, M. (2001). *"Optimal scheduling of water pipe replacement using genetic algorithms"*, Journal of Water Resources Planning & Management-ASCE, Vol. 127, No. 4, 214-223.

Dandy, G.C.; Egelhardt, M.O. (2006). *"Multi-Objective Trade-Offs between cost and reliability in the replacement of water mains"*, Journal of water resources planning and management, Vol. 132, No. 2, 79-88.

Das, V.; Lin, K.; Mannila, H.; Renganathan, G.; Smyth, V. (1998). *"Rule Discovery from Time Series"*, Knowledge Discovery and Data Mining, Pages 16-22.

Dawsey, W.J.; Minsker, B.S.; VanBlaricum, V.L. (2006). *"Bayesian Belief Networks to Integrate Monitoring Evidence of Water Distribution System Contamination"*, Journal of water resources planning and management, Vol. 132, No. 4, 234-241.

Díaz A., J.L.; Pérez G., R. (2002a), *"Estado del arte en la utilización de técnicas avanzadas para la búsqueda de información no trivial a partir de datos en los sistemas de abastecimiento de agua potable",* II SEREA, Universidade Federal da Paraiba, Brasil, Libro de Proceedings.

Díaz A., J.L.; López, J., P.A.; Pérez G., R. (2002b), *"Técnicas de minería de datos para la búsqueda de información. Aplicación a sistemas de abastecimiento de agua"*, Universidad Politécnica de Valencia, Libro de Proceedings (ISBN 84-89487-07-3), 305-320.

Díaz A., J.L.; Pérez G., R.; López, P., G. (2002c), *"Utilización de técnicas de minería de datos para la predicción de la demanda. Comparación con técnicas tradicionales",* Universidad Politécnica de Valencia, Libro de Proceedings (ISBN 84-89487-07-3), 321-336.

Díaz A., J.L.; Pérez G., R; Martínez S.; F.J.; Fuertes, V.S. (2003a), *"Pérdidas de agua y rendimientos en abastecimientos"*, Ingeniería hidráulica en los abastecimientos de agua, Grupo Multidisciplinar de Modelación de Fluidos, Capítulo 19, 749-778. ISBN 84-89487-08-1.

Díaz A., J.L.; Pérez G., R; López, J., P.A; Martínez S.; F.J.; (2003b), *"Planificación y Rehabilitación de Redes de Abastecimiento"*, Ingeniería hidráulica en los abastecimientos de agua, Grupo Multidisciplinar de Modelación de Fluidos, Capítulo 20, 779-820. ISBN 84-89487-08-1

Díaz A., J.L.; Izquierdo S., J.; López J., P.A; Pérez G., R. (2004a). *"Introducción a la Minería de Datos"*, Métodos de Análisis Inteligente de Datos. Aplicaciones Hidráulicas y Ambientales. GMMF. 1-70. ISBN 84-89487-16-2.

Díaz A., J.L.; López J., P.A; Pérez G., R.; Martínez S., F.J. (2004b). *"Métodos de Aprendizaje Automático"*, Métodos de Análisis Inteligente de Datos. Aplicaciones Hidráulicas y Ambientales. GMMF. 199-212. ISBN 84-89487-16-2.

Díaz A., J.L.; Pérez G., R.; Izquierdo S., J.; López J., P.A. (2004c), *"La Inteligencia Artificial como Apoyo a los Abastecimientos de Agua"*, Universidad Politécnica de Valencia, Libro de Proceedings (ISBN 8489487154).

Díaz A., J.L.; Pérez G., R.; Izquierdo S., J.; López J., P.A. (2004d), *"Utilización del aprendizaje automático como ayuda a la gestión de sistemas de abastecimiento de agua",* Universidad Politécnica de Valencia, Libro de Proceedings (ISBN 8489487154).

Díaz A., J.L.; Pérez G., R.; Nudelman, M.A.; Izquierdo S., J.; (2005). *"Minería de Datos (Data Mining) en los abastecimientos de agua, Casos Hipotéticos de utilización",* Universidad Politécnica de Valencia, Libro de Proceedings (ISBN 84-89487-20-0).

Díaz A., J.L.; Pérez G., R.; Izquierdo S., J.; Alonso G., C.D. (2007). *"La minería de datos como soporte en la gestión de la rehabilitación de sistemas de abastecimiento de agua"*, UPV. ISBN 84-89487-25-1.

Díaz A., J.L.; Herrera, M.; Izquierdo S., J.; Montalvo, I.; Pérez G., R.; (2008). *"A Particle Swarm Optimization derivate applied to cluster analysis"*, M. Sànchez-Marrè, J. Béjar, J. Comas, A. Rizzoli and G. Guariso Ed.4th Biennial meeting of iEMSs. http://www.iemss.org/iemss2008/index.php?n=Main.Proceedings, 3, 1782-1790.

Díaz J.L.; Herrera M.; Izquierdo J.; Pérez-G., R. (2010). *"The tasks of pre and post-processing in Data Mining applied to a real world problem"*, to appear in iEMSS 2010: International Congress on Environmental Modelling and Software, July 5-8 2010, Ottawa, Ontario, Canada.

Duch, W.; Jankowski, N. (2000). *"Taxonomy of neural transfer functions"*, Neural Networks, IJCNN 2000, Proceedings of the IEEE-INNS-ENNS International Joint Conference on , Volume 3, Pages 477-482.

Düntsch, I.; Gediga, G.; (1997a). *"The rough set engine Grobian"*, Proceedings of

the 15th IMACS World Congress, Berlin, August 1997, 613-618.

Düntsch, I.; Gediga, G.; (1997b). *"Statistical evaluation of rough set dependency analysis"*, International Journal of Human-Computer Studies, Volume 46, Number 5, Pages 589-604.

Eberhart R.C.; Simpson P.; Dobbins, R. (1995) *"Computational Intelligence PC Tools"*, Academic Press.

Eberhart R.C.; Shi, Y. (2000) *"Comparing Inertia Weights and Constriction Factors in Particle Swarm Optimization"*, Proc. CEC, San Diego, CA, 2000, 84-88.

El-Baroudy, I.; Simonovic, S.P. (2003). *"New Fuzzy Performance Indices for Reliability Analysis of Water Supply Systems"*, The University of Western Ontario, Department of Civil and Environmental Engineering, Water Resources Research Report, Report No. 045, August, 2003, tomado de: http://www.eng.uwo.ca/research/iclr/fids/publications/products/Fuzzy_Performance_Indices2.pdf, 82pp.

El-Baroudy, I.; Ahmad, S.; Simonovic, S.P. (2004). *"Fuzzy Reliability Analysis for the Evaluation of Water Resource Systems Performance"*, CD Proceedings, 59th Annual Conference – CWRA, paper C4 (2006), tomado de: http://www.eng.uwo.ca/research/iclr/fids/publications/nserc-spacial/CWRA_con.pdf, 10pp.

El-Baroudy, I.; Simonovic, S.P. (2006). *"Application of the fuzzy performance measures to the City of London water supply system"*, Canadian Journal of Civil Engineering, Vol. 33, 255-265.

Escolano R, F.; Cazorla Q, M.A.; Alfonso, G, M.I.; Colomina, P. O.; Lozano, O. M.A. (2003). *"Inteligencia Artificial, Modelos, Técnicas y Áreas de Aplicación"* International Thomson Editores Spain Paraninfo, S.A. 370pp.

Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P.; Uthurusamy, R. (Editors). (1996). *"Advances in Knowledge Discovery and Data Mining"*, AAAI Press / The MIT Press, 611pp.

Friedman, J.H. (1990). *"Multivariate Adaptive Regression Splines",* Tech Report 102, Stanford Linear Accelerator Centre and Department of Statistics, Stanford University, 158pp.

Fukuda, T.; Morimoto, Y.; Tokuyama, T.; Morishita, S.; (1996). *"Constructing efficient decision trees by using optimized numeric association rules"*, Proceedings of the 22nd VLDB conference Mumbai (Bombay), India, 11pp.

Gibert, K.; Izquierdo, J.; Holmes, G.; Athanasiadis, I.; Comas, J.; Sànchez-Marrè, M. (2008) *"On the role of pre and post-processing in environmental data mining"*, International Congress on Environmental Modelling and Software - 4th Biennial Meeting, pp. 1937-1958, Barcelona, Spain.

Girolami, M.; Cichocki, A.; Amari, SI. (1998). *"A common neural network model for unsupervised exploratory data analysis and independent component analysis"*, IEEE Transactions on Neural Networks, Volume 9, Number 6, Pages 1495-1501.

Gómez, V.; Casanovas, A. (2002). *"Fuzzy logic and meteorological variables: a case study of solar irradiance"*, Fuzzy sets and systems, Volume 126, Pages 121-128.

Grzymala-Busse, J.W.; Ziarko, W. (2000). *"Data mining and rough set theory"*, Communications of the Acm., Volume 43, Number 4, Pages 108-109.

Guo, J.; Ding, H. (2009). *"Application of data mining on water supply Management"*, 2009 International Joint Conference on Artificial Intelligence, 606-608.

Haimes, Y.Y.; Longstaff, T. (2002). *"The Role of Risk Analysis in the Protection of Critical Infrastructures against Terrorism"*, Risk Analysis, Vol. 23, No. 2, 439-444.

Hartigan, J.A.; (1975). *"Clustering algorithms"*. John Wiley & Sons, New York.

Hebb, D.O. (1949). *"The organization of behavior"*, New York: Wiley

Herrera, M.; Izquierdo, J.; Montalvo, I.; García-Armengol, J. & Roig, J.V. (2009). *"Identification of surgical practice patterns using evolutionary cluster analysis"*, Mathematical and Computer Modelling (accepted) 50, 705-712.

Herrera, M.; Pérez-García, R.; Izquierdo, J.; Montalvo, I. (2009b). *"Scrutinizing changes in the water demand behaviour"*, in: POSITIVE SYSTEMS. Lecture notes in Control and Information Sciences, Bru, R. And Romero, S. (Eds.), Springer, pp. 305,313.

Herrera, M.; Torgo, L.; Izquierdo, J.; Pérez-García, R. (2010), *"Predictive models for forecasting hourly urban water demand"*, Journal of Hydrology, accepted.

Hernández, I.; López, I. (1997). *"Lógica Fuzzy para principiantes, cuando la máquina se acerca al pensamiento humano"*, SUR A&C Omron Electronics, S.A., 61pp.

Hernández O., J.; Ramírez Q., M.J.; Ferri R., C. (2004) *"Introducción a la Minería de Datos"*, Pearson Educación, S.A. 680pp.

Hinton, G. E. and Sejnowski, T. J. (1983). *"Optimal Perceptual Inference"*. Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Washington DC, pp. 448-453.

Holland, J.H. (1975). *"Adaptation in Natural and Artificial Systems"*, University of Michigan Press, 1975.

Hruschka, E.R.; Ebecken, N.F.F. (2000). *"Using a Clustering genetic algorithm for rule extraction from artificial neural networks"*, Combinations of Evolutionary Computation and Neural Networks, 2000 IEEE Symposium on, 2000, 199-206.

Huang, J.J.; McBean, E.A. (2009). *"Data Mining to Identify Contaminant Event Locations in Water Distribution Systems"*, Journal of Water Resources Planning and Management, Vol. 135, No. 6, 466-474.

Hunt, E. B.; Marin, J.; Stone, P. J. (1966). *"Experiments in Induction"*, New York: Academic Press.

Ingeominas. (1999). *"Zonificación de Amenazas Geológicas para los municipios del Eje Cafetero afectados por el sismo del 25 de enero de 1999",* Instituto Geográfico Agustín Codazzi, Bogotá (Colombia).

Izquierdo S., J.; Díaz A., J.L.; Pérez G., R.; Jiménez L., P.A.; Mora, J de J. (2008a). *"Knowledge Discovery in Environmental Data"*, Integrated Water Management. Springer. 51-68. ISBN 978-1-4020-6551-4.

Izquierdo, J.; Montalvo, I.; Pérez, R. & Fuertes, V.S. (2008b). *"Design optimization of wastewater collection networks by PSO"*, Computer & Mathematics with Applications, 56(3), 777–784.

Izquierdo, J.; Montalvo, I.; Herrera, M. & Pérez, R. (2009a). *"A derivative of Particle Swarm Optimization with enriched diversity"*, submitted to Computational Optimization and Applications Mathematical and computer modelling.

Izquierdo, J.; Montalvo, I.; Pérez, R. & Herrera, M. (2009b). *"Robust Design of Water Supply Systems through Evolutionary Optimization". POSTA09.* Third Multidisciplinary Symposium on Positive Systems: Theory and Applications, accepted 321-330, Valencia, Spain. In. POSITIVE SYSTEMS. Lecture notes in Control and Information and Sciences. R. Bru, S. Romero, (Eds.), Springer, 321-330.

Izquierdo, J., Montalvo, I., Pérez-García, R., Alonso, C.D. (2009c). *"Swarm Intelligence for Optimization in the Urban Water Industry"*, in: Modelling, Simulation and Optimization, ISBN 978-953-7619-36-7, Ed. IN-TECH, Vienna, in print.

Jain, L.C.; Martin, N.M. (editors) (1999). *"Fusion of Neural Networks, Fuzzy Sets and Genetic Algorithms, Industrial Applications"*, CRS Press, 354pp.

Jin, Y.X.; Cheng, H.Z.; Yan, J.I.; Zhang, L. (2007). *"New discrete method for particle swarm optimization and its application in transmission network expansion planning",* Electric Power systems Research, Vol.77, No.3-4, 227-233.

Karpenko, M.; Sepehri, N. (2002). *"Neural network classifiers applied to condition*

monitoring of a pneumatic process valve actuator", Engineering Application of Artificial Intelligence, Vol. 15, 273-283.

Kennedy, J.; Eberhart, R.C. (1995). *"Particle Swarm Optimization"*, in: Proceedings of IEEE International Conference on Neural Networks, Vol. 1-6, 1942-1948.

Klösgen, W.; Zytkow, J.M. (2002). *"Hand Book of Data Mining and Knowledge discovery"*, Oxford University Press, 1026 pp.

Kok, W.W.; Gedeon, T.D.; Chun, C.F.; Wong, P.M. (2001). *"Fuzzy rules extraction using self-organising neural network and association rules"*, Electrical and Electronic Technology, TENCON., Proceedings of IEEE Region 10 International Conference on, Volume 1, Pages 403-408.

Kompare, B. (1995). *"The Use of Artificial Intelligence in Ecological Modelling"*, Ph.D. Thesis, University of Copenhaguen, Denmark.

Koza, J.R. (1992). *"Genetic Programming: On the Programming of Computers by Natural Selection"*, MIT Press, Cambridge, MA. 819 pp.

Koza, J.R. (1994). *"Genetic Programming II: Automatic Discovery of Reusable Programs"*. MIT Press, 1992.

Langley, P.; Simon, H.A. (1995). *"Applications of Machine Learning an Rule Induction"*, Comm. ACM 38 (11), 55-64.

Larock, B.E.; Jepsson, R.W.; Watters, G.Z. (2000). *"Hydraulics of Pipeline Systems"*, CRC Press, 537 pp.

Le Gat, I.; Eisenbeis, P. (2000). *"Using Maintenance records to forecast failures in water networks"*, Urban Water, Vol.2, 173-181.

León, C.; Martín, S.; Elena, J.M.; Luque, J. (2000). *"EXPLORE - Hybrid expert system for water networks management"*, Journal of water resources planning and management, March-April, 65-74.

Lertpalangsunti, N.; Chan, C.W.; Mason, R.; Tontiwachwuthikul, P. (1999). *"A toolset for construction of hybrid intelligent forecasting systems: application for water demand prediction"*, Artificial Intelligence in Engineering, Vol. 13, 21-42.

Liang, X.; Liang, Y. (2001). *"Applications of data mining in hydrology"*, Data Mining, ICDM 2001, Proceedings IEEE International Conference on, Pages 617-620.

Maulik, U.; Bandyopadhyay, S. (2002). *"Performance Evaluation of Some Clustering Algorithms and Validity Indices"*, IEEE Transactions on pattern analysis and machine intelligence, Vol.24, 12, 1650-1654.

McCulloch, W. S. and Pitts, W. H. (1943). *"A logical calculus of the ideas*

immanent in nervous activity". Bulletin of Mathematical Biophysics, 5:115-133.

Menéndez, C.; Ordieres, J.B.; Ortega, F. (1994). *"Comparación de diferentes topologías de redes neuronales para clasificación no lineal"*, I Congreso Internacional de Ingeniería de Proyectos, Asturias (España), 12pp.

Montalvo, I.; Izquierdo, J.; Pérez, R. & Tung, M.M. (2008a). *"Particle Swarm Optimization applied to the design of water supply systems"*, Computer & Mathematics with Applications, 56(3), 769–776.

Montalvo, I.; Izquierdo, J.; Pérez, R. & Iglesias, P.L. (2008b). *"A diversity-enriched variant of discrete PSO applied to the design of Water Distribution Networks"*, Engineering Optimization, 40(7), 655–668.

