UNIVERSITAT POLITÈCNICA DE VALÈNCIA

etsinf

Escola Tècnica Superior d'Enginyeria Informàtica

Escola Tècnica Superior d'Enginyeria

Informàtica

Universitat Politècnica de València

Estudio de la mortalidad en la Comunidad Valenciana mediante técnicas de Inteligencia de Negocio y Minería de Datos

Trabajo Fin de Grado

Grado en Ingeniería Informática

Autor: Jorge Pujadas Muñoz

Tutor: María José Ramírez Quintana

2018-2019

Resumen

El objetivo principal de este estudio es predecir el número de fallecimientos en un año, para las principales causas de mortalidad en la comunidad valenciana.

Para conseguir esto, se ha partido de unos datos públicos que recoge la distribución de las defunciones por municipio, sexo, causa y grupo de edad. A estos datos se les han aplicado una serie de tareas de minería de datos para conseguir unos modelos predictivos.

Las fases de preparación, modelado y evaluación se han realizado con RapidMiner.

Palabras clave: “Minería de datos”, “Datos abiertos”, “Causas mortalidad”, “RapidMiner”.

Abstract

The main objective of this study is to predict the number of deaths in a year, for the main causes of mortality in the Comunidad Valenciana.

To achieve this, it has been based on public data that includes the distribution of deaths by municipality, sex, cause and age group. A series of data mining tasks have been applied to these data to achieve predictive models.

The preparation, modeling and evaluation phases have been carried out with RapidMiner.

Keywords: “Data mining”, “Open data”, “Causes mortality”, “RapidMiner”

Tabla de contenidos

1. Introducción

1.1. Planteamiento inicial

Motivación

La idea de este estudio surge del cruce de caminos de 3 aspectos. Por un lado, la formación recibida en minería de datos con las asignaturas de “Sistemas de información estratégicos” y “Almacenes de datos y minería de datos”, la experiencia en el sector sanitario obtenida en el mundo laboral y la existencia de un portal con datos abiertos de la comunidad valenciana.

Los estudios basados en la recopilación de grandes cantidades de datos en el ámbito medico/sanitario han aumentado en gran medida en los últimos años, en parte empujados por las nuevas técnicas de minería de datos, data science, big data etc.

Los datos y la obtención de conocimiento de estos son hoy en día una herramienta más para la optimización de los procesos médicos, la obtención de diagnósticos y la mejora de la salud de la población. Viendo esto pensamos que puede ser muy útil este estudio ya sea por los resultados obtenidos, como para servir de base a futuros estudios.

Objetivos

Núcleo del objetivo: Dadas las características de un municipio de la comunidad valenciana y de su población, predecir el número de fallecidos en un año para cada causa de mortalidad.

Desarrollo de la idea: Una vez obtenida el valor numérico de la predicción, utilizar estos datos como palanca de mejora en las medidas orientadas a la mejora de la salud de la población y un mejor aprovechamiento de los recursos de personal/económico.

Preguntas a las que buscaremos dar respuesta

- ¿Existe un cambio en las causas de la mortalidad en los últimos 10 años?
- ¿Qué características son más influyentes?
- ¿Es posible a día de hoy predecir cuál será el escenario en los próximos 5 años?

Resultados esperados

De entre los resultados esperados esperamos, primero, responder a las preguntas planteadas en el apartado de objetivos. Esto se correspondería con los resultados a

nivel funcional, alguien que no quiere este estudio como ejemplo a una solución técnica de un problema, si no alguien a quien le interesa obtener conocimiento sobre el tema tratado.

Y segundo, mostrar cómo mediante técnicas de minería de datos se puede obtener conocimiento y conclusiones con unos datos iniciales en los cuales estos no aparecen. Esta solución servirá tanto para este caso como para base para otros casos con características parecidas.

1.2. Recorrido necesario

Como se va a desarrollar

Para lograr los objetivos y las preguntas a las que queremos dar respuesta, el camino por el cual vamos a pasar comienza con la toma de los datos obtenidos en el portal, con los cuales empezar un proceso de preparación y mejora de los datos, para pasar a la utilización de técnicas de minería de datos para obtener un modelo predictivo y comprobar sus resultados con los esperados.

Riesgos esperados y retos a cumplir

El principal riesgo encontrado en este estudio es si los datos serán suficientes, en cuanto a cantidad y detalle, para obtener modelos de predicción que aporten un gran valor. Este riesgo estará presente y se tendrá en cuenta a lo largo del resto de capítulos. A pesar de esto, se ha decidido seguir adelante puesto que este tipo de datos con información médica no es accesible desde fuera del ámbito sanitario. Recordemos que los datos facilitados tienen como origen un portal público de datos abiertos.

Todo esto no hace más que convertir a la realización de este estudio en un reto en el cual solventar esta desventaja mediante una buena e imaginativa fase de preparación de los datos. Para obtener un conjunto de datos lo suficientemente completo será necesario ampliar la información más allá de la registrada en los ficheros y ajustar las tareas de minería de datos para poder sacar todo el conocimiento posible.

Otro reto más es ampliar el número de estudios al respecto existentes. Los encontrados en la actualidad no utilizan técnicas de minería de datos explícitamente o están reservados al ámbito sanitario.

Metodología

Este estudio no sigue las fases de un desarrollo de software clásico, no se va a realizar un desarrollo y un entregable, vamos a intentar sacar unas conclusiones que aporten valor basándonos en los datos obtenidos.

Por lo que vamos a utilizar las fases requeridas con una tarea de minería de datos que es la cual vamos a abarcar.

Estas fases están recogidas en la metodología **CRISP-DM** que es sobre la cual nos vamos a basar para realizar la parte de tareas de minería de datos dentro del estudio. Y las fases son las siguientes:

- Comprensión del negocio o problema
- Comprensión de los datos
- Preparación de los datos
- Modelado
- Evaluación
- Implementación

En el capítulo correspondiente a la metodología se desglosará cada fase y cómo se ha aplicado en este estudio.

1.3. Estructura del estudio

Este estudio se ha estructurado en 9 capítulos, a través de estos se mostrará el recorrido por el cual hemos pasado desde la idea inicial, su planteamiento, preparación de los datos, modelos obtenidos y las conclusiones finales.

Capítulo 2 – Trabajos relacionados

En este capítulo se estudiarán los trabajos relacionados que existen actualmente, el tema de este TFG no es nuevo, pero sí el enfoque que se le dará. Esto quiere decir que ya existe información al respecto para otros ámbitos, como por ejemplo a nivel territorial, europeo y mundial, y con éstos será con los que se comparará el resultado del análisis predictivo y/o descriptivo.

También se comparará con las tendencias pasadas y futuras para situar el estado del arte en referencia al tiempo. Si, por ejemplo, las causas de mortalidad de años atrás se han superado, o si se prevé que en el futuro sean otras las dominantes.

Capítulo 3 – Metodología

Para el capítulo referido a la metodología que se ha seguido, vamos a explicar en qué consiste una metodología aplicada a un estudio de estas características, y cuáles son las diferencias con un desarrollo de software tradicional.

Capítulo 4 – Herramientas de análisis de datos

Para realizar el análisis predictivo y descriptivo se ha elegido la aplicación RapidMiner en su versión 9.3.

En este capítulo se explican los principales motivos para su elección, así como sus características principales

Capítulo 5 – Datos a analizar

Este punto es muy importante, puesto que el descubrir que existían datos abiertos sobre la mortalidad en la comunidad valenciana, se desarrolló la idea de poder utilizarlos para un estudio utilizando técnicas de minería de datos.

Estos datos están accesibles desde el portal www.dadesobertes.gva.es (Portal de Transparencia de la Generalitat Valenciana).

Durante este capítulo se desgranarán las distintas características de los datos, las transformaciones que se van a realizar y los datos añadidos para enriquecer el modelo.

Capítulo 6 – Análisis predictivo y descriptivo – Fase de preparación

El sexto capítulo junto con el séptimo, forman parte del desarrollo técnico y los que nos van a dar unos resultados sobre los que basarnos. En este se analizarán los datos y se realizarán tareas de preparación de los mismos.