Montalvo, I.; Izquierdo, J.; Pérez, R. & Herrera, M. (2010), *"Improved performance of PSO with self-adaptive parameters for computing the optimal design of Water Supply Systems"*, submitted to Computer-Aided Design Engineering Applications of Artificial Intelligence, 10.1016/j.engappai.2010.01.015.

Montesinos, P.; García-Gúzman, A.; Ayuso, J.L. (1999). *"Water distribution network optimization using a modified genetic algorithm"*, Water Resources Research, Vol. 35, No. 11, 3467-3473.

Murphy, L.J.; Simpson, A.R. (1992). *"Genetic Algorithms in Pipe Network Optimisation"*, Research Report No. R93, Department of Civil Engineering. The University of Adelaide, June 1992, 67pp.

Nguyen, H.P.; Le, L.P.; Santiprabhob, P.; De Baets, B. (2001). *"Approach to generating rules for expert systems using rough set theory"*, IFSA World Congress and 20th NAFIPS International Conference, Joint 9th , Volume 2, Pages 877-882.

Nikov, A.; Stoeva, S. (2001). *"Quick Fuzzy backpropagation algorithm"*, Neural Networks, Volume 14, Pages 231-244.

Ohrn, A. (1999). *"Discernibility and rough sets in medicine: Tools and applications"*, PhD Thesis, Norwegian University of Science and Technology, Department of computer and information science, Dec 1999, 223pp.

Olaru, C.; Geurts, P.; Wehenkel, L. (1999a). *"Data mining tools and applications in power system engineering",* Proceedings of PSCC99, the 13th power system computation conference, Throdheim, Norway, Volume 1, Pages 324-330.

Olaru, C.; Wehenkel, L.; (1999b). "*Data mining*", IEEE Computer Applications in Power, Volume 12, Number 3, Pages 19-25.

Parker, D. B., (1985). *"Learning Logic"*. Technical Report TR-47, Cambridge, MA: MIT Center for Research in computational Economics and Management

Science.

Paterson, A.; Niblett, T. (1982). *"ACLS manual"*, Version 1 (Technical Report). Glasgow, Scotland: Intelligent Terminals Limited.

Pawlak, Z.; Slowinski, R. (1994). *"Decision Analysis Using Rough Sets"*, International Transactions in Operational Research, Volume 1, Number 1, Pages 107-114.

Pawlak, Z. (1998). *"Granularity of knowledge, indiscernibility and rough sets"*, Fuzzy Systems Proceedings, IEEE World Congress on Computational Intelligence, The 1998 IEEE International Conference on, Volume 1, Pages 106-110.

Pawlak, Z. (1999). *"Rough set theory for intelligent industrial applications"*, Intelligent Processing and Manufacturing of Materials, IPMM '99.; Proceedings of the Second International Conference on, Volume 1, Pages 37-44.

Pawlak Z. (2001). *"Rough sets and their applications"*, Computational intelligence in theory and practice, Pages 73-91.

Pedrycz, W.; Gomide, F. (1998). *"An Introduction to Fuzzy Sets, Analysis and Design"*, The MIT Press. 465pp.

Quinlan, J. R. (1979). *"Discovering rules by induction from large collection of examples"*, en Expert Systems in the Microelectronic Age, Ed. D. Michie: 168-201. Edinburgh: Edinburgh University Press.

Quinlan, J. R. (1993). *"C4.5: Programs for machine learning"*, San Mateo, CA: Morgan Kaufmann.

Quinlan, J.R. (1997). *"Using C5.0: an informal tutorial"*. Available to download from WWW at: http://www.rulequest.com/see5-info.html

Rastogi, R.; Shim, K. (2000). *"PUBLIC: A decision tree classifier that integrates building and pruning"*, Data Mining & Knowledge Discovery, Volume 4, Number 4, Pages 315-344.

Referencia de nodos Clementine® 9.0 (2005). Integral Solutions Limited.

Reich, Y. (1997), *"Machine Learning techniques for civil engineering problems"*, Microcomputers in Civil Engineering, Vol. 12, No. 4, 295-310.

Reich, Y.; Barai, SV. (1999). *"Evaluating machine learning models for engineering problems"*, Artificial Intelligence in Engineering, Vol. 13, No. 3, 257-272.

Reis, L.F.R.; Porto, R.M.; Chaudhry, F.H. (1997). *"Optimal location of control valves in pipe networks by genetic algorithm"*, Journal of Water Resources

Planning & Management-ASCE., Vol. 123, No. 6, 317-326.

Resolución No. 1096 de 17 de noviembre de 2000. *"Reglamento Técnico para el sector de Agua Potable y Saneamiento Básico – RAS."*, Ministerio de Desarrollo Económico, República de Colombia, 105pp.

Revelli, R.; Ridolf, L. (2002). *"Fuzzy approach for analysis of pipe networks"*, Journal of hydraulic engineering, January 2002, 93-101.

Romero, R.F.; Kacprzyk, J.; Gomide, F. (1999). *"A biologically inspired neural network for dynamic system optimization"*, Neural Information Processing, Proceedings ICONIP '99., 6th International Conference on , Volume 3, Pages 1045-1050.

Rosenblatt, F. (1958). *"The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain"*. In, Psychological Review, Vol. 65, No. 6, pp. 386-408, November, 1958. Lancaster, PA and Washington, DC: American Psychological Association, 1958.

Rumelhart, D. E. and McClelland, J. L. eds. (1986). *"Parallel Distributed Processing: Explorations in the Microstructure of Cognition"*, 1, 318-362, MIT Press.

Santos, R.T.; Nievola, J.C.; Freitas, A.A. (2000). *"Extracting comprehensible rules from neural networks via genetic algorithms"*, Combinations of Evolutionary Computation and Neural Networks, 2000 IEEE Symposium on, Pages 130-139.

Savic, D.A., Walters, G.A, (1999). *"Hydroinformatics, Data Mining and Maintenance of UK water Networks"*, Journal of Anti-corrosion methods and materials, Vol. 46, No. 6, 415-425.

Sforna M. (2000), *"Data mining in a power company customer database"*, Electric Power Systems Research. Vol. 55, No. 3, 201-209.

Shan, N.; Ziarko, W. (1995). *"Data-Based Acquisition and Incremental Modification of Classification Rules"*, Computational Intelligence:; an international Journal, Volume 11, Number 2, Pages 357-370.

Shi, Y.; Eberhart, R.C. (1999). *"Empirical study of particle swarm optimization"*, In Proceedings of the IEEE International Congress on Evolutionary Computation, Washington, D.C., 1945-1950.

Schwefel, H.P. (1995). *"Evolution and Optimum Seeding"*, Wiley Inc. 1995.

Sheikholeslami, G.; Chatterjee, S.; Zhang A.D. (2000). *"WaveCluster: a wavelet-based clustering approach for spatial data in very large databases"*, Vldb. Journal, Volume 8, Number 3-4, Pages 289-304.

Stout, L.N. (2001). *"Finiteness notions in fuzzy sets"*, Fuzzy sets and systems,

Volume 124, Pages 25-33.

Tachibana, Y.; Ohnari, M. (1999). *"Prediction Model of Hourly Water consumption in water purification plant through categorical approach"*, Systems, Man, and Cybernetics, 1999. IEEE SMC '99 Conference Proceedings. 1999 IEEE International Conference on, 569-574.

Van Zyl, J.E.; Savic, D.A.; Walters, G.A. (2004). *"Operational Optimization of Water Distribution Systems Using a Hybrid Genetic Algorithm"*, Journal of water resources planning and management, Vol. 130, No. 2, 160-170.

Vaughn, ML. (1996). *"Interpretation and knowledge discovery from the multilayer perceptron network - opening the black box"*, Neural Computing & Applications, Volume 4, Number 2, Pages 72-82.

Vaughn, M.L.; Ong, E.; Cavill, SJ. (1997). *"Interpretation and knowledge discovery for a multilayer perceptron network that performs whole life assurance risk assessment"*, Neural Comput. & Applic., Number 6, Pages 201-213.

Vaughn, M.L. (1999). *"Derivation of the multilayer perceptron weight constraints for direct network interpretation and knowledge discovery"*, Neural Networks, Number 12, Pages 1259-1271.

Vaughn, M.L, Cavill, S.J, Taylor, S.J, Foy, M.A, Fogg, A.J.B. (2001). *"Direct explanations for the development and use of a multi-layer perceptron network that classifies low-back-pain patients"*, International Journal of Neural Systems, Volume 11, Number 4, Pages 335-347.

Vitkovsky, J.P.; Simpson, A.R. (1997). *"Calibration and Leak Detection in Pipe Networks Using Inverse Transient Analysis and Genetic Algorithms"*, Department of Civil and Environmental Engineering University of Adelaide Research report No. R157, 96 pp.

Werbos, P. J. (1974). *"Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences"*. PhD thesis, Harvard University.

Williams, C. (1983). *"A brief introduction to artificial intelligence"*, Oceans Conference Record (IEEE) 1983, 94-99.

Wu, Z.Y.; Simpson, A.R. (1996). *"Messy Genetic Algorithms for optimization of water distribution systems"*, Department of Civil and Environmental Engineering, University of Adelaide, Research Report No. R140, ISBN 0-86396-404-4 November 1996, 61pp.

Zadeh, L.A. (1965). *"Fuzzy Sets"*, Information Control 8: 338-353, 1965.

Zhang, X.; Wang, X.; Yin, Z.; Li, H. (2007). *"Research on Water Supply Reservoir Operating Rules Extraction Based on Artificial immune Recognition System"*, Intelligent Systems and Knowledge Engineering ISKE-2007 Proceedings, 8pp.

Zhong, N.; Dong,J.; Liu,C.; Ohsuga,S.; (2001). *"A hybrid model for rule discovery in data"*, Knowledge based systems, Volume 14, Pages 397-412.

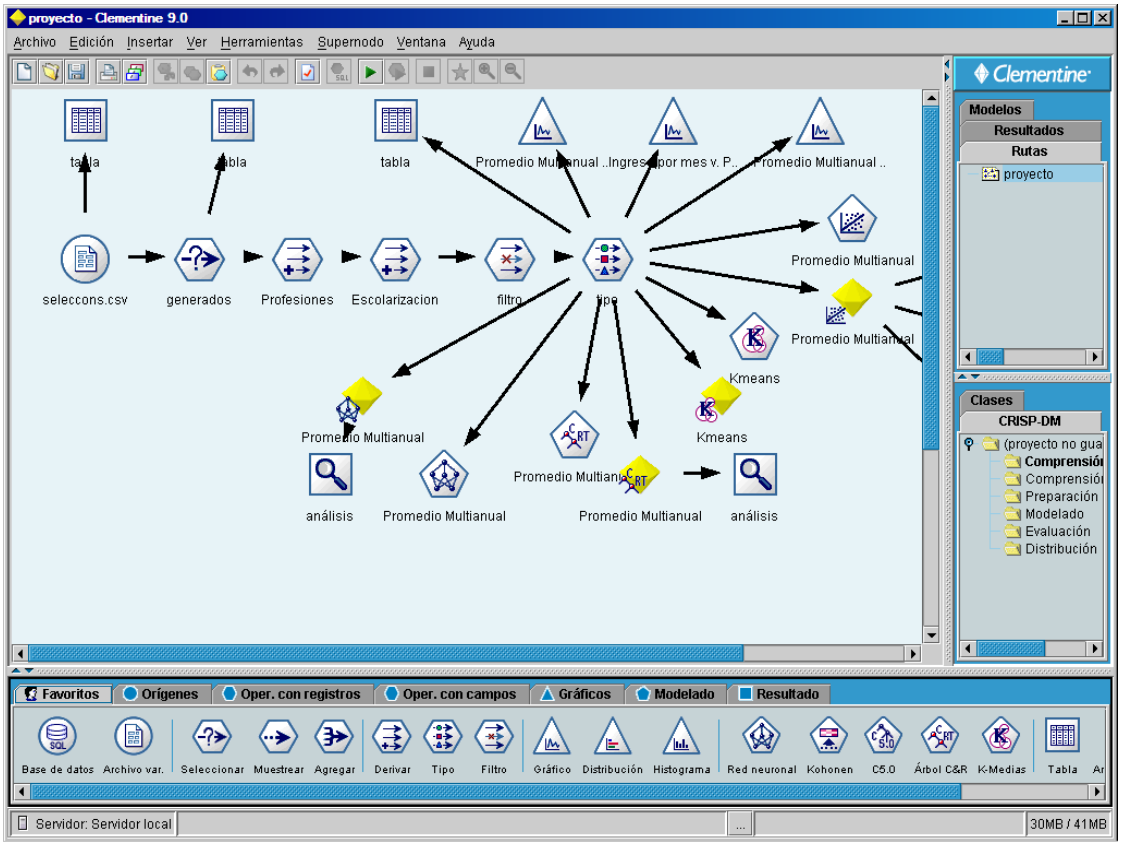
Zhou, S.M.; Xu, L.D. (2001). *"A new type of recurrent fuzzy neural network for modelling dynamic systems"*, Knowledge based systems, Volume 14, Pages 243-251.

Anexos

ANEXO 1

SPSS Clementine 9.0

La herramienta Clementine representa un conjunto de programas de minería de datos de SPSS, cuyas características de forma sucinta más relevantes son:



- Fuentes de datos (ASCII, XLS, SPSS, SAS U ODBC)
- Interfaz visual basado en rutas
- Herramientas de minería de datos: clasificación [árboles de decisión (C&R, CHAID, QUEST, C5.0), Redes Neuronales], Modelos Estadísticos [Regresión lineal, regresión logística, PCA/Factorial], Modelos conglomerados [Kohonen, K-medias, Two-Step /Bietáptico], Reglas de Asociación [GRI, A priori, CARMA, secuenciales], Modelos de extracción de texto.

- Manipulación de datos, operaciones con registros y con campos
- Nodos gráficos
- Exportación (SPSS, SAS, EXCEL)
- Generación de informes
- Diseño basado en el estándar CRISP-DM

El entorno de Clementine esta basado en *nodos* que se disponen y conectan para formar un flujo (*stream*) almacenado en rutas, ya sea por ficheros separados (.str) o por proyectos (.cpj). Estos archivos son independientes de las fuentes de datos, por tanto se pueden abrir, modificar, reejecutar o reorganizar sin afectar las bases de datos originales.

Clementine 9.0 cuenta con cuatro algoritmos de generación de árboles para realizar análisis de segmentación y clasificación. Estos algoritmos son básicamente similares: examinan todos los campos de la base de datos para detectar los que proporcionan la mejor clasificación o pronóstico dividiendo los datos en subgrupos. Este proceso se aplica de forma recursiva, dividiendo los subgrupos en unidades cada vez más pequeñas hasta completar el árbol (según se defina en determinados criterios de parada). Los campos objetivo y de entrada que son utilizados para la generación del árbol pueden ser intervalos numéricos o categóricos, según el algoritmo que se utilice. Con objetivo de rango se generan árboles de regresión, y con objetivos categóricos se generan árboles de clasificación.

El algoritmo de clasificación y regresión *C&RT* utiliza la partición reiterada para dividir los registros de entrenamiento en segmentos minimizando las impurezas en cada paso, donde un nodo se considera "puro" si el 100% de los casos corresponden a una categoría específica del campo objetivo. Los campos objetivo y predictor pueden ser de rango o categóricos. Todas las divisiones son binarias (sólo se crean dos subgrupos).

El algoritmo *CHAID* genera árboles utilizando estadísticos de chi-cuadrado para identificar las divisiones óptimas. Con este algoritmo se pueden generar árboles no binarios, lo que significa que algunas divisiones tendrán más de dos ramas. Los campos objetivo y predictor pueden ser de rango o categóricos. El algoritmo *QUEST* proporciona un método de clasificación binario y está diseñado para conseguir la reducción del tiempo de procesamiento necesario del algoritmo *C&RT*. Los campos predictores pueden ser de rango numérico, aunque el campo objetivo deber ser categórico.

El algoritmo C5.0 genera un árbol de decisión o un conjunto de reglas. El modelo divide la muestra basándose en el campo que ofrece la máxima ganancia de información en cada nivel. El campo objetivo debe ser categórico.

En cuanto a los algoritmos de redes neuronales implementados en Clementine 9.0 no se aplican restricciones a los tipos de campo, es decir la red neuronal puede gestionar entradas y salidas numéricas, simbólicas o de marcas. Incluyen análisis de sensibilidad para facilitar la interpretación de la red, la poda y la validación para evitar el sobreentrenamiento, y las redes dinámicas para buscar automáticamente arquitecturas de red adecuadas. Clementine ofrece seis métodos de entrenamiento para generar modelos de red neuronal:

- Rápido: utiliza reglas de miniaturas y características de los datos para seleccionar una forma adecuada (topología) para la red.

- Dinámico: crea una topología inicial aunque, según avanza el entrenamiento, añade o elimina unidades ocultas y modifica esta topología.

- Múltiple: crea varias redes de diferentes topologías (el número depende de los datos de entrenamiento). Luego estas redes son entrenadas conforme a un procesamiento seudoparalelo. Al final del entrenamiento, se presenta como modelo final el modelo con el nivel inferior de error cuadrático medio.

- Poda: en este método se inicia con una red de gran tamaño y elimina (o poda) las unidades más débiles de las capas ocultas y de entrada según se va

completando el entrenamiento.

- RBFN: La *red de función de base radial (RBFN)* utiliza una técnica similar al conglomerado de *K-medias* para crear una partición de los datos basándose en valores del campo objetivo.

- Poda exhaustiva: este método está relacionado con el método de poda. Se inicia con una red de gran tamaño y poda las unidades más débiles de las capas ocultas y de entrada según se va completando el entrenamiento. Los parámetros de entrenamiento son seleccionados por el método para garantizar una búsqueda exhaustiva de los posibles modelos para seleccionar el más adecuado.

Las redes de Kohonen corresponden a un tipo de red neuronal que realiza conglomerados (conocidos como *knet)* o mapas autoorganizativos. Las unidades básicas son las neuronas que se organizan en dios capas: la capa de entrada y la capa de salida (mapa de resultados).

1. Algoritmos de redes neuronales

En *Clementine* las redes neuronales utilizadas son *"feedforward"* también conocidas como "*perceptrones*" multicapas. Las neuronas en tales redes (también llamadas unidades) se organizan en capas.

a. Escalamiento de los valores de rango

Los campos son reescalados en rangos entre 0 y 1, de la siguiente forma:

$$x_i' = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}$$

donde x_i' es el valor reescalado del campo x para el registro i, x_i es el valor original de x para el registro i, $x_{\min}$ es el valor mínimo de x de todos los registros, y $x_{\max}$ es el valor máximo de x de todos los registros.

b. Codificación numérica de campos simbólicos

Se codifican los campos simbólicos como numéricos (0,1), asociando un campo numérico para cada categoría o valor del campo original.

1.1. Perceptron Multicapa

El entrenamiento del perceptron multicapa utiliza el método de retropropagación del error (*backpropagation of error*), basado en la regla *delta generalizada* (Rumelhart *et al.*, 1986.)[5]

a. Cálculos de propagación hacia adelante (feedforward)

La activación de cada neurona en una capa oculta o de salida se calcula de acuerdo con la siguiente expresión:

$$a_i = \sigma\left(\sum_j w_{ij} o_j\right)$$

donde a_i corresponde a la activación de la neurona i, j es el conjunto de neuronas en la capa precedente, w_{ij} es el peso de las conexiones entre la neurona i y la neurona j, o_j es la salida de la neurona j, y $\sigma(x)$ es la función de trasferencia;

$$\sigma(x) = \frac{1}{1+e^x}$$

b. Cálculos de retropropagación (backpropagation)

Para iniciar el entrenamiento, todos los pesos en la red se establecen como un conjunto de valores aleatorios en el intervalo $0.5 \le w_{ij} \le 0.5$.

El cambio de valor Δw para calcular los pesos se calcula como:

---

[5] Rumelhart, D.E.; McClelland, J.L. The PDP Research Group. 1986. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition.* Vol. 1: Foundations. Cambridge, MA: MIT Press.

$$\Delta w_{ij}\left(n+1\right)=\eta\delta_{pj}o_{pi}+\alpha\Delta w_{ij}\left(n\right)$$

donde η corresponde a un parámetro o índice de entrenamiento, δ_{pj} es el error propagado descrito anteriormente, o_{pi} es la neurona de salida i para el registro p, α es el parámetro de momentum, y $\Delta w_{ji}\left(n\right)$ es el valor de cambio para w_{ij} en el ciclo previo.

El valor de α se fija durante el entrenamiento, pero el valor de η varía entre ciclos de entrenamiento. η (*eta*) inicia con un valor especificado por el usuario, que va decreciendo de manera logarítmica hasta un valor bajo, revirtiendo a un valor alto y decreciendo de nuevo a un *eta* bajo. El valor de η es calculado como:

$$\eta\left(t\right)=\eta\left(t-1\right)\cdot\exp\left(\log\left(\frac{\eta_{bajo}}{\eta_{alto}}\right)\Big/d\right)$$

donde d es el número de ciclos de decaimiento de *eta* especificado por el usuario. Si $\eta\left(t-1\right)<\eta_{bajo}$, entonces $\eta\left(t\right)$ se dispone como $\eta_{alto}\cdot\eta$ continuando el ciclo así hasta que el entrenamiento sea completado.

El valor del error *retropropagado* δ_{pj} se calcula de acuerdo con el lugar donde la conexión se encuentra en la red. Para conexiones de neuronas de salida se calcula como:

$$\delta_{pj}=\left(t_{pj}-o_{pj}\right)o_{pj}\left(1-o_{pj}\right)$$

donde t_{pj} es el valor objetivo de la salida j para el registro p.

Para pesos que no conectan las neuronas de salida, los cálculos de δ_{pj} toman en cuenta la propagación hacia atrás del error:

$$\delta_{pj}=o_{pj}\left(1-o_{pj}\right)\sum_{k}\delta_{pk}w_{kj}$$

donde k es el conjunto de neuronas para la cual la neurona de salida j es conectada, w_{kj} es el peso entre la neurona actual y la neurona k, y δ_{pk} es el error propagado para el peso del registro actual de entrada.

Los pesos son actualizados inmediatamente en cuanto cada registro es presentado a la red durante el entrenamiento.

1.2. Redes de función de base radial

Compuestas por tres capas: entrada, oculta (también denominada capa receptor) y salida.

a. Estimación de los centros de la función de base radial

Son entrenados haciendo uso del algoritmo de *k-medias*, en cada iteración del algoritmo, cada registro es asignado al clúster cuyo centro es el más cercano. La cercanía es medida por la distancia Euclídea.

$$d_{ij} = \left\| X_i - C_j \right\|^2 = \sum_{q=1}^{Q} \left(x_{qi} - c_{qi} \right)^2$$

donde X_i es el vector de campos de entrada codificados para el registro i, C_j es el vector del centro del clúster para el clúster j, Q es el número de campos de entrada codificados, x_{qi} es el valor del $q_{iésimo}$ campo de entrada codificado para el $i_{ésimo}$ registro, y c_{qj} es el valor del $q_{iésimo}$ campo de entrada codificado para el $j_{iésimo}$ registro.

Para cada registro, la distancia entre el registro y cada centro de clúster es calculada, y el centro de clúster cuya distancia del registro es menor se asigna como nuevo registro del clúster. Cuando todos los registros han sido asignados, los centros de clúster se actualizan.

Luego de que los registros han sido reasignados a sus *clusters* más cercanos, se actualizan los centros de clúster. Este centro de clúster, se calcula

como el vector medio de los registros asignados al clúster:

$$C_j = \overline{X}_j$$

donde los componentes del vector medio $\overline{X}_j$ son calculados de la forma:

$$\overline{x}_{qj} = \frac{\sum_{i=1}^{n_j} x_{qi}(j)}{n_j}$$

donde n_j es el número de registros en el clúster j, $x_{qi}(j)$ es el $q_{iésimo}$ valor de campo codificado para el registro i el cual es asignado al clúster j.

b. Asignación de los anchos de la función base

Cada neurona receptora tiene una función de base radial asociada a ella. La función base utilizada en Clementine es una Gaussiana multidimensional

$$\exp\left(-\frac{d_i^2}{2\sigma_i^2}\right)$$

donde d_i es la distancia del centro del clúster i y, σ_i corresponde a un parámetro de escala que describe el tamaño de la función de base radial al clúster/neurona i.

El parámetro de escala σ_i se calcula basándose en las distancias de los dos *clusters* más cercanos.

$$\sqrt{\frac{d_1 + d_2}{2}}$$

donde d_1 corresponde a la distancia entre el centro del clúster y el centro del otro clúster más cercano y d_2 corresponde a la distancia al próximo centro del clúster más cercano. Así, los *clusters* que son cercanos a otros *clusters* tendrán pequeños campos receptivos, mientras que aquellos que se encuentran lejos de otros *clusters* tendrán grandes campos receptivos.

c. Entrenamiento de los pesos de salida para una RBFN

La activación para la neurona receptora *j* es calculada como

$$a_j = \exp\left(-\frac{\|r - c\|^2}{2\sigma_j^2 h} \right)$$

donde *r* es el vector de registros de entrada y *c* es el vector de centro de clúster.

La(s) neurona(s) de salida está(n) completamente interconectadas con la receptora o neuronas ocultas. Las neuronas receptoras pasan sobre sus valores de activación los cuales son pesados y sumados por la(s) neurona(s) de salida.

$$o_k = \sum_j W_{jk} a_j$$

Los pesos de salida W_{jk}, son entrenados de forma similar al entrenamiento de una red de retropropagación de dos capas. Los pesos se inicializan con pequeños valores aleatorios en el rango $-0.001 \le w_{ij} \le 0.001$, y entonces se actualizan en cada ciclo *t* por la formula:

$$w_{jk}(t) = w_{jk}(t-1) + \Delta w_{jk}(t)$$

El valor de cambio $\Delta w_{jk}(t)$ es calculado como:

$$\Delta w_{jk}(t) = \eta (r_k - o_k) a_i + \alpha \Delta w_{jk}(t-1)$$

análoga a la utilizada en la formula del método de retropropagación, el valor de η se fija a través del entrenamiento.

2. Algoritmos C&RT

C&RT son las siglas en inglés de **Árboles de Clasificación y Regresión,** originalmente descritas en el libro del mismo nombre (Breiman *et al.*, 1984)[6]. El algoritmo parte los datos en dos subconjuntos tal que los registros dentro de cada subconjunto sean más homogéneos que el subconjunto previo.

Parámetros del modelo

Los siguientes pasos son los utilizados para construir un árbol C&RT (comenzando con el nodo raíz que contiene todos los registros):

Encontrar la mejor división predictora. Para cada campo predictor, se encuentra la posible mejor división de cada campo, como sigue:

- **Campos de rango (numéricos).** Se clasifican los valores de los registros de menor a mayor. Se elige cada punto como un punto de división, y se calcula la impureza estadística para los nodos hijos resultantes de la partición. Se selecciona el mejor punto de partición del campo como el que produce el mayor decrecimiento en impureza relativa para la impureza del nodo que se está dividiendo.

- **Campos Simbólicos (Categóricos).** Se examina cada posible combinación de valores como dos subconjuntos. Para cada combinación, se calcula la impureza de los nodos filiales para la partición sobre esta combinación. Se selecciona el mejor punto de partición del campo como el que produce el mayor decrecimiento en impureza relativa para la impureza del nodo que se está dividiendo.

3. Algoritmos CHAID

Este algoritmo está fundamentado en el *Detector de Interacción Automático de Chi-Cuadrado*, que corresponde a una técnica estadística eficiente para

[6] Breiman, L.; Friedman, J.H.; Olshen, R.A.; Stone, C.J. 1984. *Classification and Regression Trees*. New York: Chapman & Hall/CRC.

segmentación o crecimiento de árboles desarrollado por *Kass*[7], 1980. Haciendo uso de la significancia de un test estadístico como criterio, CHAID evalúa todos los valores de un posible campo predictor. El algoritmo combina valores que son juzgados como estadísticamente homogéneos (similares) con respecto a la variable objetivo y mantiene todos los otros valores que son heterogéneos (diferentes).

CHAID Exhaustivo

Corresponde a una modificación del algoritmo, desarrollada para manejar algunas de las debilidades del método (Biggs *et al.*, 1991)[8]. En particular, CHAID algunas veces puede no encontrar la partición óptima para una variable, entonces finaliza combinando categorías tan pronto como encuentra que las categorías restantes son estadísticamente diferentes. El CHAID exhaustivo soluciona esto continuando la combinación de categorías de la variable estimadora hasta que se dejan solo dos súper categorías, luego se examinan estas combinaciones y se encuentra el conjunto de categorías que dan la asociación más fuerte con la variable objetivo, y se calcula un valor de *p* ajustado para la asociación. Así, el algoritmo encuentra la mejor partición para cada predictor y entonces se elige cual predictor dividir comparando los valores de *p* ajustados.

4. Algoritmos QUEST

QUEST (Quick, Unbiased, Efficient Statistical Tree) *árbol estadístico eficiente, rápido, imparcial*. Corresponde a un algoritmo binario de desarrollo de árboles relativamente novedoso (Loh y Shih, 1997)[9]. Este algoritmo trata con la selección por separado de un campo de división y un punto de división. Esta

---

[7] Kass, G. 1980. An exploratory technique for investigating large quantities of categorical data. Applied Statistics, 29:2, 119-127.

[8] Biggs, D.; de Ville, B.; Suen, E. 1991. *A method of choosing multiway partitions for classification and decision tress.* Journal of Applied Statistics, 18, 49-62.

[9] Loh, W. Y.; Shih, Y. S. 1997. Split selection methods for classification trees. *Statistica Sinica*, 7, 815-840.

división univariada en QUEST ejecuta una selección de campo aproximadamente imparcial. Es decir, si todos los campos predictores son igualmente informativos con respecto al campo objetivo, QUEST selecciona cualesquiera de los campos predictores con igual probabilidad.

5. Algoritmos de Kohonen

Los modelos de *Kohonen* (Kohonen, 2001) pertenecen a una clase especial de modelo de redes neuronales que ejecutan ***aprendizaje no supervisado***. Estos modelos toman los vectores de entrada y realizan un tipo de clústering organizado espacialmente o mapeo de características, para agrupar registros similares y colapsar el espacio de entrada a un espacio bidimensional que aproxime las relaciones de proximidad multidimensional entre *clusters*.

El modelo neuronal de Kohonen está compuesto de dos capas de neuronas o unidades: una capa de entrada y una de salida. La capa de entrada está totalmente conectada a la capa de salida y cada conexión tiene un peso asociado. Otra forma de ver la estructura de la red es pensando en que cada unidad capa de salida tiene un centro asociado, representado por un vector de entradas para el cual responde lo más fuertemente posible (donde cada elemento del vector central tiene un peso de la unidad de salida correspondiente a la unidad de entrada).

a. Parámetros del modelo

En un modelo de Kohonen, los parámetros se representan como ***pesos*** entre unidades de entrada y de salida, o igualmente, como un ***centro de clúster*** asociado con cada unidad de salida. Los registros de entrada son *presentados* a la red y los centros de clúster son actualizados de forma similar a las de un modelo de *k-medias*, con la diferencia de que los *clusters* son organizados en una cuadricula bidimensional y cada registro no afecta solo la unidad (clúster) a la cual es asignado sino también a unidades vecinas a la unidad ganadora.

El entrenamiento de las redes de Kohonen se realiza de la siguiente forma:

- La red se inicializa con pequeños pesos aleatorios.

- Los registros de entrada son presentados a la red en orden aleatorio. A la vez que cada registro es presentado, la unidad de salida con el centro más cercano al vector de entrada se identifica como la unidad de entrada.

- Los pesos de la unidad de entrada son ajustados para mover el centro del clúster más cercano al vector de entrada.

- Si el tamaño de la vecindad es mayor de cero, otras unidades de salida que están dentro de la vecindad de la unidad ganadora son también actualizadas así sus centros son cercanos al vector de entrada.

- Al final de cada ciclo, el parámetro tasa de aprendizaje η (*eta*) es actualizado.

- Este proceso se repite hasta que uno de los criterios de parada es colocado. El entrenamiento se realiza en dos fases, una fase de estructura gruesa y una fina. La primera fase en general tiene un tamaño de vecindad relativamente grande y un *eta* alto para aprender la estructura general de los datos, la segunda fase utiliza un tamaño de vecindad y *eta* pequeños para afinar los centros de clúster.

b. Distancias

Las distancias en una red de Kohonen son calculadas como distancias *Euclídeas* entre el vector de entrada codificado y el centro del clúster de la unidad de salida,

$$d_{i,j} = \sqrt{\sum_k \left(x_{ik} - w_{jk} \right)^2}$$

donde x_{ik} corresponde al valor del *k-ésimo* campo de entrada del *i-ésimo* registro, y w_{jk} corresponde al peso del *k-ésimo* campo de entrada de la *j-ésima* unidad de salida.

La activación de una unidad de salida corresponde simplemente a la distancia *Euclídea* entre el vector de pesos de la unidades de salida (su centro) y el vector de entrada. En las redes de Kohonen, la unidad de salida con la activación más baja es la unidad ganadora, esto en contraste con otro tipo de redes neuronales, donde las activaciones más altas representan una respuesta más fuerte.

c. Vecindades

La función de vecindad está basada en la distancia de *Chebychev*, la cual solo considera la distancia máxima en una única dimensión:

$$d_c\left(x,y\right)=\max_i\left|x_i-y_i\right|$$

donde x_i es la ubicación de la unidad *x* en la dimensión *i* de la cuadrícula de salida, y y_i de la otra unidad *y* sobre la misma dimensión.