Capítulo 7 – Análisis predictivo y descriptivo – Fase de modelización y evaluación

Una vez tengamos los datos preparados, se aplicarán distintas técnicas de minería de datos para obtener y analizar los resultados.

Se describirán, justificaran y evaluaran las diferentes técnicas empleadas para darle validez a los datos y poder comprarlas entre ellas.

Capítulo 8 – Conclusiones

Tras hacer el recorrido por el resto de capítulos, en este se expondrán las conclusiones a las cuales se ha llegado tras analizar los datos.

Tomando como referencia los objetivos iniciales y los puntos clave encontrados en otros trabajos relacionados

Capítulo 9 – Bibliografía

En el último capítulo nos limitaremos a enumerar las referencias bibliográficas y en qué parte se han utilizado.

2. Contexto y Trabajos Relacionados

Antes de avanzar con el estudio, vamos a situarnos en el contexto, para entender la evolución de las causas de mortandad a lo largo de la historia y qué se espera en el futuro.

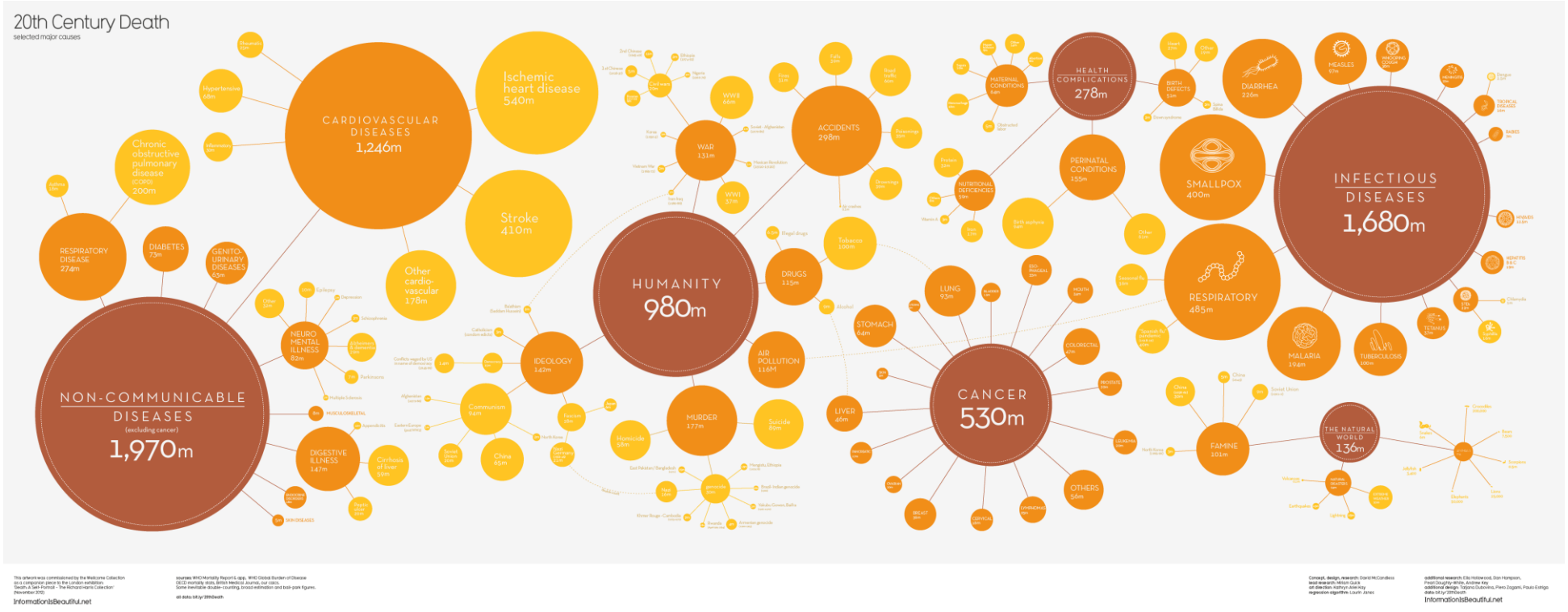
En este capítulo no se va a realizar una búsqueda exhaustiva del tema, puesto que no es lo que se pretende, por lo que las fuentes y formatos de la información mostrada son diversas. Aun así, se va a mostrar de forma resumida y lo más clara posible un tema tan amplio.

Para ello vamos a resumir las principales características de 6 escenarios que hemos creído más importantes y son los siguientes:

2.1. Pasado (Siglo XX)

El primer escenario es el estudio de las principales causas de defunción en el siglo XX. Esto nos servirá como punto de inicio viendo cómo ha evolucionado la sociedad en este aspecto.

En esta gráfica vienen representadas estas causas en millones y agrupadas por similitud o relación.



Fuente: https://roulive.com/category/medical/

Si observamos su distribución y agrupación podemos sacar 5 grandes grupos y sus principales causas [1]:

- Enfermedades no transmisibles
  - Enfermedades respiratorias
  - Diabetes
  - Enfermedades genitourinarias
  - Enfermedad neuro mental
  - Enfermedad digestiva
- Enfermedades cardiovasculares
  - Enfermedad isquémica del corazón
  - Accidente cerebrovascular
- Humanidad
  - Guerra
  - Accidentes
  - Drogas
  - Polución
  - Asesinato
  - Ideología
- Cáncer
- Enfermedades infecciosas
  - Tuberculosis
  - Malaria
  - Respiratorias
  - Viruela
  - Diarrea

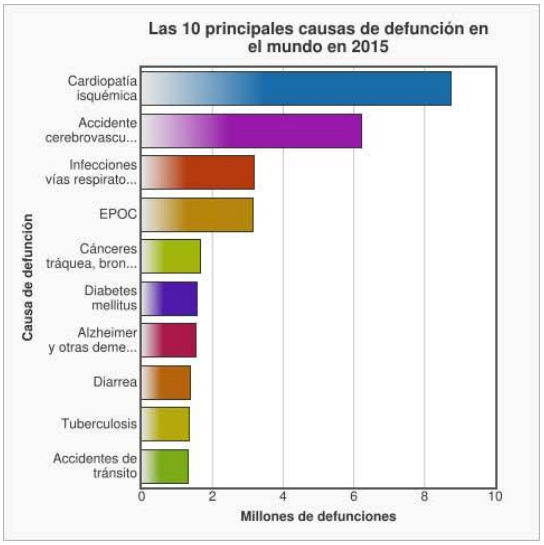
2.2. Actualidad a nivel mundial

El segundo escenario es el estudio de las principales causas de defunción en la actualidad (datos tomados en 2015) en todo el mundo. Antes de bajar al nivel de detalle de una comunidad autónoma en España, es bueno tomar de referencia a toda la población mundial.

Según un estudio de la Organización Mundial de la Salud, estas son las 10 principales causas de muerte registradas en 2015 ordenadas de mayor a menor [2]:

1. Cardiopatías isquémicas
2. Accidentes cerebrovasculares
3. Infecciones de vías respiratorias
4. Enfermedad pulmonar obstructiva crónica
5. Tumores de pulmón, tráquea y bronquios
6. Diabetes
7. Alzheimer
8. Diarrea
9. Tuberculosis

10. Accidentes de tráfico



Fuente: https://www.phmk.es/oms-ranking-10-principales-causas-de-muerte/

2.3. Actualidad en Europa

Desde la unión europea se están desarrollando planes para detectar y tratar las principales causas de muerte derivadas de enfermedades crónicas, como el recogido en el siguiente enlace:

https://ec.europa.eu/health/non_communicable_diseases/overview_es

Para el escenario europeo vamos a seguir el mismo esquema que en el apartado anterior, mostrar las 10 causas más importantes [3]:

1. Cardiopatías isquémicas
2. Accidentes cerebrovasculares
3. Tumores de pulmón, tráquea y bronquios
4. Infecciones de vías respiratorias
5. Accidentes
6. Cáncer de colon
7. Neumonía
8. Diabetes
9. Cáncer de páncreas
10. Cáncer de pecho

2.4. Actualidad en España

Para buscar las principales causas en el territorio español, vamos a utilizar los datos publicados por el Instituto Nacional de Estadística [4].