Una unidad de salida o_j se considera que pertenece a la vecindad de otra unidad de salida o_i si $d_c\left(o_i,o_j\right)<n$, donde *n* corresponde al tamaño de la vecindad.

El tamaño de la vecindad permanece constante durante cada fase, pero diferentes tamaños de vecindades son usualmente utilizados en diferentes fases. Por defecto, *n*=2 para la fase 1 y *n*=2 para la fase 2.

d. Actualización de pesos

Para el nodo de salida ganador y sus vecinos si la vecindad es mayor de cero, los pesos se ajustan añadiendo una porción de la diferencia entre el vector de entrada y el vector actual de peso. La magnitud de este cambio se determina por medio del parámetro de la tasa de aprendizaje η (*eta*). El cambio de peso se calcula como:

$$\Delta W=\eta\cdot(W-I)$$

donde *W* es el vector peso para la unidad de salida que se está actualizando, *I* es el vector de entrada, y η es el parámetro de tasa de aprendizaje. En términos individuales de unidad,

$$\Delta w_j = \eta \cdot \left(w_j - i_j \right)$$

donde w_j es el peso correspondiente a la unidad de entrada *j*, e i_j corresponde a la *j-ésima* unidad de entrada.

e. Decaimiento de Eta

Al finalizar cada ciclo, el valor de η es actualizado, este valor generalmente decrece entre ciclos de entrenamiento. El usuario puede controlar la tasa de decrecimiento seleccionando decaimiento lineal o exponencial.

Decaimiento Lineal. Corresponde a la tasa de decaimiento por defecto. Cuando se selecciona esta opción, el valor de η decrece de forma lineal decayendo una cantidad fija en cada ciclo, de acuerdo a la formula:

$$\eta(t+1) = \eta(t) - \left(\frac{\eta(0) - \eta_{bajo}}{c} \right)$$

donde $\eta(0)$ corresponde al valor inicial de *eta* para la fase actual, y η_{bajo} es el valor bajo de *eta* para la fase de entrenamiento actual, calculado como el menor de los valores iniciales de *eta* entre la fase actual y la fase siguiente, y *c* es el número de ciclos establecidos para la fase actual.

Decaimiento exponencial. Cuando se selecciona esta opción, el valor de η decrece de forma exponencial, decayendo una proporción fija en cada ciclo, de acuerdo con la formula:

$$\eta(t+1) = \eta(t) \cdot \exp\left(\frac{\log\left(\frac{\eta_{bajo}}{\eta(0)}\right)}{c}\right)$$

donde el valor de η_{bajo} tiene un valor mínimo de 0.0001 para prevenir errores aritméticos en el algoritmo.

f. Modelo generado

Clúster

El clúster para un nuevo registro es derivado presentando el vector de entrada del registro a la red e identificando la neurona de salida con el vector peso más cercano, como se describió anteriormente. El valor estimado es devuelto como las coordenadas *x* e *y* de la neurona ganadora en la cuadrícula de salida.

ANEXO 2A

Modelo Generado

MATERIAL PQR in [""] [Moda: Fuga llave de paso] (47)

FUNCIONARIO MULTIPROPÓSITO (1) in [6 15 18 21 25] [Moda: Fuga llave de paso] => Fuga llave de paso (29; 0,724)

FUNCIONARIO MULTIPROPÓSITO (1) in [1 2 4 5 8 9 10 16 17 29] [Moda: Fuga tubo roto] (18)

FUNCIONARIO MULTIPROPÓSITO (1) in [5 8 10 17] [Moda: Fuga por universal] => Fuga por universal (5; 0,4)

FUNCIONARIO MULTIPROPÓSITO (1) in [1 2 4 9 16 29] [Moda: Fuga tubo roto] => Fuga tubo roto (13; 0,769)

MATERIAL PQR in ["HG" "PAD" "PEAD" "PF" "PVC"] [Moda: Fuga tubo roto] (176)

FUNCIONARIO MULTIPROPÓSITO (1) in [5 13 17 19 20 21 26 28] [Moda: Fuga, acometida en mal estado] (21)

MATERIAL PQR in ["HG"] [Moda: Fuga, acometida en mal estado] => Fuga, acometida en mal estado (8; 0,75)

MATERIAL PQR in ["PF" "PVC"] [Moda: Fuga adaptador] (13)

Longitud Tramo (m) <= 72.545 [Moda: Fuga, acometida en mal estado] (6)

FUNCIONARIO MULTIPROPÓSITO (1) in [5 20] [Moda: Fuga, acometida en mal estado] => Fuga, acometida en mal estado (3; 1,0)

FUNCIONARIO MULTIPROPÓSITO (1) in [17 28] [Moda: Fuga acometida] => Fuga acometida (3; 0,667)

Longitud Tramo (m) > 72.545 [Moda: Fuga adaptador] => Fuga adaptador (7; 1,0)

FUNCIONARIO MULTIPROPÓSITO (1) in [1 2 3 4 6 7 8 9 10 11 12 14 16 18 25 27 29] [Moda: Fuga tubo roto] (155)

MATERIAL PQR in ["PAD" "PF"] [Moda: Fuga adaptador] (52)

NIVEL RIESGO in ["Bajo"] [Moda: Fuga tubo roto] (43)

FUNCIONARIO MULTIPROPÓSITO (1) in [4 8 9 12 27] [Moda: Fuga adaptador] => Fuga adaptador (12; 0,583)

FUNCIONARIO MULTIPROPÓSITO (1) in [2 6 7 10 16 18 25 29] [Moda: Fuga tubo roto] (31)

Caudal medio (L/S) <= 0.186 [Moda: Fuga unión] => Fuga adaptador (11; 0,364)

Caudal medio (L/S) > 0.186 [Moda: Fuga tubo roto] => Fuga tubo roto (20; 0,7)

NIVEL RIESGO in ["" "Medio"] [Moda: Fuga adaptador] => Fuga adaptador (9; 0,889)

MATERIAL PQR in ["HG" "PEAD" "PVC"] [Moda: Fuga tubo roto] (103)

P(mm) Bella <= 42.850 [Moda: Fuga tubo roto] => Fuga tubo roto (96; 0,698)

P(mm) Bella > 42.850 [Moda: Fuga, acometida en mal estado] => Fuga, acometida en mal estado (7; 0,714)

Resumen de Configuración

Análisis
- Profundidad del árbol: 6

Campos
- Objetivo
  - TIPO DAÑO (1)
- Entradas
  - DIÁMETRO PQR(mm)
  - MATERIAL PQR
  - FUNCIONARIO MULTIPROPÓSITO (1)
  - NIVEL RIESGO
  - AMENAZAS
  - T (ºC) Bella
  - HR(%) Bella
  - BS(Hr) Bella
  - P(mm) Bella
  - P(mm) Jardin
  - Material Red
  - Diámetro Red (mm)
  - Longitud Tramo (m)
  - Rugosidad
  - Caudal medio (L/S)
  - Pérdida media (m/Km)
  - Presión PQR media (mca)

Configuración de creación
- Utilizar los datos en particiones: falso
- Utilizar frecuencia: falso
- Utilizar ponderación: falso
- Niveles por debajo del raíz: 20
- Modo: Experto
- Número máximo de sustitutos: 5
- Cambio mínimo en la impureza: 0,0
- Medida de impureza para objetivos categóricos: Gini
- Criterios de parada: Utilizar porcentaje
- Número mínimo de registros en rama parental (%): 2
- Número mínimo de registros en rama filial (%): 1
- Podar árbol: verdadero
- Utilizar regla de error típico: verdadero
- Multiplicador: 1,0
- Probabilidades previas: Basadas en datos de entrenamiento
- Corregir previas por costes de clasificación errónea: falso
- Utilizar costes de clasificación errónea: falso

ANEXO 2B

Modelo Generado

DIÁMETRO PQR in ["16" "16mm" "20mm" "31/2" "5/4"] [Moda: Fuga adaptador] (51)

FUNCIONARIO MULTIPROPÓSITO (2) in ["Alex, Carlos" "Carlos" "Juan C" "Juan C., Luis" "JuanC., Robinson" "Robinson"] [Moda: Fuga tubo roto] (27)

Coordenada Y <= 991064.250 [Moda: Fuga tubo roto] (19)

DESCRIPCIÓN in ["Filtración" "Fuga Agua" "Fuga medidor Goteo"] [Moda: Fuga unión] => Fuga unión (3; 1,0)

DESCRIPCIÓN in ["Filtración de agua en el anden" "Fuga agua" "Fuga de agua" "Fuga medidor" "Tobo roto" "Tubo roto" "Tubo roto, se informó a Juan Carlos García" "Tubo roto, se reventó manguera"] [Moda: Fuga tubo roto] => Fuga tubo roto (16; 0,75)

Coordenada Y > 991064.250 [Moda: Fuga adaptador] => Fuga adaptador (8; 0,75)

FUNCIONARIO MULTIPROPÓSITO (2) in ["Alex, Juan C." "Carlos, Henry" "Carlos, Luis" "Carlos, Miguel" "Carlos, Robinson" "Henry" "Henry, Juan C." "Juan C., Miguel" "Robinson, Alex" "Robinson, Miguel"] [Moda: Fuga adaptador] => Fuga adaptador (24; 0,625)

DIÁMETRO PQR in ["1" "1/2" "10" "11/2" "2" "3" "3/4" "6"] [Moda: Fuga tubo roto] (172)

Distancia del PQR al nodo de inicio <= 26.046 [Moda: Fuga tubo roto] (52)

DESCRIPCIÓN in ["Filtración de agua en la vía" "Fuga Medidor" "Fuga de agua" "Fuga medidor" "Fuga por 2 válvulas"] [Moda: Fuga llave de paso] => Fuga llave de paso (7; 0,571)

DESCRIPCIÓN in ["Fuga" "Fuga Agua" "Fuga acometida" "Fuga agua" "Fuga medidor, Sin agua" "Tubo Roto" "Tubo roto" "Tubo roto " "Tubo roto llevar llave de paso"] [Moda: Fuga tubo roto] => Fuga tubo roto (45; 0,8)

Distancia del PQR al nodo de inicio > 26.046 [Moda: Fuga tubo roto] (120)

Presión PQR (mca) 00:00 <= 56.485 [Moda: Fuga tubo roto] (68)

DESCRIPCIÓN in ["Fuga de agua" "Fuga de agua en lave de paso" "Fuga en medidor, ya se informó a Robinson Meza" "Fuga medidor" "Fuga medidor, sin agua (posible cambio de llave)" "Tubo roto, llevar llave de paso"] [Moda: Fuga llave de paso] => Fuga llave de paso (14; 0,786)

DESCRIPCIÓN in ["Daño en el contador" "Filtración" "Filtración, tubo roto" "Fuga Agua" "Fuga Medidor" "Fuga Medidor sin agua" "Fuga acometida" "Fuga agua" "Fuga agua medidor" "Fuga agua por robo medidor" "Fuga de agua en la vía" "Se reportó tubo roto" "Tubo Roto" "Tubo roto" "Tubo roto Contratistas" "Tubo roto, se le informó al disponible, RG3"] [Moda: Fuga tubo roto] (54)

FUNCIONARIO MULTIPROPÓSITO (2) in ["Carlos" "Carlos, Luis" "Henry" "Henry, Luis" "Juan C., Miguel" "Luis" "Luis, Robinson" "Robinson, Alex"] [Moda: Fuga, acometida en mal estado] (17)

FUNCIONARIO MULTIPROPÓSITO (2) in ["Carlos" "Carlos, Luis" "Henry"] [Moda: Fuga adaptador] => Fuga adaptador (6; 0,833)

FUNCIONARIO MULTIPROPÓSITO (2) in ["Henry, Luis" "Juan C., Miguel" "Luis" "Luis, Robinson" "Robinson, Alex"] [Moda: Fuga, acometida en mal estado] (11)

Presión Nodo Inicial (mca) 00:00 <= 47.050 [Moda: Fuga adaptador] => Fuga adaptador (3; 0,667)

Presión Nodo Inicial (mca) 00:00 > 47.050 [Moda: Fuga, acometida en mal estado] => Fuga, acometida en mal estado (8; 1,0)

FUNCIONARIO MULTIPROPÓSITO (2) in ["Alex" "Alex, Henry" "Alex, Juan C." "Alex, Miguel" "Carlos, Henry" "Henry, Juan C." "Juan C" "Juan C., Luis" "JuanC., Robinson" "Robinson" "Robinson, Miguel"] [Moda: Fuga tubo roto] (37)

Presión Nodo Final (mca) 6:00 <= 50.740 [Moda: Fuga tubo roto] (28)

Coordenada Y <= 991825.289 [Moda: Fuga adaptador] (13)

ID <= 298 [Moda: Fuga adaptador] => Fuga adaptador (5; 1,0)

ID > 298 [Moda: Fuga tubo roto] (8)

Presión Nodo Inicial (mca) 00:00 <= 43.405 [Moda: Fuga llave de paso] => Fuga llave de paso (5; 0,6)

Presión Nodo Inicial (mca) 00:00 > 43.405 [Moda: Fuga tubo roto] => Fuga tubo roto (3; 1,0)

Coordenada Y > 991825.289 [Moda: Fuga tubo roto] => Fuga tubo roto (15; 0,933)

Presión Nodo Final (mca) 6:00 > 50.740 [Moda: Fuga, acometida en mal estado] (9)

Presión PQR (mca) 13:00 <= 51.770 [Moda: Fuga, acometida en mal estado] => Fuga, acometida en mal estado (3; 1,0)

Presión PQR (mca) 13:00 > 51.770 [Moda: Daño interno] => Daño interno (6; 0,333)

Presión PQR (mca) 00:00 > 56.485 [Moda: Fuga tubo roto] (52)

DESCRIPCIÓN in ["Fuga" "Fuga acometida" "Fuga agua" "Fuga red principal" "Reporte de tubo roto, se informo a R-3" "Tubo Roto" "Tubo roto" "Tubo roto en andén"] [Moda: Fuga tubo roto] => Fuga tubo roto (40; 0,75)

DESCRIPCIÓN in ["Cambio de llave, solo mano de obra" "Daño en medidor, ya se informó a Robinson Mesa" "Fuga Válvula" "Fuga agua por universal" "Fuga de agua" "Fuga de agua en medidor" "Fuga medidor" "Tubo roto por hurto de medidor" "Tubo roto, se le informo al fontanero Robinson Mesa" "Tubo roto, ya se informó a Robinson Mesa"] [Moda: Fuga llave de paso] (12)

DESCRIPCIÓN in ["Cambio de llave, solo mano de obra" "Fuga Válvula" "Fuga agua por universal" "Fuga de agua" "Fuga de agua en medidor" "Fuga medidor" "Tubo roto por hurto de medidor"] [Moda: Fuga llave de paso] (9)

DESCRIPCIÓN in ["Cambio de llave, solo mano de obra" "Fuga Válvula" "Fuga medidor"] [Moda: Fuga llave de paso] => Fuga llave de paso (4; 0,75)

DESCRIPCIÓN in ["Fuga agua por universal" "Fuga de agua" "Fuga de agua en medidor" "Tubo roto por hurto de medidor"] [Moda: Fuga unión] => Fuga por universal (5; 0,4)

DESCRIPCIÓN in ["Daño en medidor, ya se informó a Robinson Mesa" "Tubo roto, se le informo al fontanero Robinson Mesa" "Tubo roto, ya se informó a Robinson Mesa"] [Moda: Fuga adaptador] => Fuga adaptador (3; 1,0)

Resumen de Configuración

Análisis
- Profundidad del árbol: 9

Campos
- Objetivo
  - TIPO DAÑO (1)
- Entradas
  - ID
  - Coordenada X
  - Coordenada Y
  - DESCRIPCIÓN
  - DIÁMETRO PQR
  - TIEMPO REPARACIÓN
  - FUNCIONARIO MULTIPROPÓSITO (2)
  - AMENAZAS
  - Coordenada Nudo inicio X
  - Coordenada Nudo inicio Y
  - Coordenada Nudo fin X
  - Coordenada Nudo fin Y
  - Distancia del PQR al nodo de inicio
  - Caudal (L/s) 00:00
  - Presión Nodo Inicial (mca) 00:00
  - Demanda Nodo Final (L/s) 00:00
  - Presión Nodo Final (mca) 00:00
  - Presión PQR (mca) 00:00
  - Caudal (L/s) 01:00
  - Presión Nodo Inicial (mca) 01:00
  - Demanda Nodo Final (L/s) 01:00
  - Presión Nodo Final (mca) 01:00
  - Presión PQR (mca) 1:00
  - Caudal (L/s) 02:00
  - Presión Nodo Inicial (mca) 02:00
  - Demanda Nodo Final (L/s) 02:00
  - Presión Nodo Final (mca) 02:00
  - Presión PQR (mca) 2:00
  - Caudal (L/s) 3:00
  - Presión Nodo Inicial (mca) 3:00
  - Presión Nodo Final (mca) 3:00
  - Presión PQR (mca) 3:00
  - Caudal (L/s) 4:00
  - Presión Nodo Inicial (mca) 4:00
  - Presión Nodo Final (mca) 4:00
  - Presión PQR (mca) 4:00
  - Presión Nodo Inicial (mca) 5:00
  - Presión Nodo Final (mca) 5:00
  - Presión PQR (mca) 5:00
  - Demanda Nodo Inicial (L/s) 6:00
  - Presión Nodo Final (mca) 6:00
  - Demanda Nodo Inicial (L/s) 7:00
  - Demanda Nodo Inicial (L/s) 8:00
  - Presión Nodo Final (mca) 8:00
  - Presión Nodo Final (mca) 9:00
  - Presión Nodo Final (mca) 10:00
  - Presión Nodo Final (mca) 12:00
  - Presión Nodo Final (mca) 13:00
  - Presión PQR (mca) 13:00
  - Presión PQR (mca) 14:00
  - Presión PQR (mca) 15:00
  - Presión PQR (mca) 16:00

Presión PQR (mca) 17:00
Presión PQR (mca) 18:00
Presión PQR (mca) 19:00
Presión PQR (mca) 20:00
Presión Nodo Inicial (mca) 23:00
Presión PQR (mca) 24:00

Configuración de creación

Utilizar los datos en particiones: falso
Utilizar frecuencia: falso
Utilizar ponderación: falso
Niveles por debajo del raíz: 20
Modo: Experto
Número máximo de sustitutos: 5
Cambio mínimo en la impureza: 0,0
Medida de impureza para objetivos categóricos: Gini
Criterios de parada: Utilizar porcentaje
Número mínimo de registros en rama parental (%): 2
Número mínimo de registros en rama filial (%): 1
Podar árbol: verdadero
Utilizar regla de error típico: verdadero
Multiplicador: 1,0
Probabilidades previas: Basadas en datos de entrenamiento
Corregir previas por costes de clasificación errónea: falso
Utilizar costes de clasificación errónea: falso

ANEXO 2C

Modelo Generado

$KX-Kohonen <= 5.500 [Moda: Fuga adaptador] (98)

$KX-Kohonen <= 1.500 [Moda: Fuga tubo roto] (39)

$KY-Kohonen <= 4 [Moda: Fuga tubo roto] (35)

$KY-Kohonen <= 0.500 [Moda: Fuga tubo roto] (26)

$KX-Kohonen <= 0.500 [Moda: Fuga tubo roto] => Fuga tubo roto (20; 0,55)

$KX-Kohonen > 0.500 [Moda: Fuga adaptador] => Fuga adaptador (6; 0,667)

$KY-Kohonen > 0.500 [Moda: Fuga tubo roto] => Fuga tubo roto (9; 1,0)

$KY-Kohonen > 4 [Moda: Fuga llave de paso] => Fuga llave de paso (4; 1,0)

$KX-Kohonen > 1.500 [Moda: Fuga adaptador] (59)

$KX-Kohonen <= 2.500 [Moda: Fuga adaptador] => Fuga adaptador (21; 0,905)

$KX-Kohonen > 2.500 [Moda: Fuga llave de paso] (38)

$KY-Kohonen <= 2.500 [Moda: Fuga llave de paso] => Fuga llave de paso (15; 0,867)

$KY-Kohonen > 2.500 [Moda: Fuga, acometida en mal estado] => Fuga, acometida en mal estado (23; 0,348)

$KX-Kohonen > 5.500 [Moda: Fuga tubo roto] (125)

$KX-Kohonen <= 7.500 [Moda: Fuga tubo roto] (115)

$KX-Kohonen <= 6.500 [Moda: Fuga tubo roto] (23)

$KY-Kohonen <= 2.500 [Moda: Fuga, acometida en mal estado] => Fuga, acometida en mal estado (12; 0,417)

$KY-Kohonen > 2.500 [Moda: Fuga tubo roto] => Fuga tubo roto (11; 0,909)

$KX-Kohonen > 6.500 [Moda: Fuga tubo roto] => Fuga tubo roto (92; 0,707)

$KX-Kohonen > 7.500 [Moda: Fuga adaptador] (10)

$KY-Kohonen <= 2.500 [Moda: Fuga adaptador] => Fuga adaptador (6; 0,5)

$KY-Kohonen > 2.500 [Moda: Fuga llave de paso] => Fuga llave de paso (4; 0,75)

Resumen de Configuración

Análisis
- Profundidad del árbol: 5

Campos
- Objetivo
  - TIPO DAÑO (1)
- Entradas
  - $KX-Kohonen
  - $KY-Kohonen

Configuración de creación
- Utilizar los datos en particiones: falso
- Utilizar frecuencia: falso
- Utilizar ponderación: falso
- Niveles por debajo del raíz: 20
- Modo: Experto
- Número máximo de sustitutos: 5
- Cambio mínimo en la impureza: 0,0
- Medida de impureza para objetivos categóricos: Gini
- Criterios de parada: Utilizar porcentaje
- Número mínimo de registros en rama parental (%): 2
- Número mínimo de registros en rama filial (%): 1
- Podar árbol: verdadero
- Utilizar regla de error típico: verdadero
- Multiplicador: 1,0
- Probabilidades previas: Basadas en datos de entrenamiento
- Corregir previas por costes de clasificación errónea: falso
- Utilizar costes de clasificación errónea: falso

ANEXO 2D

Modelo Generado

$KX-Kohonen <= 4.500 [Moda: Fuga pitorra] (430)

$KX-Kohonen <= 0.500 [Moda: Fuga tubo roto] (124)

$KY-Kohonen <= 3.500 [Moda: Fuga tubo roto] => Fuga tubo roto (69; 0,754)

$KY-Kohonen > 3.500 [Moda: Fuga llave de paso] => Fuga llave de paso (55; 0,491)

$KX-Kohonen > 0.500 [Moda: Fuga pitorra] (306)

$KY-Kohonen <= 2.500 [Moda: Fuga adaptador] (149)

$KX-Kohonen <= 2.500 [Moda: Fuga adaptador] (76)

$KY-Kohonen <= 1.500 [Moda: Fuga adaptador] => Fuga adaptador (45; 0,911)

$KY-Kohonen > 1.500 [Moda: Fuga tubo roto] => Fuga tubo roto (31; 0,548)

$KX-Kohonen > 2.500 [Moda: Fuga llave de paso] (73)

$KY-Kohonen <= 0.500 [Moda: Fuga pitorra] => Fuga pitorra (32; 0,344)

$KY-Kohonen > 0.500 [Moda: Fuga llave de paso] => Fuga llave de paso (41; 0,976)

$KY-Kohonen > 2.500 [Moda: Fuga pitorra] (157)

$KY-Kohonen <= 5.500 [Moda: Fuga pitorra] (98)

$KX-Kohonen <= 3.500 [Moda: Fuga pitorra] => Fuga pitorra (80; 0,9)

$KX-Kohonen > 3.500 [Moda: Fuga acometida] => Fuga acometida (18; 0,278)

$KY-Kohonen > 5.500 [Moda: Fuga, acometida en mal estado] (59)

$KX-Kohonen <= 2.500 [Moda: Fuga acometida] => Fuga acometida (28; 0,357)

$KX-Kohonen > 2.500 [Moda: Fuga, acometida en mal estado] => Fuga, acometida en mal estado (31; 0,581)

$KX-Kohonen > 4.500 [Moda: Fuga tubo roto] (416)

$KX-Kohonen <= 8.500 [Moda: Fuga tubo roto] (273)

$KY-Kohonen <= 4.500 [Moda: Fuga tubo roto] (204)

$KX-Kohonen <= 6.500 [Moda: Fuga tubo roto] (92)

$KY-Kohonen <= 3.500 [Moda: Fuga, acometida en mal estado] (67)

$KY-Kohonen <= 0.500 [Moda: Fuga tubo roto] => Fuga tubo roto (34; 0,324)

$KY-Kohonen > 0.500 [Moda: Fuga, acometida en mal estado] => Fuga, acometida en mal estado (33; 0,424)

$KY-Kohonen > 3.500 [Moda: Fuga tubo roto] => Fuga tubo roto (25; 0,68)

$KX-Kohonen > 6.500 [Moda: Fuga tubo roto] => Fuga tubo roto (112; 0,705)

$KY-Kohonen > 4.500 [Moda: Fuga tubo roto] => Fuga tubo roto (69; 0,797)

$KX-Kohonen > 8.500 [Moda: Fuga pitorra] (143)

$KY-Kohonen <= 5 [Moda: Fuga pitorra] (123)

$KY-Kohonen <= 3.500 [Moda: Fuga pitorra] => Fuga pitorra (89; 0,303)

$KY-Kohonen > 3.500 [Moda: Fuga llave de paso] => Fuga llave de paso (34; 0,353)

$KY-Kohonen > 5 [Moda: Fuga tubo roto] => Fuga tubo roto (20; 0,4)

Resumen de Configuración

Análisis
- Profundidad del árbol: 6

Campos
- Objetivo
  - TIPO DAÑO (1)
- Entradas
  - $KX-Kohonen
  - $KY-Kohonen

Configuración de creación
- Utilizar los datos en particiones: falso
- Utilizar frecuencia: falso
- Utilizar ponderación: falso
- Niveles por debajo del raíz: 20
- Modo: Experto
- Número máximo de sustitutos: 5
- Cambio mínimo en la impureza: 0,0
- Medida de impureza para objetivos categóricos: Gini
- Criterios de parada: Utilizar porcentaje
- Número mínimo de registros en rama parental (%): 2
- Número mínimo de registros en rama filial (%): 1
- Podar árbol: verdadero
- Utilizar regla de error típico: verdadero
- Multiplicador: 1,0
- Probabilidades previas: Basadas en datos de entrenamiento
- Corregir previas por costes de clasificación errónea: falso
- Utilizar costes de clasificación errónea: falso

ANEXO 2E

Modelo Generado

MATERIAL PQR in ["" " "] [Moda: Fuga tubo roto] (611)

SE CORRIGIÓ DAÑO in ["No"] [Moda: Daño interno] (62)

FUNCIONARIO MULTIPROPÓSITO (2) in ["Alex, Carlos" "Carlos" "Carlos, Robinson" "Henry" "Henry, Juan C." "Henry, Miguel"] [Moda: no hay daño] => no hay daño (22; 0,545)

FUNCIONARIO MULTIPROPÓSITO (2) in ["" "#N/A" "Alex, Juan C." "Carlos, Henry" "Carlos, Luis" "Juan C" "Juan C., Luis" "JuanC., Robinson" "Luis" "Robinson" "Robinson, Alex"] [Moda: Daño interno] (40)

Pérdida (m/Km) 0:00 <= 0.280 [Moda: Fuga tubo roto] => Fuga tubo roto (25; 0,32)

Pérdida (m/Km) 0:00 > 0.280 [Moda: Daño interno] => Daño interno (15; 0,667)

SE CORRIGIÓ DAÑO in ["Si"] [Moda: Fuga tubo roto] (549)

DIÁMETRO PQR(mm) <= 14.350 [Moda: Fuga pitorra] (457)

Coordenada Nudo fin X <= 1160156.940 [Moda: Fuga pitorra] (427)

Coordenada Nudo fin X <= 1158867 [Moda: Fuga llave de paso] (102)

FUNCIONARIO MULTIPROPÓSITO (2) in ["" "Alex, Henry" "Alex, Juan C." "Carlos" "Carlos, Henry" "Henry, Juan C." "Henry, Miguel" "Juan C., Luis" "JuanC., Robinson"] [Moda: Fuga llave de paso] (65)

Presión PQR (mca) 16:00 <= 52.633 [Moda: Fuga llave de paso] => Fuga llave de paso (55; 0,582)

Presión PQR (mca) 16:00 > 52.633 [Moda: Fuga tubo roto] => Fuga tubo roto (10; 0,6)

FUNCIONARIO MULTIPROPÓSITO (2) in ["Alex, Carlos" "Carlos, Luis" "Carlos, Robinson" "Henry" "Henry, Robinson" "Juan C" "Juan C., Carlos" "Juan C., Miguel" "Luis" "Robinson"] [Moda: Fuga pitorra] => Fuga pitorra (37; 0,297)

Coordenada Nudo fin X > 1158867 [Moda: Fuga pitorra] (325)

Material Red in ["AC" "AP" "HG"] [Moda: Fuga pitorra] (159)

FUNCIONARIO MULTIPROPÓSITO (2) in ["" "Alex, Carlos" "Alex, Henry" "Carlos, Henry" "Henry"] [Moda: Fuga tubo roto] => Fuga tubo roto (23; 0,609)

FUNCIONARIO MULTIPROPÓSITO (2) in ["#N/A" "Alex" "Alex, Juan C." "Alex, Miguel" "Carlos" "Henry, Juan C." "Henry, Luis" "Henry, Robinson" "Juan C" "Juan C., Luis" "Juan C., Miguel" "JuanC., Robinson" "Luis" "Luis, Edgar" "Luis, Robinson" "Miguel" "Robinson" "Robinson, Miguel"] [Moda: Fuga pitorra] (136)

FUNCIONARIO MULTIPROPÓSITO (2) in ["#N/A" "Alex" "Alex, Juan C." "Alex, Miguel" "Carlos" "Henry, Juan C." "Henry, Robinson" "Juan C" "Juan C., Luis" "JuanC., Robinson" "Luis" "Robinson"] [Moda: Fuga llave de paso] (126)

Caudal (L/s) 00:00 <= 0.590 [Moda: Fuga llave de paso] (81)

FUNCIONARIO MULTIPROPÓSITO (2) in ["Alex, Juan C." "Alex, Miguel" "Carlos"] [Moda: Fuga tubo roto] (35)

Longitud Tramo (m) <= 76.685 [Moda: Fuga pitorra] => Fuga

pitorra (20; 0,35)

Longitud Tramo (m) > 76.685 [Moda: Fuga tubo roto] => Fuga tubo roto (15; 0,467)

FUNCIONARIO MULTIPROPÓSITO (2) in ["#N/A" "Alex" "Henry, Juan C." "Henry, Robinson" "Juan C" "Juan C., Luis" "JuanC., Robinson" "Luis" "Robinson"] [Moda: Fuga llave de paso] => Fuga llave de paso (46; 0,543)

Caudal (L/s) 00:00 > 0.590 [Moda: Fuga pitorra] => Fuga pitorra (45; 0,4)

FUNCIONARIO MULTIPROPÓSITO (2) in ["Henry, Luis" "Juan C., Miguel" "Luis, Edgar" "Luis, Robinson" "Miguel" "Robinson, Miguel"] [Moda: Fuga pitorra] => Fuga pitorra (10; 0,9)

Material Red in ["" "PAD" "PVC"] [Moda: Fuga pitorra] (166)

Longitud Tramo (m) <= 163.325 [Moda: Fuga tubo roto] (155)

Presión Nodo Inicial (mca) 01:00 <= 41.800 [Moda: Fuga acometida] (26)

FUNCIONARIO MULTIPROPÓSITO (2) in ["Alex, Carlos" "Carlos, Henry" "Henry" "Henry, Luis" "Juan C" "Miguel"] [Moda: Fuga acometida] => Fuga acometida (11; 0,727)

FUNCIONARIO MULTIPROPÓSITO (2) in ["Alex, Henry" "Alex, Juan C." "Alex, Miguel" "Carlos" "Carlos, Luis" "Juan C., Carlos" "Juan C., Miguel" "JuanC., Robinson" "Luis" "Luis, Robinson" "Robinson"] [Moda: Fuga llave de paso] => Fuga llave de paso (15; 0,333)

Presión Nodo Inicial (mca) 01:00 > 41.800 [Moda: Fuga tubo roto] (129)

FUNCIONARIO MULTIPROPÓSITO (2) in ["" "#N/A" "Alex" "Alex, Henry" "Alex, Juan C." "Alex, Miguel" "Carlos" "Carlos, Henry" "Carlos, Miguel" "Carlos, Robinson" "Henry" "Henry, Juan C." "Henry, Robinson" "Juan C" "Juan C., Luis" "Juan C., Miguel" "Luis, Robinson" "Luis-Miguel" "Miguel"] [Moda: Fuga tubo roto] => Fuga tubo roto (116; 0,353)

FUNCIONARIO MULTIPROPÓSITO (2) in ["Alex, Carlos" "Carlos, Luis" "JuanC., Robinson" "Luis, Edgar" "Robinson" "Robinson, Alex"] [Moda: Fuga pitorra] => Fuga pitorra (13; 0,923)

Longitud Tramo (m) > 163.325 [Moda: Fuga pitorra] => Fuga pitorra (11; 0,818)

Coordenada Nudo fin X > 1160156.940 [Moda: Fuga, acometida en mal estado] (30)

FUNCIONARIO MULTIPROPÓSITO (2) in ["Carlos" "Carlos, Henry" "Henry, Juan C." "Henry, Miguel" "Henry, Robinson" "Juan C" "Juan C., Luis" "Luis" "Luis, Alex"] [Moda: Fuga tubo roto] => Fuga tubo roto (20; 0,5)