1. Enfermedades del sistema circulatorio
2. Tumores
3. Enfermedades del sistema respiratorio
4. Enfermedades del sistema nervioso y de los órganos de los sentidos
5. Trastornos mentales y del comportamiento
6. Enfermedades del sistema digestivo
7. Causas externas de mortalidad
8. Enfermedades endocrinas, nutricionales y metabólicas
9. Enfermedades del sistema genitourinario
10. Síntomas, signos y hallazgos anormales clínicos y de laboratorio

Después de ver las causas en los tres principales ámbitos como son nivel mundial, Europa y España, podemos destacar que en los tres la principal causa de mortalidad está relacionada con las cardiopatías.

Destacar que a nivel mundial siguen apareciendo causas como diarrea y tuberculosis las cuales en Europa y España ocupan posiciones muy bajas. En cambio, en estos dos ámbitos aparecen causas como el cáncer y los accidentes (causas externas no derivadas de enfermedad) que están relacionadas con una mayor esperanza de vida que no se da a nivel mundial.

2.5. Predicción

Existen pruebas médicas especificas destinadas a predecir porcentualmente el riesgo de muerte sobre una causa en concreto. Como por ejemplo en el artículo “Causas de muerte y predicción de mortalidad en la EPOC” [5], pero estas no se hacen servir de técnicas de minería de datos para obtener sus resultados.

En cambio, existen aplicaciones como

https://flowingdata.com/2016/01/19/how-you-will-die/

que sí utilizan técnicas de minería de datos (Según el creador de la web, la preparación de los datos y el modelo están creados en R) para obtener la edad y causa de la muerte de un usuario según los datos introducidos en cuanto a edad, sexo y raza.

Otro ejemplo sería https://population.io, que, introduciendo fecha de nacimiento, país y sexo, predice el tiempo de vida. No especifica que técnica utilizan.

En nuestro estudio desarrollaremos un modelo predictivo con el cual intentaremos representar la evolución en cuanto a las principales causas en la CV y compararla con los resultados en España, la UE y a nivel mundial, así como viendo las previsiones que se han descrito para el futuro en otros artículos, como de cerca o lejos están nuestros resultados.

2.6. Futuro

Basándose en los estudios actuales, ya sean con técnicas tradicionales o de minería de datos, las causas más comunes en Europa en 2030 serán [6]:

1. Cardiopatías isquémicas
2. Accidentes cerebrovasculares
3. Cáncer de pulmón
4. Cáncer de colon
5. Alzheimer

Con todo esto queremos destacar que no existe un número suficiente de artículos que describan un proceso de predicción sobre las causas de mortalidad ni cómo se ha realizado este proceso.

3. Metodología empleada

En un desarrollo de software clásico, si se implementan los procesos correctos y se siguen las pruebas necesarias, se obtiene un producto que cumplirá con los requisitos iniciales. En un proceso de extracción de conocimiento o "Knowledge Discovery in Databases", como es el que nos ocupa, esto no está asegurado. Existe una incertidumbre en cuanto a los resultados obtenidos a pesar de partir de una premisa válida, realizar todos los pasos correctamente y tener claro cuál es el objetivo a buscar, ya que todo el proceso depende de la cantidad y calidad de los datos.

Esto hace de los procesos de desarrollo orientados a la extracción de conocimiento diferentes a lo que tradicionalmente se ha realizado y estudiado en cuanto a desarrollo de software, es por esto que se necesita un tipo de metodología distinto para abordar un estudio como el presente.

3.1. El proceso de Extracción de conocimiento

El modelo KDD

KDD (del inglés, *Knowledge Discovery in Databases*) es el proceso de extraer conocimiento útil y novedoso a partir de una base de datos. Este proceso consta de varias etapas que se realizan de forma secuencial, aunque en todas ellas podemos volver hacia atrás por lo que el proceso es también iterativo. Más concretamente, se distinguen las siguientes seis etapas [7]:

- "*Selección*
  - *Consiste en la integración de diferentes fuentes de datos en una misma base de datos (puede ser en forma de datawarehouse) o sistema de ficheros.*
- *Procesamiento previo*
  - *Los datos integrados deben de ser tratados antes de realizar el proceso de minería de datos. Debe realizarse una selección de aquellos datos que van a utilizarse.*
- *Transformación*
  - *Sobre ese subconjunto de datos hay que realizar un proceso de limpieza y transformado para dejarlos en condiciones de ser tratados en fases posteriores. El objetivo de esta fase es obtener una vista minable para la fase siguiente.*
- *Minería de Datos*
  - *Es considerada la fase más importante del proceso de KDD, se define como el proceso de exploración y análisis, por medios automáticos o*

semiautomáticos, de los datos existentes en la vista minable obtenida en la fase anterior, con el fin de descubrir patrones/modelos significativos y reglas. El resultado de la fase son los patrones/modelos de ese análisis.

- *Interpretación / Evaluación*
  - *El primer paso de esta fase es la evaluación de los patrones y modelos obtenidos, ya que, antes de ser interpretados para la obtención de conocimiento, debe de comprobarse que tienen la calidad suficiente para poder realizar la interpretación.*
- *Despliegue*
  - *Una vez obtenidos unos modelos y se ha comprobado que se corresponden con el objetivo buscado, aplicarlos sobre el contexto en el que se realizan para obtener conocimiento."*

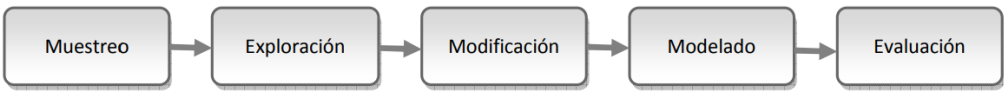
Como puede observarse, la minería de datos es una de las etapas del KDD, en concreto la etapa de creación de los modelos o patrones. Quizás, como estos modelos son el conocimiento buscado, a menudo al proceso de KDD se le suele denominar Minería de Datos, en un intento de enfatizar la importancia de esta etapa.

3.2. Metodologías actuales y sus características

Dicho lo cual, desde el punto de vista de la metodología, existe un gran número de modelos ampliamente aceptados y con múltiples proyectos realizados. De entre todos vamos a destacar 2 que son los más utilizados o sobre los cuales se han basado el resto.

La metodología SEMMA

"Desarrollada por SAS, el nombre de esta metodología corresponde al acrónimo de las cinco fases básicas del proceso: Sample (Muestreo), Explore (Exploración), Modify (Modificación), Model (Modelado) y Assess (Valoración)." [8]



Fuente: [8]

Desarrollo de las fases [26]:

1. Muestreo: El proceso comienza con el muestreo de datos, por ejemplo, seleccionando el conjunto de datos para el modelado.
2. Exploración: Comprensión de los datos mediante el descubrimiento de relaciones con la ayuda de la visualización de datos.

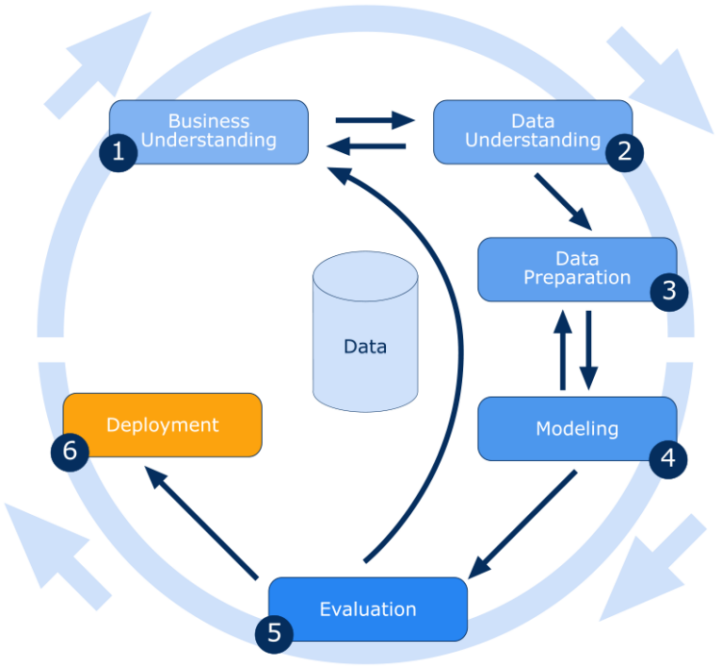
3. Modificación: Contiene métodos para seleccionar, crear y transformar variables en preparación para el modelado de datos.
4. Modelado: El enfoque se centra en aplicar varias técnicas de modelado en las variables preparadas para crear modelos que posiblemente proporcionen el resultado deseado.
5. Evaluación: La evaluación de los resultados del modelado muestra la fiabilidad y la utilidad de los modelos creados.