FUNCIONARIO MULTIPROPÓSITO (2) in ["Alex, Carlos" "Alex, Juan C." "Carlos, Luis" "Henry" "Juan C., Miguel" "JuanC., Robinson" "Robinson"] [Moda: Fuga, acometida en mal estado] => Fuga, acometida en mal estado (10; 0,6)

DIÁMETRO PQR(mm) > 14.350 [Moda: Fuga tubo roto] => Fuga tubo roto (92; 0,391)

MATERIAL PQR in ["HF" "HG" "PAD" "PEAD" "PF" "PVC"] [Moda: Fuga tubo roto] (235)

MATERIAL PQR in ["PAD" "PF"] [Moda: Fuga adaptador] => Fuga adaptador (87; 0,414)

MATERIAL PQR in ["HF" "HG" "PEAD" "PVC"] [Moda: Fuga tubo roto] (148)

DIÁMETRO PQR(mm) <= 15.875 [Moda: Fuga tubo roto] (115)

FUNCIONARIO MULTIPROPÓSITO (2) in ["Alex, Miguel" "Henry, Luis" "Luis" "Luis, Robinson" "Luis-Miguel"] [Moda: Fuga, acometida en mal estado] => Fuga, acometida en mal estado (14; 0,786)

FUNCIONARIO MULTIPROPÓSITO (2) in ["Alex" "Alex, Carlos" "Alex, Henry" "Alex, Juan C." "Carlos" "Carlos, Henry" "Carlos, Miguel" "Carlos, Robinson" "Henry" "Henry, Juan C." "Henry, Miguel" "Juan C" "Juan C., Luis" "Juan C., Miguel" "JuanC., Robinson" "Robinson" "Robinson, Alex" "Robinson-Samuel"] [Moda: Fuga tubo roto] (101)

P(mm) Jardin <= 38.500 [Moda: Fuga tubo roto] => Fuga tubo roto (91; 0,593)

P(mm) Jardin > 38.500 [Moda: Fuga, acometida en mal estado] => Fuga, acometida en mal estado (10; 0,7)

DIÁMETRO PQR(mm) > 15.875 [Moda: Fuga tubo roto] => Fuga tubo roto (33; 0,788)

Resumen de Configuración

Análisis
- Profundidad del árbol: 11

Campos
- Objetivo
  - TIPO DAÑO (1)
- Entradas
  - Coordenada X
  - Coordenada Y
  - RED
  - DIÁMETRO PQR(mm)
  - MATERIAL PQR
  - SE CORRIGIÓ DAÑO
  - FUNCIONARIO MULTIPROPÓSITO (2)
  - NIVEL RIESGO
  - HR(%) Bella
  - BS(Hr) Bella
  - P(mm) Bella
  - P(mm) Jardin
  - Material Red
  - Diámetro Red (mm)
  - Longitud Tramo (m)
  - Rugosidad
  - Coordenada Nudo inicio X
  - Coordenada Nudo inicio Y
  - Coordenada Nudo fin X
  - Coordenada Nudo fin Y
  - Distancia del PQR al nodo de inicio
  - Caudal (L/s) 00:00
  - Pérdida (m/Km) 0:00
  - Demanda Nodo Inicial (L/s) 00:00
  - Presión Nodo Final (mca) 00:00
  - Presión PQR (mca) 00:00
  - Caudal (L/s) 01:00
  - Pérdida (m/Km) 01:00
  - Demanda Nodo Inicial (L/s) 01:00
  - Presión Nodo Inicial (mca) 01:00
  - Presión Nodo Final (mca) 01:00
  - Presión PQR (mca) 1:00
  - Caudal (L/s) 02:00
  - Pérdida (m/Km) 02:00

Demanda Nodo Inicial (L/s) 02:00
Presión Nodo Inicial (mca) 02:00
Presión Nodo Final (mca) 02:00
Presión PQR (mca) 2:00
Caudal (L/s) 3:00
Pérdida (m/Km) 3:00
Presión Nodo Final (mca) 3:00
Presión PQR (mca) 3:00
Caudal (L/s) 4:00
Presión PQR (mca) 4:00
Caudal (L/s) 5:00
Demanda Nodo Inicial (L/s) 5:00
Presión Nodo Inicial (mca) 5:00
Presión PQR (mca) 5:00
Presión Nodo Inicial (mca) 6:00
Presión PQR (mca) 6:00
Demanda Nodo Inicial (L/s) 7:00
Presión Nodo Inicial (mca) 7:00
Presión PQR (mca) 7:00
Presión PQR (mca) 8:00
Presión PQR (mca) 13:00
Presión PQR (mca) 14:00
Presión PQR (mca) 15:00
Presión Nodo Inicial (mca) 16:00
Presión PQR (mca) 16:00
Presión Nodo Inicial (mca) 17:00
Presión PQR (mca) 17:00
Presión Nodo Inicial (mca) 18:00
Presión PQR (mca) 18:00
Presión Nodo Final (mca) 19:00
Presión PQR (mca) 19:00
Presión Nodo Inicial (mca) 20:00
Presión PQR (mca) 20:00
Presión PQR (mca) 21:00
Caudal (L/s) 22:00
Presión PQR (mca) 22:00
Caudal (L/s) 23:00
Pérdida (m/Km) 23:00
Presión Nodo Inicial (mca) 23:00
Caudal (L/s) 24:00
Pérdida (m/Km) 24:00

Configuración de creación
Utilizar los datos en particiones: falso
Utilizar frecuencia: falso
Utilizar ponderación: falso
Niveles por debajo del raíz: 20
Modo: Experto
Número máximo de sustitutos: 5
Cambio mínimo en la impureza: 0,0
Medida de impureza para objetivos categóricos: Gini
Criterios de parada: Utilizar porcentaje
Número mínimo de registros en rama parental (%): 2
Número mínimo de registros en rama filial (%): 1
Podar árbol: verdadero
Utilizar regla de error típico: verdadero
Multiplicador: 1,0
Probabilidades previas: Basadas en datos de entrenamiento
Corregir previas por costes de clasificación errónea: falso
Utilizar costes de clasificación errónea: falso

ANEXO 2F

Modelo Generado

MATERIAL PQR in ["" " "] [Moda: Fuga tubo roto] (611)

SE CORRIGIÓ DAÑO in ["No"] [Moda: Daño interno] (62)

Pérdida (m/Km) 01:00 <= 0.010 [Moda: Fuga tubo roto] => Fuga tubo roto (16; 0,25)

Pérdida (m/Km) 01:00 > 0.010 or Pérdida (m/Km) 01:00 IS MISSING [Moda: Daño interno] (46)

Demanda Nodo Inicial (L/s) 01:00 <= 0.060 [Moda: no hay daño] => no hay daño (9; 0,333)

Demanda Nodo Inicial (L/s) 01:00 > 0.060 or Demanda Nodo Inicial (L/s) 01:00 IS MISSING [Moda: Daño interno] => Daño interno (37; 0,405)

SE CORRIGIÓ DAÑO in ["Si"] [Moda: Fuga tubo roto] (549)

DIÁMETRO PQR(mm) <= 12.700 [Moda: Fuga llave de paso] => Fuga llave de paso (41; 0,537)

DIÁMETRO PQR(mm) > 12.700 or DIÁMETRO PQR(mm) IS MISSING [Moda: Fuga pitorra] (508)

Coordenada Nudo fin Y <= 990868 [Moda: Fuga tubo roto] (95)

Pérdida (m/Km) 22:00 <= 0 [Moda: Fuga tubo roto] => Fuga adaptador (10; 0,3)

Pérdida (m/Km) 22:00 > 0 [Moda: Fuga tubo roto] (85)

Material Red in ["PAD"] [Moda: Fuga tubo roto] => Fuga tubo roto (62; 0,468)

Material Red in ["PVC"] [Moda: Fuga llave de paso] => Fuga llave de paso (23; 0,261)

Coordenada Nudo fin Y > 990868 and Coordenada Nudo fin Y <= 992378 [Moda: Fuga pitorra] (190)

Demanda Nodo Inicial (L/s) 01:00 <= 0.030 [Moda: Fuga llave de paso] => Fuga llave de paso (21; 0,238)

Demanda Nodo Inicial (L/s) 01:00 > 0.030 and Demanda Nodo Inicial (L/s) 01:00 <= 0.040 [Moda: Fuga, acometida en mal estado] => Fuga, acometida en mal estado (26; 0,308)

Demanda Nodo Inicial (L/s) 01:00 > 0.040 and Demanda Nodo Inicial (L/s) 01:00 <= 0.060 [Moda: Fuga llave de paso] => Fuga llave de paso (22; 0,5)

Demanda Nodo Inicial (L/s) 01:00 > 0.060 and Demanda Nodo Inicial (L/s) 01:00 <= 0.110 [Moda: Fuga acometida] => Fuga acometida (29; 0,31)

Demanda Nodo Inicial (L/s) 01:00 > 0.110 [Moda: Fuga pitorra] => Fuga pitorra (92; 0,348)

Coordenada Nudo fin Y > 992378 [Moda: Fuga pitorra] (208)

Presión Nodo Final (mca) 16:00 <= 51.670 [Moda: Fuga pitorra] => Fuga pitorra (64; 0,344)

Presión Nodo Final (mca) 16:00 > 51.670 [Moda: Fuga pitorra] (144)

P(mm) Jardin <= 28 [Moda: Fuga pitorra] (127)

Distancia del PQR al nodo de inicio <= 16.997 [Moda: Fuga pitorra] => Fuga

pitorra (20; 0,35)

Distancia del PQR al nodo de inicio > 16.997 and Distancia del PQR al nodo de inicio <= 46.917 [Moda: Fuga llave de paso] => Fuga llave de paso (52; 0,327)

Distancia del PQR al nodo de inicio > 46.917 [Moda: Fuga tubo roto] => Fuga tubo roto (55; 0,364)

P(mm) Jardin > 28 [Moda: Fuga acometida] => Fuga acometida (9; 0,556)

or P(mm) Jardin IS MISSING [Moda: Fuga tubo roto] => Fuga tubo roto (8; 0,625)

or Coordenada Nudo fin Y IS MISSING [Moda: Fuga tubo roto] => Fuga tubo roto (15; 0,333)

MATERIAL PQR in ["HF" "PEAD" "PVC"] [Moda: Fuga tubo roto] => Fuga tubo roto (100; 0,57)

MATERIAL PQR in ["HG"] [Moda: Fuga tubo roto] (48)

RED in ["" "Principal"] [Moda: Fuga tubo roto] => Fuga tubo roto (14; 0,857)

RED in ["Domiciliaria"] [Moda: Fuga, acometida en mal estado] => Fuga, acometida en mal estado (34; 0,5)

MATERIAL PQR in ["PAD" "PF"] [Moda: Fuga adaptador] (87)

Presión Nodo Inicial (mca) 00:00 <= 46.350 [Moda: Fuga adaptador] => Fuga adaptador (14; 0,5)

Presión Nodo Inicial (mca) 00:00 > 46.350 or Presión Nodo Inicial (mca) 00:00 IS MISSING [Moda: Fuga adaptador] => Fuga adaptador (73; 0,397)

Resumen de Configuración

Análisis
- Profundidad del árbol: 7

Campos
- Objetivo
  - TIPO DAÑO (1)
- Entradas
  - RED
  - DIÁMETRO PQR(mm)
  - MATERIAL PQR
  - SE CORRIGIÓ DAÑO
  - P(mm) Jardin
  - Material Red
  - Coordenada Nudo fin Y
  - Distancia del PQR al nodo de inicio
  - Presión Nodo Inicial (mca) 00:00
  - Pérdida (m/Km) 01:00
  - Demanda Nodo Inicial (L/s) 01:00
  - Presión Nodo Final (mca) 16:00
  - Pérdida (m/Km) 22:00

Configuración de creación
- Utilizar los datos en particiones: falso
- Utilizar frecuencia: falso
- Utilizar ponderación: falso
- Niveles por debajo del raíz: 20
- Modo: Experto

Alfa para división: 0,05
Alfa para fusión: 0,05
Épsilon para convergencia: 0,001
Número máximo de iteraciones para la convergencia: 100
Utilizar corrección de Bonferroni: verdadero
Permitir división de categorías fusionadas: falso
Método de chi-cuadrado: Pearson
Criterios de parada: Utilizar porcentaje
Número mínimo de registros en rama parental (%): 2
Número mínimo de registros en rama filial (%): 1
Utilizar costes de clasificación errónea: falso

ANEXO 2G

Modelo Generado

MATERIAL PQR in ["" " "] [Moda: 1] (436)

SE CORRIGIÓ DAÑO = Si [Moda: 4] (387)

NIVEL RIESGO in [""] [Moda: 1] => 1 (30; 0,433)

NIVEL RIESGO in ["Alto" "Bajo" "Medio"] [Moda: 4] (357)

Coordenada Y <= 990621.782 [Moda: 1] (36)

Material Red in ["" "AC" "AP" "HG"] [Moda: 1] => 1 (2; 0,5)

Material Red in ["PAD"] [Moda: 1] (33)

Presión Nodo Final (mca) 00:00 <= 50.370 [Moda: 4] (12)

Presión Nodo Final (mca) 00:00 <= 47.710 [Moda: 1] (6)

P(mm) Bella <= 0.100 [Moda: 4] => 4 (3; 0,667)

P(mm) Bella > 0.100 [Moda: 1] => 1 (3; 1,0)

Presión Nodo Final (mca) 00:00 > 47.710 [Moda: 4] (6)

BS(Hr) Bella <= 1.500 [Moda: 4] => 4 (3; 1,0)

BS(Hr) Bella > 1.500 [Moda: 5] => 5 (3; 0,667)

Presión Nodo Final (mca) 00:00 > 50.370 [Moda: 1] (21)

CFuncionario(1) in ["10" "16" "24"] [Moda: 3] => 3 (3,15; 0,952)

CFuncionario(1) in ["11" "29" "4" "6"] [Moda: 1] (17,85)

Longitud Tramo (m) <= 93.350 [Moda: 1] (10)

P(mm) Quebradanegra <= 4 [Moda: 10] => 10 (4,444; 0,9)

P(mm) Quebradanegra > 4 [Moda: 1] => 1 (5,556; 0,82)

Longitud Tramo (m) > 93.350 [Moda: 1] => 1 (7,85; 0,637)

CFuncionario(1) in ["1" "12" "13" "14" "15" "17" "18" "19" "2" "20" "21" "22" "23" "25" "26" "27" "28" "3" "30" "5" "7" "8" "9"] [Moda: 1] => 1 (0)

Material Red in ["PVC"] [Moda: 5] => 5 (1; 1,0)

Coordenada Y > 990621.782 [Moda: 4] (321)

CFuncionario(1) in ["1"] [Moda: 4] (8,257)

Diámetro Red (mm) <= 75 [Moda: 4] (5,232)

Presión Nodo Inicial (mca) 01:00 <= 52.790 [Moda: 4] => 4 (3,051; 0,983)

Presión Nodo Inicial (mca) 01:00 > 52.790 [Moda: 6] => 6 (2,18; 0,953)

Diámetro Red (mm) > 75 [Moda: 12] => 12 (3,026; 0,33)

CFuncionario(1) in ["10"] [Moda: 13] => 13 (2,064; 0,484)

CFuncionario(1) in ["11"] [Moda: 1] (8,257)

Diámetro Red (mm) <= 25 [Moda: 12] => 12 (2; 0,5)

Diámetro Red (mm) > 25 [Moda: 1] => 1 (6,257; 0,811)

CFuncionario(1) in ["12"] [Moda: 6] (11,354)

Caudal (L/s) 00:00 <= 0.570 [Moda: 6] => 6 (7,248; 0,847)

Caudal (L/s) 00:00 > 0.570 [Moda: 13] (4,106)

Presión Nodo Final (mca) 01:00 <= 51.890 [Moda: 13] => 13 (2,035; 0,983)

Presión Nodo Final (mca) 01:00 > 51.890 [Moda: 1] => 1 (2,071; 0,517)

CFuncionario(1) in ["14" "21"] [Moda: 6] => 6 (5,161; 0,597)

CFuncionario(1) in ["15"] [Moda: 1] (10,322)

NIVEL RIESGO in ["Bajo"] [Moda: 6] (6,161)

Diámetro Red (mm) <= 50 [Moda: 6] => 6 (2; 1,0)

Diámetro Red (mm) > 50 [Moda: 1] (4,161)

Demanda Nodo Final (L/s) 00:00 <= 0.060 [Moda: 12] => 12 (2,032; 0,492)

Demanda Nodo Final (L/s) 00:00 > 0.060 [Moda: 1] => 1 (2,129; 0,97)

NIVEL RIESGO in ["Medio"] [Moda: 4] (4,161)

HR(%) Bella <= 78.300 [Moda: 1] => 1 (2,129; 0,485)

HR(%) Bella > 78.300 [Moda: 4] => 4 (2,032; 0,984)

NIVEL RIESGO in ["Alto"] [Moda: 1] => 1 (0)