Resaltar que en esta metodología la etapa de Selección y comprensión del problema son excluidas del ciclo, esto es debido a que está más orientada a aspectos técnicos. Esta característica es debida a que fue principalmente creado para la utilización del software de Explotación de Datos de la compañía SAS.

La metodología CRISP-DM

Un grupo de empresas europeas (entre las cuales destacamos SPSS, Teradata, NCR, AG) proponen, basándose en diferentes versiones de KDD, el desarrollo de una guía de referencia de libre distribución denominada CRISP-DM (Cross Industry Standard Process for Data Mining). Este modelo se basa en 6 fases [27]:

1. “Comprensión del negocio: Convierte este conocimiento de los datos en la definición de un problema de minería de datos
2. Comprensión de los datos: Colección de datos inicial y continúa con las actividades que permiten familiarizarse con los datos, identificar los problemas de calidad, descubrir conocimiento preliminar sobre los datos.
3. Preparación de datos: Cubre todas las actividades necesarias para construir el conjunto final de datos a partir de los datos en bruto iniciales.
4. Modelado: Se seleccionan y aplican las técnicas de modelado que sean pertinentes al problema, y se calibran sus parámetros a valores óptimos.
5. Evaluación: Antes de proceder al despliegue final del modelo, es importante evaluarlo a fondo y revisar los pasos ejecutados para crearlo, comparar el modelo obtenido con los objetivos de negocio.
6. Implementación: El conocimiento obtenido tendrá que organizarse y presentarse para que el cliente pueda usarlo.”



Fuente: [10]

3.3. Selección del proceso a seguir

Una vez mostradas las características principales del proceso de KDD y de las dos metodologías, vamos a compararlas, seleccionar una, y dar los motivos para esta elección.

Deseamos una metodología/proceso que sea de amplio uso entre profesionales por lo cual recurrimos a un análisis de https://www.kdnuggets.com/ en el que se puede ver la preferencia de 200 usuarios en el uso de una metodología de minería de datos.

What main methodology are you using for your analytics, data mining, or data science projects ? [200 votes total] 2014 poll 2007 poll	
CRISP-DM (86)	43% 42%
My own (55)	27.5% 19%
SEMMA (17)	8.5% 13%
Other, not domain-specific (16)	8% 4%
KDD Process (15)	7.5% 7.3%
My organizations' (7)	3.5% 5.3%
A domain-specific methodology (4)	2% 4.7%
None (0)	0% 4.7%

Fuente [11]

Vemos que en primer lugar de preferencia se encuentra la metodología CRISP – DM con un 43%, en segundo lugar, SEMMA con un 8.5% y en tercer lugar KDD con 7.5%

Un aspecto en contra de SEMMA, es que no es una metodología independiente y depende del mantenimiento y actualizaciones que se hagan por parte de SAS. Y en cuanto a un hándicap sobre KDD, diremos que las metodologías actuales están mejor estructuradas para las tareas que se dan hoy en día.

Con todo esto vamos a seleccionar la metodología CRISP-DM para basarnos y tomar como referencia en la parte de proceso, modelado y validación de los datos. No queriendo decir que se vayan a seguir todos los pasos, puesto que puede que no sean necesarios.

3.4. Metodología CRISP-DM en detalle

Una vez comentadas las tres principales metodologías utilizadas en proyectos de minería de datos y seleccionada CRISP-DM como la elegida para tomar como referencia, vamos a pasar a detallar que ocurre dentro de cada fase.

**Definición y desglose tomado de los apuntes de la asignatura "Sistemas de información estratégicos" Tema 1 (Parte DM) [12]*

- ✓ "Comprensión del negocio
  - Entender los objetivos y requerimientos del proyecto desde una perspectiva de negocio. Sub-fases:
    - establecimiento de los objetivos de negocio (contexto inicial, objetivos y criterios de éxito)
    - evaluación de la situación (inventario de recursos, requerimientos, suposiciones y restricciones, riesgos y contingencias, terminología y costes y beneficios),
    - establecimiento de los objetivos de minería de datos (objetivos de minería de datos y criterios de éxito)
    - generación del plan del proyecto (plan del proyecto y evaluación inicial de herramientas y técnicas)
- ✓ Comprensión de los datos:
  - Recopilar y familiarizarse con los datos, identificar los problemas de calidad de datos y ver las primeras potencialidades o subconjuntos de datos que puede ser interesante analizar (según los objetivos de negocio en la fase anterior). Sub-fases:
    - recopilación inicial de datos (informe de recopilación)
    - descripción de datos (informe de descripción)
    - exploración de datos (informe de exploración)
    - verificación de calidad de datos (información de calidad)

- ✓ Preparación de los datos:
  - El objetivo de esta fase es obtener la "vista minable". Aquí se incluye la integración, selección, limpieza y transformación. Sub-fases:
    - selección de datos (razones de inclusión / exclusión)
    - limpieza de datos (informe de limpieza de datos)
    - construcción de datos (atributos derivados, registros generados)
    - integración de datos (datos mezclados) y
    - formateo de datos (datos reformateados)
- ✓ Modelado:
  - Es la aplicación de técnicas de modelado o de minería de datos propiamente dichas a las vistas minables anteriores. Sub-fases:
    - selección de la técnica de modelado (técnica de modelado, suposiciones de modelado),
    - diseño de la evaluación (diseño del test)
    - construcción del modelo (parámetros elegidos, modelos, descripción de los modelos)
    - evaluación del modelo (medidas del modelo, revisión de los parámetros elegidos)
- ✓ Evaluación:
  - Es necesario evaluar (desde el punto de vista de la finalidad) los modelos de la fase anterior. Es decir, si el modelo nos sirve para responder a algunos de los requerimientos del negocio. Sub-fases:
    - evaluación de resultados (evaluación de los resultados de minería de datos, modelos aprobados)
    - revisar el proceso (revisión del proceso)
    - establecimiento de los siguientes pasos (lista de posibles acciones, decisión)
- ✓ Despliegue:
  - Se trata de explotar la potencialidad de los modelos, integrarlos en los procesos de toma de decisión de la organización, difundir informes sobre el conocimiento extraído, etc. Sub-fases:
    - planificación del despliegue (plan del despliegue),
    - planificación de la monitorización y del mantenimiento (plan de la monitorización y del despliegue)
    - generación del informe final (informe final, presentación final)
    - revisión del proyecto (documentación de la experiencia)"

Sobre las fases y los puntos que se han desglosado, en el siguiente capítulo donde se pasara a la parte técnica del estudio, se comentara en cada paso a qué fase y sub-fase corresponde. Recordemos que no todas las fases y/o sub-fases tienen porque aparecer en un estudio de estas características.

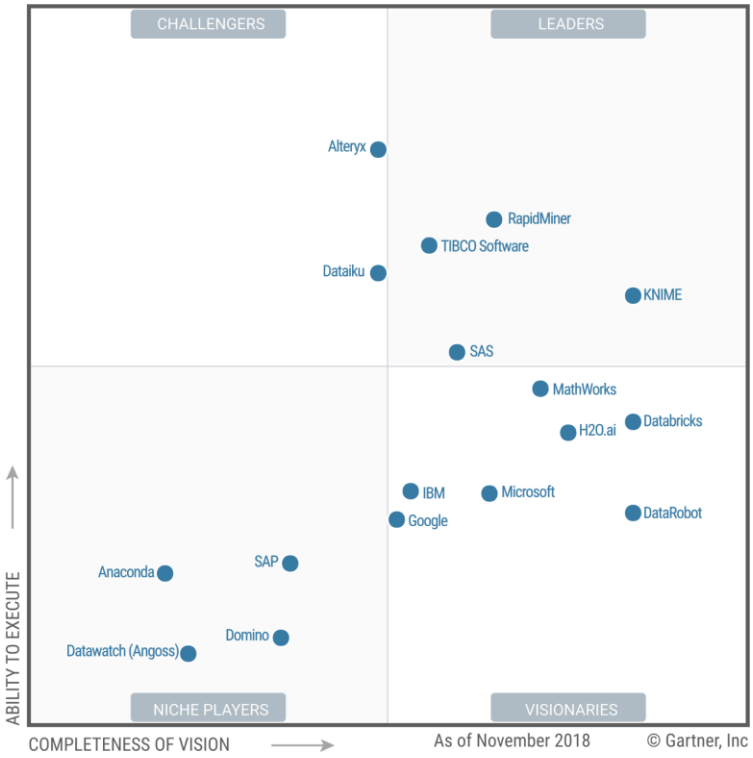
4. Herramientas de análisis de datos

Para realizar el estudio se van a utilizar principalmente los programas Excel, Notepad++ y RapidMiner. Para los dos primeros no vamos a entrar en detalles y explicaciones puesto que son programas ampliamente conocidos y/o fáciles de usar. Este capítulo se va a centrar en los motivos y principales características que nos ha llevado a la elección de RapidMiner sobre otras herramientas de minería de datos.

Motivos

La primera toma de contacto con RapidMiner se produce en la asignatura de "Sistemas de información estratégicos" durante el curso 2017/18. En esta, se utiliza la versión 5.3 en los laboratorios de prácticas. Al realizar los ejercicios, nos damos cuenta que posee una interface más intuitiva y ágil que el programa que conocíamos anteriormente en este ámbito, que era Weka [28].

Buscando información sobre la aplicación, nos damos cuenta de que está muy bien valorada, está siendo utilizada en muchos proyectos, tanto profesionales como académicos, como muestra el cuadrante de Gartner, una visualización de la posición que ocupan las herramientas de Data Science y Machine Learnig.



Fuente: Gartner (enero 2019) [13]

En cuanto a la importancia y relevancia de RapidMiner en el ámbito de la minería de datos y ciencia de datos, vemos que ocupa un lugar destacado entre los líderes en el cuadrante de Gartner para plataformas de Data Science.

Trabajan con más de 30.000 organizaciones en industrias como automoción, salud y comunicaciones entre otras. Y en múltiples casos de uso que van desde segmentación de usuarios, detección de fraude y optimización de precios.

Buscando para su descarga, vemos que en el 2019 esta publicada la versión 9.3 Studio, que mantiene la forma de trabajo de la versión 5.3, con la que estamos familiarizados, y además añade características nuevas. Nos decidimos por esta versión para realizar las tareas técnicas del estudio.

Esta y otras versiones existentes se pueden descargar desde:

https://my.rapidminer.com/nexus/account/index.html#downloads

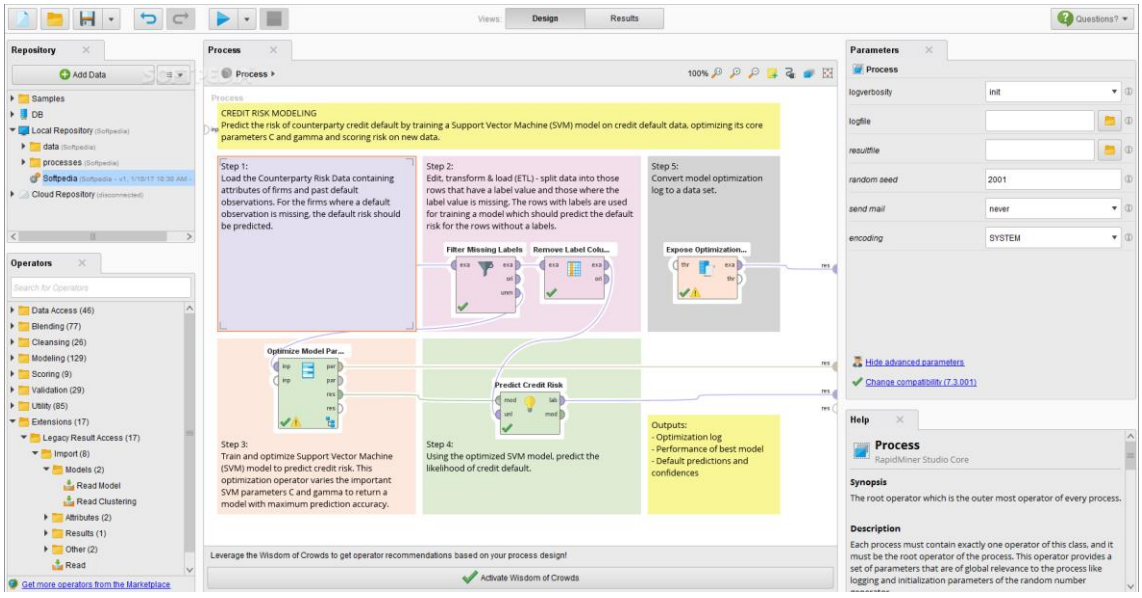
Antes de entrar en detalles sobre sus características y versiones disponibles vamos a indicar la definición disponible en Wikipedia [14]:

"*RapidMiner* *es un programa informático para el análisis y minería de datos. Permite el desarrollo de procesos de análisis de datos mediante el encadenamiento de operadores a través de un entorno gráfico. Se usa en investigación, educación, capacitación, creación rápida de prototipos y en aplicaciones empresariales.*

La versión inicial fue desarrollada por el departamento de inteligencia artificial de la Universidad de Dortmund en 2001. Se distribuye bajo licencia AGPL y está hospedado en SourceForge desde el 2004.

RapidMiner proporciona más de 500 operadores orientados al análisis de datos, incluyendo los necesarios para realizar operaciones de entrada y salida, preprocesamiento de datos y visualización."

Ejemplo del entorno de trabajo en RapidMiner 9.3:



4.1. RapidMiner – Versiones disponibles

Entre las versiones que podemos encontrar están:

RapidMiner Studio 9.3

Versión utilizada en este estudio, se centra en la preparación de datos, creación de modelos, y validación de modelos. Es de carácter local y el punto de partida para el resto de versiones.

RapidMiner Server 9.3

Dedicado a dar un lugar donde compartir y re-utilizar modelos predictivos, procesos automáticos y desplegar modelos a través de diferentes entornos. Ofrece programación, seguridad y escalado.

RapidMiner Server in the Cloud

Imágenes de RapidMiner Server disponibles para desplegar en Amazon AWS y Microsoft Azure

RapidMiner Radoop 9.3

Extensión de RapidMiner que puede ser instalada en la versión estudio y Server. Añade procesamiento y análisis de Big Data con un cluster de Hadoop. Una vez instalado también se pueden utilizar características de Hive y Spark.

4.2. RapidMiner – Sistema de licencias

RapidMiner no es un software gratuito, existen distintas licencias con los cuales se puede utilizar.

Free

Sin coste ni fecha de caducidad, pero las limitaciones vienen en las funcionalidades pasados el mes de prueba. Se limita a 10.000 filas las usadas en los procesos de preparación y modelado. Límite de un procesador lógico en la ejecución de las tareas.

Educational

Sin coste, pero con licencia de utilización valida durante un año. No existen las limitaciones en cuanto número de filas o procesadores de la versión Free. Es necesario registrase con una cuenta de correo que pertenezca a una universidad, en este caso la UPV, y describir en pocas palabras la práctica que se va a realizar con el programa.