CFuncionario(1) in ["16"] [Moda: 6] (19,611)

P(mm) Bella <= 0.500 [Moda: 1] (12,244)

P(mm) Quebradanegra <= 15 [Moda: 1] (9,244)

Diámetro Red (mm) <= 50 [Moda: 6] => 6 (2,122; 0,5)

Diámetro Red (mm) > 50 [Moda: 1] (7,122)

Rugosidad <= 0.002 [Moda: 8] => 8 (3,061; 0,653)

Rugosidad > 0.002 [Moda: 1] => 1 (4,061; 0,754)

P(mm) Quebradanegra > 15 [Moda: 10] => 10 (3; 0,333)

P(mm) Bella > 0.500 [Moda: 6] => 6 (7,367; 0,983)

CFuncionario(1) in ["17"] [Moda: 4] (9,289)

NIVEL RIESGO in ["Bajo"] [Moda: 6] (5,145)

Pérdida (m/Km) 0:00 <= 0.080 [Moda: 6] => 6 (2,116; 0,514)

Pérdida (m/Km) 0:00 > 0.080 [Moda: 10] => 10 (3,029; 0,33)

NIVEL RIESGO in ["Medio"] [Moda: 4] (4,145)

Caudal (L/s) 00:00 <= 0.090 [Moda: 1] => 1 (2,029; 0,493)

Caudal (L/s) 00:00 > 0.090 [Moda: 4] => 4 (2,116; 0,945)

NIVEL RIESGO in ["Alto"] [Moda: 4] => 4 (0)

CFuncionario(1) in ["18"] [Moda: 6] (14,45)

T (ºC) Bella <= 22.300 [Moda: 6] => 6 (12,405; 0,502)

T (ºC) Bella > 22.300 [Moda: 13] => 13 (2,045; 0,978)

CFuncionario(1) in ["19" "30"] [Moda: 12] => 12 (2,064; 0,969)

CFuncionario(1) in ["2"] [Moda: 6] (10,322)

Pérdida (m/Km) 5:00 <= 0.410 [Moda: 6] => 6 (5,161; 0,794)

Pérdida (m/Km) 5:00 > 0.410 [Moda: 1] => 1 (5,161; 0,788)

CFuncionario(1) in ["20"] [Moda: 6] (4,129)

Presión Nodo Final (mca) 01:00 <= 57.010 [Moda: 6] => 6 (2,064; 0,994)

Presión Nodo Final (mca) 01:00 > 57.010 [Moda: 4] => 4 (2,064; 0,969)

CFuncionario(1) in ["13" "22" "23" "26" "27" "28"] [Moda: 4] => 4 (4,129; 0,969)

CFuncionario(1) in ["24"] [Moda: 10] (5,161)

Presión Nodo Final (mca) 01:00 <= 51.030 [Moda: 10] => 10 (3,032; 0,66)

Presión Nodo Final (mca) 01:00 > 51.030 [Moda: 1] => 1 (2,129; 0,492)

CFuncionario(1) in ["25"] [Moda: 4] (12,386)

Demanda Nodo Final (L/s) 00:00 <= 0.070 [Moda: 4] (6,154)

Distancia del PQR al nodo de inicio <= 33.337 [Moda: 13] => 13 (2,077; 0,481)

Distancia del PQR al nodo de inicio > 33.337 [Moda: 4] => 4 (4,077; 0,981)

Demanda Nodo Final (L/s) 00:00 > 0.070 [Moda: 6] (6,232)

Coordenada Nudo inicio X <= 1159401.250 [Moda: 6] => 6 (4,193; 0,734)

Coordenada Nudo inicio X > 1159401.250 [Moda: 8] => 8 (2,039; 0,981)

CFuncionario(1) in ["29"] [Moda: 4] (50,576)

Material Red in ["AC"] [Moda: 8] => 8 (1; 1,0)

Material Red in ["HG"] [Moda: 4] (15,473)

Demanda Nodo Inicial (L/s) 01:00 <= 0.070 [Moda: 10] (6,158)

Distancia del PQR al nodo de inicio <= 53.124 [Moda: 12] => 12 (3,158; 0,633)

Distancia del PQR al nodo de inicio > 53.124 [Moda: 10] => 10 (3; 0,667)

Demanda Nodo Inicial (L/s) 01:00 > 0.070 [Moda: 4] (9,315)

TIEMPO REPARACIÓN <= 0 [Moda: 6] (7,315)

Demanda Nodo Inicial (L/s) 01:00 <= 0.100 [Moda: 4] => 4 (3,158; 0,633)

Demanda Nodo Inicial (L/s) 01:00 > 0.100 [Moda: 6] => 6 (4,158; 0,759)

TIEMPO REPARACIÓN > 0 [Moda: 4] => 4 (2; 1,0)

Material Red in ["" "AP" "PAD"] [Moda: 4] => 4 (2,158; 0,463)

Material Red in ["PVC"] [Moda: 4] (31,945)

Coordenada Nudo fin Y <= 992477 [Moda: 4] (21,788)

Coordenada Nudo fin Y <= 990781.940 [Moda: 2] => 2 (2; 1,0)

Coordenada Nudo fin Y > 990781.940 [Moda: 4] (19,788)

Demanda Nodo Final (L/s) 00:00 <= 0.150 [Moda: 13] (15,63)

TIEMPO REPARACIÓN <= 0 [Moda: 13] (11,315)

Demanda Nodo Final (L/s) 02:00 <= 0.070 [Moda: 4] => 4 (5,158; 0,776)

Demanda Nodo Final (L/s) 02:00 > 0.070 [Moda: 13] => 13 (6,158; 0,812)

TIEMPO REPARACIÓN > 0 [Moda: 4] (4,315)

Demanda Nodo Final (L/s) 00:00 <= 0.050 [Moda: 8] => 8 (2,158; 0,927)

Demanda Nodo Final (L/s) 00:00 > 0.050 [Moda: 4] => 4 (2,158; 0,927)

Demanda Nodo Final (L/s) 00:00 > 0.150 [Moda: 6] (4,158)

Presión Nodo Inicial (mca) 10:00 <= 64.820 [Moda: 1] => 1 (2,158; 0,537)

Presión Nodo Inicial (mca) 10:00 > 64.820 [Moda: 6] => 6 (2; 1,0)

Coordenada Nudo fin Y > 992477 [Moda: 1] (10,158)

Pérdida (m/Km) 01:00 <= 0.020 [Moda: 4] => 4 (2; 1,0)

Pérdida (m/Km) 01:00 > 0.020 [Moda: 1] => 1 (8,158; 0,736)

CFuncionario(1) in ["3"] [Moda: 6] (7,225)

TIEMPO REPARACIÓN <= 1 [Moda: 6] (5,145)

NIVEL RIESGO in ["Bajo"] [Moda: 1] => 1 (3,084; 0,661)

NIVEL RIESGO in ["Alto" "Medio"] [Moda: 6] => 6 (2,061; 0,989)

TIEMPO REPARACIÓN > 1 [Moda: 6] => 6 (2,08; 0,975)

CFuncionario(1) in ["4"] [Moda: 4] (40,254)

Material Red in ["AC"] [Moda: 6] (6)

BS(Hr) Bella <= 3.300 [Moda: 1] => 1 (3; 0,333)

BS(Hr) Bella > 3.300 [Moda: 6] => 6 (3; 0,667)

Material Red in ["HG"] [Moda: 4] (8,376)

P(mm) Bella <= 4.100 [Moda: 4] (5,251)

Demanda Nodo Inicial (L/s) 00:00 <= 0.070 [Moda: 4] => 4 (2,125; 0,941)

Demanda Nodo Inicial (L/s) 00:00 > 0.070 [Moda: 10] => 10 (3,125; 0,64)

P(mm) Bella > 4.100 [Moda: 6] => 6 (3,125; 0,36)

Material Red in ["PAD"] [Moda: 1] => 1 (1,125; 1,0)

Material Red in ["PVC"] [Moda: 4] (24,752)

Caudal (L/s) 00:00 <= 0.020 [Moda: 10] => 10 (2; 0,5)

Caudal (L/s) 00:00 > 0.020 [Moda: 4] (22,752)

T (ºC) Bella <= 19.600 [Moda: 6] (5,623)

Distancia del PQR al nodo de inicio <= 24.386 [Moda: 6] => 6 (2,498; 1,0)

Distancia del PQR al nodo de inicio > 24.386 [Moda: 2] => 2 (3,125; 0,64)

T (ºC) Bella > 19.600 [Moda: 4] (17,129)

Presión PQR (mca) 1:00 <= 38.001 [Moda: 1] => 1 (2; 0,5)

Presión PQR (mca) 1:00 > 38.001 [Moda: 4] (15,129)

NIVEL RIESGO in ["Bajo"] [Moda: 4] (7,878)

Presión Nodo Final (mca) 00:00 <= 52.120 [Moda: 4] => 4 (4; 1,0)

Presión Nodo Final (mca) 00:00 > 52.120 [Moda: 12] => 12 (3,878; 0,516)

NIVEL RIESGO in ["Medio"] [Moda: 4] (7,251)

Coordenada Nudo fin Y <= 991757 [Moda: 13] => 13 (2,125; 0,941)

Coordenada Nudo fin Y > 991757 [Moda: 4] => 4 (5,125; 0,976)

NIVEL RIESGO in ["Alto"] [Moda: 4] => 4 (0)

Material Red in ["" "AP"] [Moda: 4] => 4 (0)

CFuncionario(1) in ["5"] [Moda: 12] (4,129)

HR(%) Bella <= 74 [Moda: 12] => 12 (2,077; 0,481)

HR(%) Bella > 74 [Moda: 3] => 3 (2,051; 0,487)

CFuncionario(1) in ["6"] [Moda: 6] (68,122)

Presión Nodo Final (mca) 00:00 <= 41.330 [Moda: 12] (8,212)

Demanda Nodo Final (L/s) 00:00 <= 0.130 [Moda: 12] (6,212)

Longitud Tramo (m) <= 75.360 [Moda: 13] => 13 (2,212; 0,904)

Longitud Tramo (m) > 75.360 [Moda: 12] (4)

Caudal (L/s) 10:00 <= 1.720 [Moda: 4] => 4 (2; 1,0)

Caudal (L/s) 10:00 > 1.720 [Moda: 12] => 12 (2; 1,0)

Demanda Nodo Final (L/s) 00:00 > 0.130 [Moda: 10] => 10 (2; 0,5)

Presión Nodo Final (mca) 00:00 > 41.330 [Moda: 6] (59,91)

Diámetro Red (mm) <= 19 [Moda: 12] => 12 (3; 0,667)

Diámetro Red (mm) > 19 [Moda: 6] (56,91)

Presión Nodo Inicial (mca) 10:00 <= 48.100 [Moda: 6] (18,212)

T (ºC) Bella <= 20.700 [Moda: 6] (14)

P(mm) Bella <= 10.500 [Moda: 6] => 6 (10; 1,0)

P(mm) Bella > 10.500 [Moda: 1] (4)

Caudal (L/s) 00:00 <= 0.160 [Moda: 1] => 1 (2; 1,0)

Caudal (L/s) 00:00 > 0.160 [Moda: 6] => 6 (2; 1,0)

T (ºC) Bella > 20.700 [Moda: 4] => 4 (4,212; 0,712)

Presión Nodo Inicial (mca) 10:00 > 48.100 [Moda: 1] (38,698)

Caudal (L/s) 6:00 <= 0.190 [Moda: 1] (12,849)

Presión Nodo Inicial (mca) 11:00 <= 58 [Moda: 1] (8,849)

Pérdida (m/Km) 6:00 <= 0.010 [Moda: 2] => 2 (2; 0,5)

Pérdida (m/Km) 6:00 > 0.010 [Moda: 1] => 1 (6,849; 0,907)

Presión Nodo Inicial (mca) 11:00 > 58 [Moda: 3] (4)

Caudal (L/s) 00:00 <= 0.070 [Moda: 6] => 6 (2; 1,0)

Caudal (L/s) 00:00 > 0.070 [Moda: 3] => 3 (2; 1,0)

Caudal (L/s) 6:00 > 0.190 [Moda: 4] (25,849)

Presión Nodo Final (mca) 02:00 <= 70.380 [Moda: 4] (17,849)

Longitud Tramo (m) <= 105.660 [Moda: 3] (9,849)

Pérdida (m/Km) 01:00 <= 0.180 [Moda: 4] => 4 (4,212; 0,712)

Pérdida (m/Km) 01:00 > 0.180 [Moda: 3] => 3 (5,637; 0,71)

Longitud Tramo (m) > 105.660 [Moda: 4] => 4 (8; 0,875)

Presión Nodo Final (mca) 02:00 > 70.380 [Moda: 1] => 1 (8; 0,625)

CFuncionario(1) in ["7"] [Moda: 13] (15,482)

TIEMPO REPARACIÓN <= 0 [Moda: 6] (11,312)

Presión Nodo Inicial (mca) 01:00 <= 44.840 [Moda: 13] => 13 (3,048; 0,656)

Presión Nodo Inicial (mca) 01:00 > 44.840 [Moda: 1] (8,263)

Caudal (L/s) 11:00 <= 0.290 [Moda: 1] => 1 (4,167; 0,728)

Caudal (L/s) 11:00 > 0.290 [Moda: 6] => 6 (4,096; 0,732)

TIEMPO REPARACIÓN > 0 [Moda: 13] (4,171)

Caudal (L/s) 00:00 <= 0.090 [Moda: 13] => 13 (2,096; 0,954)

Caudal (L/s) 00:00 > 0.090 [Moda: 3] => 3 (2,074; 0,964)

CFuncionario(1) in ["8"] [Moda: 1] => 1 (3,096; 0,655)

CFuncionario(1) in ["9"] [Moda: 13] (5,161)

P(mm) Bella <= 48.900 [Moda: 4] => 4 (2,161; 0,926)

P(mm) Bella > 48.900 [Moda: 13] => 13 (3; 0,667)

SE CORRIGIÓ DAÑO = No [Moda: 1] (49)

Coordenada Nudo inicio X <= 1157962 [Moda: 8] (6)

Demanda Nodo Final (L/s) 00:00 <= 0.060 [Moda: 8] => 8 (2; 1,0)

Demanda Nodo Final (L/s) 00:00 > 0.060 [Moda: 9] (4)

Caudal (L/s) 00:00 <= 0.120 [Moda: 13] => 13 (2; 0,5)

Caudal (L/s) 00:00 > 0.120 [Moda: 9] => 9 (2; 1,0)

Coordenada Nudo inicio X > 1157962 [Moda: 1] (43)

NIVEL RIESGO in ["" "Alto"] [Moda: 1] => 1 (3; 0,667)

NIVEL RIESGO in ["Bajo"] [Moda: 7] (17)

Demanda Nodo Final (L/s) 4:00 <= 0.190 [Moda: 7] (15)

Rugosidad <= 0.030 [Moda: 7] (9)

Coordenada Nudo fin Y <= 990530 [Moda: 5] => 5 (2; 0,5)

Coordenada Nudo fin Y > 990530 [Moda: 7] => 7 (7; 0,857)

Rugosidad > 0.030 [Moda: 9] (6)

Coordenada Nudo fin X <= 1159320.750 [Moda: 9] => 9 (3; 1,0)

Coordenada Nudo fin X > 1159320.750 [Moda: 1] => 1 (3; 0,667)

Demanda Nodo Final (L/s) 4:00 > 0.190 [Moda: 5] => 5 (2; 0,5)

NIVEL RIESGO in ["Medio"] [Moda: 12] (23)

Coordenada Nudo fin X <= 1159936 [Moda: 12] (20)

Longitud Tramo (m) <= 69.720 [Moda: 9] => 9 (3; 1,0)

Longitud Tramo (m) > 69.720 [Moda: 12] (17)

Pérdida (m/Km) 9:00 <= 10.540 [Moda: 1] => 1 (12; 0,5)

Pérdida (m/Km) 9:00 > 10.540 [Moda: 12] => 12 (5; 0,8)

Coordenada Nudo fin X > 1159936 [Moda: 10] => 10 (3; 0,667)

MATERIAL PQR in ["HF" "HG" "PAD" "PEAD" "PF" "PVC"] [Moda: 1] (161)

MATERIAL PQR in ["HF" "PEAD"] [Moda: 1] => 1 (2; 1,0)

MATERIAL PQR in ["HG"] [Moda: 1] (31)

Demanda Nodo Final (L/s) 5:00 <= 0.300 [Moda: 1] (28)

RED in ["" "Principal"] [Moda: 1] => 1 (13; 0,846)

RED in ["Domiciliaria"] [Moda: 10] (15)

Presión Nodo Final (mca) 7:00 <= 48.240 [Moda: 1] => 1 (3; 1,0)

Presión Nodo Final (mca) 7:00 > 48.240 [Moda: 10] => 10 (12; 0,917)

Demanda Nodo Final (L/s) 5:00 > 0.300 [Moda: 7] => 7 (3; 0,667)

MATERIAL PQR in ["PAD"] [Moda: 1] (5)