Nota: Esta es la licencia utilizada durante el desarrollo del estudio.

Empresarial

Para la utilización con fines comerciales, es necesario adquirir una licencia empresarial. Las mostradas en la web son:

- Professional – 5.500$ al año por usuario. Todas las funcionalidades, pero limitaciones en cuanto a filas y procesadores.
- Enterprise – 11.000$ al año por usuario. Todas las funcionalidades sin limitaciones en cuanto a filas y procesadores.

4.3. Ayudas a la creación de modelos

Entre las facilidades que incorpora a la hora del procesado de datos y creación de modelos, existen múltiples funciones y paquetes ya creados que se pueden utilizar sin llegar a escribir líneas de código y en pocos minutos.

Destacamos por encima los modos “Turbo Prep” y “Auto Model” que realizan las siguientes funciones:

Turbo Prep

Permite trabajar directamente sobre una vista de los datos donde crear visualizaciones, limpieza de datos, combinar, crear atributos, etc. Creando un historial de cambios sobre los cuales se puede volver atrás.

Auto Model

Según reza su eslogan, crea modelos predictivos en pocos clics. Incluye elementos de Turbo Prep y además analiza automáticamente los datos para identificar problemas, elegir la técnica más propicia y representa un modelo de forma aplicable a un simulador de decisiones. Estos procesos no son cajas negras totalmente, se puede editan y mostrar el proceso completo.

Nota: Estas funcionalidades no van a ser utilizadas como método para realizar el presente estudio, puesto que no existían cuando se recibió la formación en los laboratorios, y encasilla el proceso y lo hace menos flexible.

5. Análisis exploratorio de los datos

Antes de empezar el estudio, vamos a describir los datos con los cuales vamos a trabajar. En un estudio de estas características, los datos son el elemento central y sobre el cual giran el resto de aspectos.

Tanto es así, que una de las razones por las cuales elegimos el tema del estudio y hacia dónde dirigirlo es precisamente debido a que encontramos los datos que ahora vamos a describir.

El origen de estos datos es el Portal de Transparencia de la Generalitat Valenciana (http://www.dadesobertes.gva.es). En él se encuentran un gran número de datos disponibles para todos los usuarios que quieran acceder. A su vez, el portal de Dades Obertes está incluido en el Portal de Transparencia de la Generalitat Valenciana (http://www.gvaoberta.gva.es/es).

Temas con conjuntos de datos:

- Urbanismo e infraestructura
- Educación
- Medio Ambiente
- Demografía
- Turismo
- Legislación y justicia
- Salud
- Sociedad y bienestar
- Economía
- Empleo
- Sector publico
- Industria
- Energía

Estos datos tienen como característica en común el ser datos abiertos u Open Data

¿Qué es Open Data? –Definición de Dades Obertes [15]

> *"Open Data es una filosofía y práctica a nivel global cuyo objetivo es poner a disposición de los ciudadanos el mayor volumen de información posible, especialmente aquella información perteneciente a las administraciones públicas, respetando los límites establecidos por la legalidad vigente (LOPD, secreto estadístico, etc…)*
>
> *La información publicada se ofrece adicionalmente siguiendo estándares abiertos y reutilizables, lo cual posibilita que tanto ciudadanos, como empresas o instituciones de cualquier índole puedan consultarla, o incluso crear a partir de la misma nuevas aplicaciones o servicios.*
>
> *Esta filosofía se enmarca dentro de la demanda creciente por parte de la sociedad de una mayor transparencia por parte de todas las Administraciones Públicas, razón por la cual es una de las principales iniciativas tenida en cuenta*

en cualquier política de Gobierno Abierto. A su vez, gracias al creciente desarrollo de las nuevas tecnologías, que cada vez más facilitan el manejo y tratamiento automatizado de cantidades ingentes de información, la publicación de datos está alcanzando una notable popularidad por constituirse en una herramienta real de transparencia y difusión.”

Este tipo de portales nos permiten acceder a conjuntos de datos muy valiosos con los cuales hemos podido realizar este estudio, pero también tienen su contraposición.

Al tener que cumplir con la legalidad vigente en materia de anonimización, los datos que tenemos no tienen todo el detalle original.

¿Qué quiere decir esto?

- Los datos pueden no tener el valor exacto original, se han sustituido por su equivalente dentro de un rango.
- Los datos están agrupados o resumidos al nivel de detalle resultante del punto anterior.
- Los datos puede que hayan pasado por un proceso de limpieza o preparación previo sobre el cual no se tiene conocimiento del estado anterior.

Esto es un **riesgo muy a tener en cuenta** cuando se utilizan este tipo de datos, ya sea para un estudio o para fines comerciales. Esta situación ya se comenta en el capítulo 1 y será un aspecto a tener en cuenta a lo largo del estudio.

De entre los temas mencionados que están disponibles en el portal Dades Obertes, hemos seleccionado los correspondientes al apartado de salud.

En este apartado se pueden encontrar 12 conjuntos de datos, que se corresponden con datos de mortalidad, 10 de ellos, y datos de centros y servicios sanitarios, 2 de ellos. Los utilizados en este estudio se corresponden a los ficheros con información sobre la mortalidad por cada causa.

5.1. Descripción de los datos utilizados

Resumen de características de los datos escogidos

- Fuente: http://www.dadesobertes.gva.es
- Descripción: Datos sobre la mortalidad en los años 2007-2016.
- Formatos (csv, json, xml)
- Tipos de causas:
    - CV (86 causas, lista abreviada de la Comunidad Valenciana)
    - INE-CCAA (102 causas, lista abreviada del INE-CCAA)

  - CIE10 (códigos de 3 dígitos, jerarquía de 3 niveles)
- Estructura sanitaria departamental en la comunidad valenciana (Hospital/Centro de atención en la localidad/zona)

Estos datos se pueden descargar directamente desde la web. También existe una API con la cual acceder a los datos para su consulta, este método es cada vez más utilizado para aplicaciones de Big Data, Data Science, Data Near-Real-Time, etc.

Descripción del dato

Una vez descargados los datos, este es el aspecto que tienen sin realizar ninguna transformación.

- Vista fichero plano:

```
ANYO;COD_PROV;NOM_PROV;COD_MUN;NOM_MUN;COD_CAUSA;DESC_CAUSA;COD_GENERO;DESC_GENERO;RANGO_EDAD;NUM_FALLECIDOS
2016;03;ALICANTE;009;ALCOY;001;ENFERMEDADES INFECCIOSAS INTESTINALES;H;HOMBRE;80 - 84;1
2016;03;ALICANTE;009;ALCOY;004;SEPTICEMIA;H;HOMBRE;80 - 84;2
2016;03;ALICANTE;009;ALCOY;004;SEPTICEMIA;M;MUJER;> 84;6
2016;03;ALICANTE;009;ALCOY;004;SEPTICEMIA;M;MUJER;75 - 79;1
2016;03;ALICANTE;009;ALCOY;004;SEPTICEMIA;M;MUJER;80 - 84;3
2016;03;ALICANTE;009;ALCOY;005;HEPATITIS VÍRICA;M;MUJER;> 84;2
2016;03;ALICANTE;009;ALCOY;005;HEPATITIS VÍRICA;M;MUJER;70 - 74;1
2016;03;ALICANTE;009;ALCOY;006;SIDA;H;HOMBRE;30 - 34;1
2016;03;ALICANTE;009;ALCOY;006;SIDA;H;HOMBRE;50 - 54;1
2016;03;ALICANTE;009;ALCOY;009;TUMOR MALIGNO DEL LABIO, DE LA CAVIDAD BUCAL Y DE LA FARINGE;H;HOMBRE;65 - 69;1
```

Como se puede observar con este ejemplo, los datos descargados figuran en formato .CSV (*comma-separated values*) con la primera línea con los títulos de los campos, y las siguientes con los registros.

- Vista Excel

	A	B	C	D	E	F	G	H	I	J	K
1	ANYO	COD_PROV	NOM_PROV	COD_MUN	NOM_MUN	COD_CAUSA	DESC_CAUSA	COD_GENERO	DESC_GENERO	RANGO_EDAD	NUM_FALLECIDOS
2	2016	3	ALICANTE	9	ALCOY	1	ENFERMEDADES INFECCIOSAS INTESTINALES	H	HOMBRE	80 - 84	1
3	2016	3	ALICANTE	9	ALCOY	4	SEPTICEMIA	H	HOMBRE	80 - 84	2
4	2016	3	ALICANTE	9	ALCOY	4	SEPTICEMIA	M	MUJER	> 84	6
5	2016	3	ALICANTE	9	ALCOY	4	SEPTICEMIA	M	MUJER	75 - 79	1
6	2016	3	ALICANTE	9	ALCOY	4	SEPTICEMIA	M	MUJER	80 - 84	3
7	2016	3	ALICANTE	9	ALCOY	5	HEPATITIS VÍRICA	M	MUJER	> 84	2
8	2016	3	ALICANTE	9	ALCOY	5	HEPATITIS VÍRICA	M	MUJER	70 - 74	1
9	2016	3	ALICANTE	9	ALCOY	6	SIDA	H	HOMBRE	30 - 34	1
10	2016	3	ALICANTE	9	ALCOY	6	SIDA	H	HOMBRE	50 - 54	1
11	2016	3	ALICANTE	9	ALCOY	9	TUMOR MALIGNO DEL LABIO, DE LA CAVIDAD BUC	H	HOMBRE	65 - 69	1

Para conocer mejor los datos con los que vamos a trabajar, vamos a mostrar un resumen con la información del metadato resultante.

Fichero		Causas CV	Causas INE	Causas CIE10
Campos	**Año**	10 Valores Distintos	10 Valores Distintos	10 Valores Distintos
	Provincia Cod + Nombre	3 Valores Distintos	3 Valores Distintos	*Sin valores
	Municipio Cod + Nombre	13 Valores Distintos	13 Valores Distintos	*Sin valores
	Causa Cod + Nombre	81 Valores Distintos	98 Valores Distintos	650 Valores Distintos
	Genero Cod + Nombre	2 Valores Distintos	2 Valores Distintos	2 Valores Distintos
	Rango Edad	16 Valores Distintos	16 Valores Distintos	16 Valores Distintos
	Numero Fallecidos	Indicador	Indicador	Indicador

Numero de filas en cada fichero por año y tipo de causa

Año	Causas CV	Causas INE	Causas CIE10
2007	27014	30713	41818
2008	26488	30896	40985
2009	27058	31124	42277
2010	27704	31599	41231
2011	26756	30497	42467
2012	27077	30611	40871
2013	27115	31295	41270
2014	27419	12764	41612
2015	11974	13131	7514
2016	11866	12999	7492

*Nota: Para los años 2015, 2016, solo hay registros para las combinaciones con datos, en el resto de año aparecen todas las combinaciones incluyendo los valores a 0

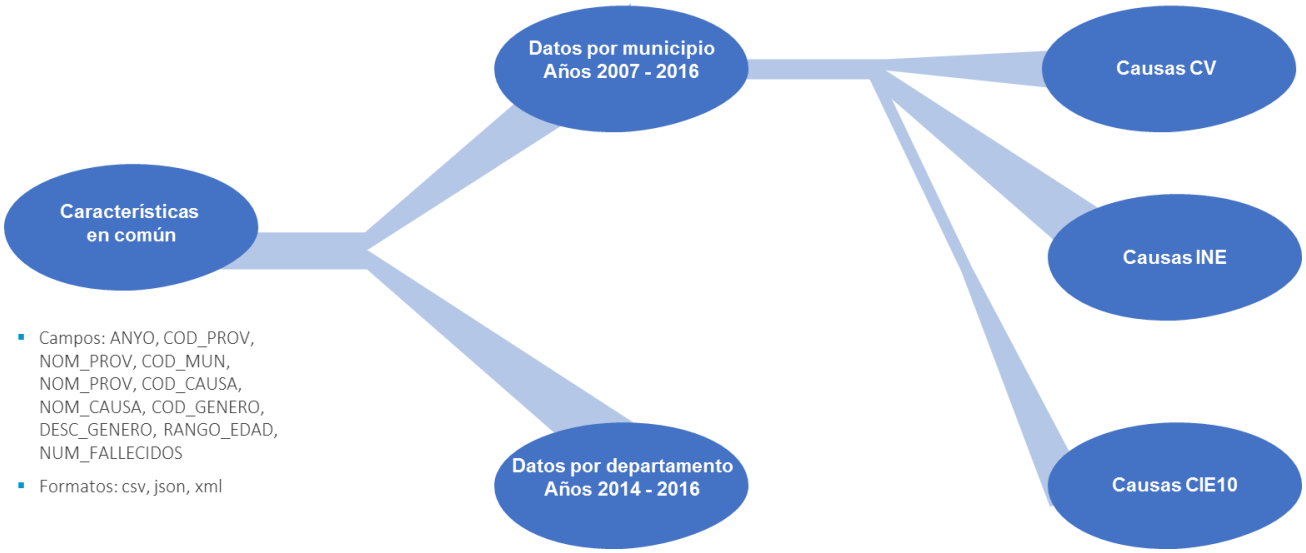
5.2. Datos seleccionados para el estudio

Una vez ya hemos descrito las características de los datos disponibles, vamos a desarrollar cuál ha sido el proceso para la selección de los que serán utilizados.

Existe un gran número de ficheros disponibles en la web de Dades Obertes' (45 en total) con datos sobre las causas de la mortalidad en la comunidad valenciana en los

últimos años, a continuación, vamos a describir las características de cada uno, indicar con cuales vamos a trabajar y los motivos.

Estos ficheros poseen características y contenido en común, por lo cual, para enumerar mejor el listado de ficheros, su relación y características en común o diferentes vamos a situarlos en el siguiente gráfico.



Ruta de decisiones

Tomando como referencia el grafico anterior, vamos a ir explicando qué conjuntos de datos se van a seleccionar para el estudio.

- **La primera bifurcación** nos da a elegir principalmente entre datos agrupados por municipio o datos agrupados por departamento de salud.

***Porque **SI** elegimos datos por municipio*

Seleccionar esta agrupación nos permite ubicar sobre el mapa los datos con los cuales estamos trabajando, esto es muy útil si queremos añadir información que inicialmente no está incluida en los conjuntos de datos.

Por ejemplo, la situación del municipio (costa, interior), tamaño del municipio, densidad, cercanía a polígonos industriales, centrales nucleares, canteras, etc., pueden ser incluidos para darle un valor añadido a los datos.

Otro factor es el número de años con datos, en este caso desde 2008 a 2016.

***Porque **NO** elegimos datos por departamento*

El total del territorio de la comunidad valenciana se reparte en 24 departamentos, cada uno con sus municipios y/o zonas de municipios dentro de los cuales se encuentra la población asignada. Esto provoca que en un mismo departamento se encuentren municipios con distintas características, lo cual puede dificultar el estudio.

A favor de esta agrupación diremos que los datos están más agregados que en el caso de los municipios. Esto puede facilitar las tareas de análisis y/o minería de datos.

Otro inconveniente es el número de años con datos, en este caso desde 2014 a 2016.

- **La segunda bifurcación** nos da a elegir entre tres tipos de agrupación de los datos teniendo en cuenta la asignación de la causa de la muerte, estos tres tipos son: Causas CV, Causas INE y Causas CIE10.

Porque ***SI* elegimos datos* Causas INE

La agrupación de los datos se realiza mediante las causas de defunción recogidas por el INE (Instituto Nacional de Estadística), esto permite poder comparar los datos con otras provincias o ampliar el estudio a todo el país.