BS(Hr) Bella <= 2.800 [Moda: 3] => 3 (2; 1,0)

BS(Hr) Bella > 2.800 [Moda: 1] => 1 (3; 0,667)

MATERIAL PQR in ["PF"] [Moda: 3] (56)

Presión Nodo Inicial (mca) 10:00 <= 44.420 [Moda: 3] (12,444)

Demanda Nodo Final (L/s) 4:00 <= 0.120 [Moda: 3] => 3 (10,37; 0,771)

Demanda Nodo Final (L/s) 4:00 > 0.120 [Moda: 10] => 10 (2,074; 0,482)

Presión Nodo Inicial (mca) 10:00 > 44.420 [Moda: 3] (43,556)

CFuncionario(1) in ["1" "10" "11" "12" "13" "14" "15" "17" "19" "20" "22" "23" "24" "26" "28" "3" "30" "5" "8" "9"] [Moda: 3] => 3 (4; 1,0)

CFuncionario(1) in ["16" "2" "21" "25" "7"] [Moda: 1] => 1 (9; 0,667)

CFuncionario(1) in ["18"] [Moda: 3] (4)

Caudal (L/s) 00:00 <= 0.470 [Moda: 3] => 3 (2; 1,0)

Caudal (L/s) 00:00 > 0.470 [Moda: 1] => 1 (2; 0,5)

CFuncionario(1) in ["27"] [Moda: 2] => 2 (2; 0,5)

CFuncionario(1) in ["29"] [Moda: 1] (6,778)

Pérdida (m/Km) 01:00 <= 0.160 [Moda: 1] => 1 (4,519; 0,557)

Pérdida (m/Km) 01:00 > 0.160 [Moda: 2] => 2 (2,259; 0,885)

CFuncionario(1) in ["4"] [Moda: 3] (11)

NIVEL RIESGO in ["Bajo"] [Moda: 3] (9)

Coordenada Nudo inicio Y <= 990497.250 [Moda: 2] (4)

Caudal (L/s) 00:00 <= 0.120 [Moda: 2] => 2 (2; 1,0)

Caudal (L/s) 00:00 > 0.120 [Moda: 1] => 1 (2; 0,5)

Coordenada Nudo inicio Y > 990497.250 [Moda: 3] => 3 (5; 1,0)

NIVEL RIESGO in ["" "Alto" "Medio"] [Moda: 2] => 2 (2; 1,0)

CFuncionario(1) in ["6"] [Moda: 2] (6,778)

Coordenada Nudo fin X <= 1158626 [Moda: 1] => 1 (2,259; 0,885)

Coordenada Nudo fin X > 1158626 [Moda: 2] => 2 (4,519; 1,0)

MATERIAL PQR in ["PVC"] [Moda: 1] (67)

CFuncionario(1) in ["1" "5"] [Moda: 10] => 10 (3; 0,667)

CFuncionario(1) in ["11" "28"] [Moda: 3] => 3 (2; 1,0)

CFuncionario(1) in ["10" "12" "13" "14" "15" "16" "17" "18" "19" "21" "22" "23" "24" "26" "27" "29" "3" "30" "8" "9"] [Moda: 1] => 1 (24; 0,833)

CFuncionario(1) in ["2"] [Moda: 6] => 6 (1; 1,0)

CFuncionario(1) in ["20"] [Moda: 10] (4)

BS(Hr) Bella <= 3.500 [Moda: 3] => 3 (2; 1,0)

BS(Hr) Bella > 3.500 [Moda: 10] => 10 (2; 1,0)

CFuncionario(1) in ["25"] [Moda: 1] (10)

Diámetro Red (mm) <= 100 [Moda: 1] => 1 (8; 0,75)

Diámetro Red (mm) > 100 [Moda: 3] => 3 (2; 1,0)

CFuncionario(1) in ["4"] [Moda: 1] (13)

Diámetro Red (mm) <= 75 [Moda: 1] (11)

Coordenada Nudo inicio Y <= 991529 [Moda: 3] => 3 (3; 0,667)

Coordenada Nudo inicio Y > 991529 [Moda: 1] => 1 (8; 1,0)

Diámetro Red (mm) > 75 [Moda: 10] => 10 (2; 0,5)

CFuncionario(1) in ["6"] [Moda: 3] (9)

Demanda Nodo Inicial (L/s) 02:00 <= 0.090 [Moda: 10] => 10 (2; 0,5)

Demanda Nodo Inicial (L/s) 02:00 > 0.090 [Moda: 3] (7)

TIEMPO REPARACIÓN <= 0 [Moda: 1] => 1 (4; 0,75)

TIEMPO REPARACIÓN > 0 [Moda: 3] => 3 (3; 1,0)

CFuncionario(1) in ["7"] [Moda: 11] => 11 (1; 1,0)

Resumen de Configuración

Análisis
- Profundidad del árbol: 12
- Validación cruzada
  - Media: 28,5
  - Error típico: 1,3

Campos
- Objetivo
  - CTipo Daño
- Entradas
  - MATERIAL PQR
  - SE CORRIGIÓ DAÑO
  - NIVEL RIESGO
  - Coordenada Y
  - Material Red
  - Presión Nodo Final (mca) 00:00
  - P(mm) Bella
  - BS(Hr) Bella
  - CFuncionario(1)
  - Longitud Tramo (m)
  - P(mm) Quebradanegra
  - Diámetro Red (mm)
  - Presión Nodo Inicial (mca) 01:00
  - Caudal (L/s) 00:00
  - Presión Nodo Final (mca) 01:00
  - Demanda Nodo Final (L/s) 00:00
  - HR(%) Bella
  - Rugosidad
  - Pérdida (m/Km) 0:00
  - T (ºC) Bella
  - Pérdida (m/Km) 5:00
  - Distancia del PQR al nodo de inicio
  - Coordenada Nudo inicio X
  - Demanda Nodo Inicial (L/s) 01:00
  - TIEMPO REPARACIÓN
  - Coordenada Nudo fin Y
  - Demanda Nodo Final (L/s) 02:00
  - Presión Nodo Inicial (mca) 10:00
  - Pérdida (m/Km) 01:00
  - Demanda Nodo Inicial (L/s) 00:00
  - Presión PQR (mca) 1:00
  - Caudal (L/s) 10:00
  - Caudal (L/s) 6:00
  - Presión Nodo Inicial (mca) 11:00
  - Pérdida (m/Km) 6:00
  - Presión Nodo Final (mca) 02:00
  - Caudal (L/s) 11:00
  - Demanda Nodo Final (L/s) 4:00
  - Coordenada Nudo fin X
  - Pérdida (m/Km) 9:00

Demanda Nodo Final (L/s) 5:00
RED
Presión Nodo Final (mca) 7:00
Coordenada Nudo inicio Y
Demanda Nodo Inicial (L/s) 02:00

Configuración de creación
Utilizar los datos en particiones: verdadero
Partición: Partición
Tipo de resultados: Árbol de decisión
Agrupar simbólicos: verdadero
Utilizar aumento: falso
Efectuar validación cruzada: verdadero
Número de veces: 10
Modo: Experto
Gravedad de la poda: 75
Número mínimo de registros por rama filial: 2
Valoración incial de atributos: falso
Utilizar costes de clasificación errónea: falso

ANEXO 2H

Modelo Generado

MATERIAL PQR in ["" " "] [Moda: 1] (436)

SE CORRIGIÓ DAÑO in ["No"] [Moda: 1] (49)

CFuncionario(1) in ["10" "16" "18" "20" "25" "26" "29" "7"] [Moda: 7] (32)

Distancia del PQR al nodo de inicio <= 95.855 [Moda: 7] (24)

Pérdida (m/Km) 0:00 <= 0.280 [Moda: 1] (18)

Demanda Nodo Final (L/s) 00:00 <= 0.065 [Moda: 1] => 1 (9; 0,667)

Demanda Nodo Final (L/s) 00:00 > 0.065 [Moda: 7] => 5 (9; 0,222)

Pérdida (m/Km) 0:00 > 0.280 [Moda: 7] => 7 (6; 0,833)

Distancia del PQR al nodo de inicio > 95.855 [Moda: 12] => 12 (8; 0,625)

CFuncionario(1) in ["11" "12" "14" "6" "9"] [Moda: 9] (17)

Coordenada Nudo fin Y <= 992315.155 [Moda: 9] => 9 (7; 1,0)

Coordenada Nudo fin Y > 992315.155 [Moda: 1] => 1 (10; 0,3)

SE CORRIGIÓ DAÑO in ["Si"] [Moda: 4] (387)

Coordenada Y <= 990736.494 [Moda: 1] => 1 (54; 0,463)

Coordenada Y > 990736.494 [Moda: 4] (333)

Coordenada Nudo inicio X <= 1158706 [Moda: 6] (62)

CFuncionario(1) in ["12" "16" "17" "18" "2" "3" "6" "7"] [Moda: 6] (39)

Presión Nodo Inicial (mca) 6:00 <= 51.455 [Moda: 6] => 6 (32; 0,688)

Presión Nodo Inicial (mca) 6:00 > 51.455 [Moda: 3] => 1 (7; 0,286)

CFuncionario(1) in ["10" "15" "19" "25" "29" "4" "9"] [Moda: 4] => 13 (23; 0,217)

Coordenada Nudo inicio X > 1158706 [Moda: 4] (271)

Longitud Tramo (m) <= 27.135 [Moda: 10] => 10 (8; 0,75)

Longitud Tramo (m) > 27.135 [Moda: 4] (263)

Coordenada Nudo fin Y <= 993953 [Moda: 4] (256)

Presión Nodo Inicial (mca) 19:00 <= 41.025 [Moda: 13] (26)

CFuncionario(1) in ["1" "20" "21" "25" "3" "4" "5" "6"] [Moda: 6] => 4 (17; 0,294)

CFuncionario(1) in ["11" "15" "24" "29" "7"] [Moda: 13] => 13 (9; 0,778)

Presión Nodo Inicial (mca) 19:00 > 41.025 [Moda: 4] (230)

Material Red in ["AC" "HG"] [Moda: 6] (106)

CFuncionario(1) in ["11" "2" "3"] [Moda: 1] => 1 (13; 0,692)

CFuncionario(1) in ["1" "12" "15" "16" "17" "18" "20" "21" "23" "24" "25" "27" "29" "4" "5" "6" "7"] [Moda: 6] => 6 (93; 0,312)

Material Red in ["" "AP" "PAD" "PVC"] [Moda: 1] (124)

Longitud Tramo (m) <= 163.325 [Moda: 1] (115)

CFuncionario(1) in ["11" "12" "14" "15" "16" "17" "2" "21" "22" "24" "25" "29" "3" "30" "4" "6" "7" "8"] [Moda: 1] (103)

Demanda Nodo Inicial (L/s) 00:00 <= 0.055 [Moda: 6] (34)

CFuncionario(1) in ["12" "14" "2" "3" "4" "7"] [Moda: 6] => 6 (14; 0,571)

CFuncionario(1) in ["11" "16" "17" "29" "30" "6" "8"] [Moda: 4] (20)

P(mm) Bella <= 0.150 [Moda: 4] => 4 (11; 0,545)

P(mm) Bella > 0.150 [Moda: 1] => 1 (9; 0,333)

Demanda Nodo Inicial (L/s) 00:00 > 0.055 [Moda: 1] (69)

P(mm) Bella <= 9.850 [Moda: 1] => 1 (51; 0,569)

P(mm) Bella > 9.850 [Moda: 4] => 4 (18; 0,278)

CFuncionario(1) in ["1" "10" "18" "23" "26" "9"] [Moda: 4] => 4 (12; 0,667)

Longitud Tramo (m) > 163.325 [Moda: 4] => 4 (9; 0,889)

Coordenada Nudo fin Y > 993953 [Moda: 4] => 4 (7; 0,857)

MATERIAL PQR in ["HF" "HG" "PAD" "PEAD" "PF" "PVC"] [Moda: 1] (161)

DIÁMETRO PQR(mm) <= 22.700 [Moda: 1] (132)

DIÁMETRO PQR(mm) <= 14.350 [Moda: 1] (79)

CFuncionario(1) in ["1" "20" "21" "26" "30" "5"] [Moda: 10] => 10 (14; 0,786)

CFuncionario(1) in ["10" "11" "16" "17" "18" "2" "25" "28" "29" "4" "6" "7" "8" "9"] [Moda: 1] (65)

P(mm) Jardin <= 12.500 [Moda: 1] => 1 (49; 0,653)

P(mm) Jardin > 12.500 [Moda: 3] => 3 (16; 0,438)

DIÁMETRO PQR(mm) > 14.350 [Moda: 3] (53)

CFuncionario(1) in ["12" "17" "18" "2" "21" "25" "29" "6"] [Moda: 1] (28)

Longitud Tramo (m) <= 84.705 [Moda: 1] => 1 (14; 0,714)

Longitud Tramo (m) > 84.705 [Moda: 2] => 2 (14; 0,5)

CFuncionario(1) in ["10" "11" "16" "27" "4" "7" "8" "9"] [Moda: 3] (25)

Pérdida (m/Km) 0:00 <= 0.015 [Moda: 2] => 2 (13; 0,462)

Pérdida (m/Km) 0:00 > 0.015 [Moda: 3] => 3 (12; 0,833)

DIÁMETRO PQR(mm) > 22.700 [Moda: 1] => 1 (29; 0,724)

Resumen de Configuración

Análisis
- Profundidad del árbol: 13

Campos
- Objetivo
  - CTipo Daño
- Entradas
  - Coordenada X
  - Coordenada Y
  - RED
  - DIÁMETRO PQR(mm)
  - MATERIAL PQR
  - TIEMPO REPARACIÓN
  - SE CORRIGIÓ DAÑO
  - NIVEL RIESGO
  - T (ºC) Bella
  - HR(%) Bella
  - BS(Hr) Bella
  - P(mm) Bella
  - P(mm) Jardin
  - P(mm) Quebradanegra
  - Material Red
  - Longitud Tramo (m)
  - Rugosidad
  - Coordenada Nudo inicio X
  - Coordenada Nudo inicio Y
  - Coordenada Nudo fin X
  - Coordenada Nudo fin Y
  - Distancia del PQR al nodo de inicio
  - Caudal (L/s) 00:00
  - Pérdida (m/Km) 0:00
  - Demanda Nodo Inicial (L/s) 00:00
  - Presión Nodo Inicial (mca) 00:00
  - Demanda Nodo Final (L/s) 00:00
  - Presión Nodo Final (mca) 00:00
  - Presión PQR (mca) 00:00
  - Pérdida (m/Km) 01:00
  - Demanda Nodo Inicial (L/s) 01:00
  - Presión Nodo Inicial (mca) 01:00
  - Presión Nodo Final (mca) 01:00
  - Pérdida (m/Km) 02:00
  - Demanda Nodo Inicial (L/s) 02:00
  - Presión Nodo Inicial (mca) 02:00
  - Demanda Nodo Final (L/s) 02:00
  - Presión Nodo Final (mca) 02:00
  - Pérdida (m/Km) 3:00
  - Demanda Nodo Inicial (L/s) 3:00
  - Demanda Nodo Final (L/s) 3:00
  - Pérdida (m/Km) 4:00
  - Demanda Nodo Inicial (L/s) 4:00
  - Presión Nodo Inicial (mca) 4:00
  - Demanda Nodo Final (L/s) 4:00

Presión Nodo Final (mca) 4:00
Caudal (L/s) 5:00
Pérdida (m/Km) 5:00
Demanda Nodo Final (L/s) 5:00
Pérdida (m/Km) 6:00
Presión Nodo Inicial (mca) 6:00
Presión Nodo Inicial (mca) 7:00
Presión Nodo Inicial (mca) 8:00
Pérdida (m/Km) 9:00
Presión Nodo Inicial (mca) 9:00
Presión Nodo Inicial (mca) 10:00
Presión Nodo Inicial (mca) 11:00
Presión PQR (mca) 11:00
Presión PQR (mca) 13:00
Demanda Nodo Final (L/s) 14:00
Presión PQR (mca) 15:00
Presión PQR (mca) 17:00
Demanda Nodo Inicial (L/s) 19:00
Presión Nodo Inicial (mca) 19:00
Presión PQR (mca) 19:00
Pérdida (m/Km) 24:00
CFuncionario(1)

Configuración de creación
Utilizar los datos en particiones: verdadero
Partición: Partición
Utilizar frecuencia: falso
Utilizar ponderación: falso
Niveles por debajo del raíz: 100
Modo: Experto
Número máximo de sustitutos: 5
Cambio mínimo en la impureza: 0,0
Medida de impureza para objetivos categóricos: Gini
Criterios de parada: Utilizar porcentaje
Número mínimo de registros en rama parental (%): 2
Número mínimo de registros en rama filial (%): 1
Podar árbol: verdadero
Utilizar regla de error típico: falso
Probabilidades previas: Basadas en datos de entrenamiento
Corregir previas por costes de clasificación errónea: falso
Utilizar costes de clasificación errónea: falso