Estos datos se pueden encontrar en la página web del INE [16] siguiendo esta ruta:

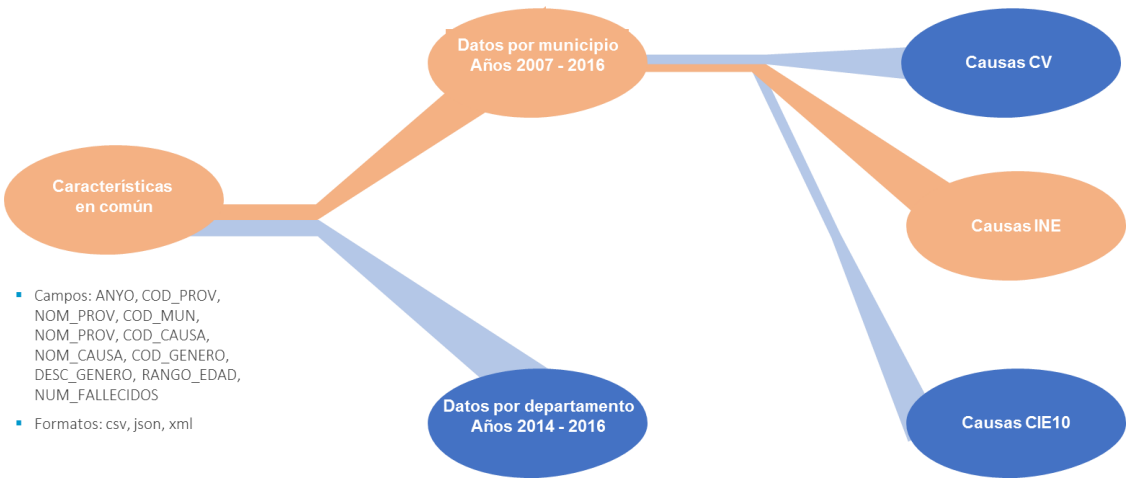
INEbase -> Sociedad -> Salud -> Estadística de defunciones según la causa de muerte

Porque ***NO* elegimos datos* Causas CV

La agrupación de estos datos es concreta para la comunidad valenciana, y esto impediría poder ampliar o extrapolar el estudio al resto de comunidades del país.

Porque ***NO* elegimos datos* Causas CIE10

Su agrupación es por provincia, no por municipio, lo que nos lleva a los mismos motivos que no haber elegido los datos por departamento. En este caso ese efecto se encuentra incluso más agravado pues se reducen a 3 los valores de asignación.



Resumen de decisiones

5.3. Descripción Fichero Seleccionado

Años: 2008 a 2016

Formatos: CSV, JSON, XML

El formato elegido será .CSV

Definición del formato .CSV [17]:

"Los archivos son un tipo de documento en formato abierto sencillo para representar datos en forma de tabla, en las que las columnas se separan por comas y las filas por saltos de línea."

Descripción: Distribución de las defunciones por municipio, sexo, causa (agrupadas en 102 causas, lista abreviada del INE-CCAA) y grupos de edad

Relación de campos y su descripción:

Nombre	Descripción
ANYO	Año de los datos
COD_PROV	Código INE de la provincia
NOM_PROV	Nombre de la provincia
COD_MUN	Código INE del municipio
NOM_MUN	Nombre del municipio
COD_CAUSA	Código de la causa de fallecimiento
DESC_CAUSA	Descripción de la causa de fallecimiento
COD_GENERO	Código del género
DESC_GENERO	Descripción del genero
RANGO	Rango de edad
NUM_FALLECIDOS	Número de fallecidos en el departamento, causa de fallecimiento, género y rango de edad.

5.4. Calidad del dato

El conjunto de datos utilizado para este estudio viene suministrado por el portal de transparencia de la Generalitat Valenciana, esto quiere decir que ya ha pasado un proceso de validación y consolidación de los datos.

Estos ya vienen anonimizados, agregados y revisados antes de su publicación. Lo que por una parte puede resultar un inconveniente en cuanto a cantidad de datos o detalle, por otra nos proporciona una seguridad de que el dato recogido este en buen estado.

A pesar de esto, en todo proceso de obtención de conocimiento mediante técnicas de minería de datos, se debe realizar una validación en busca de valores anómalos, repetidos, incompletos, etc.

Vamos a repasar con estadísticas y gráficamente el estado de cada campo. Para ello nos vamos a apoyar en la herramienta RapidMiner y su capacidad de extraer información y gráficas de los datos facilitados.

Vamos a tomar como ejemplo el fichero correspondiente al año 2014 al tener mayor número de registros que los de los años posteriores. Recordamos que los correspondientes al año 2015 y 2016 no tienen valor 0 en los registros sin actividad.

Campo ANYO

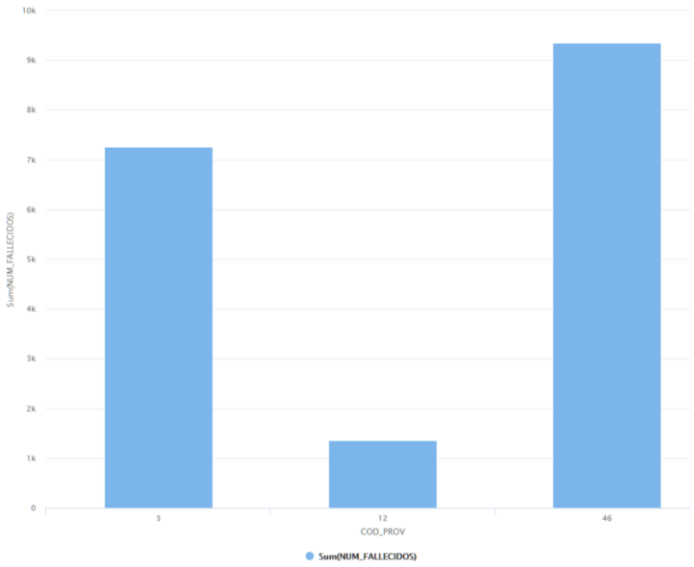
Todos los registros tienen el valor 2014.

0 valores nulos o desconocidos

Campo COD_PROV, NOM_PROV

Todos los registros tienen un valor entre los 3 posibles (3, 12 o 46), (ALICANTE, CASTELLÓN DE LA PLANA, VALENCIA) que se corresponden con el código y el nombre de cada provincia

0 valores nulos o desconocidos

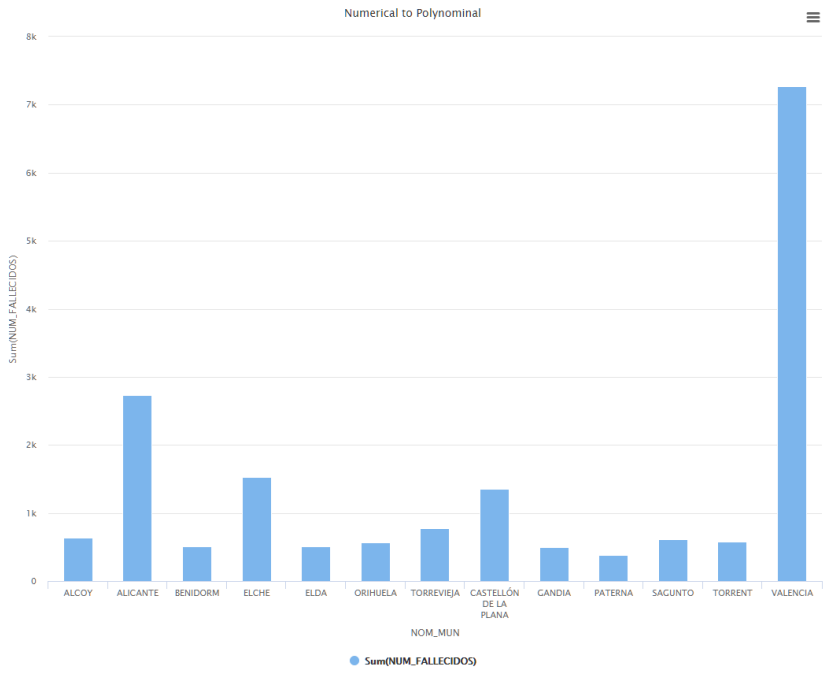


Distrubución de la suma del campo NUM_FALLECIDOS por COD_PROV o NOM_PROV

Campo COD_MUN, NOM_PROV

Todos los registros tienen un valor entre los 9 posibles (9, 14, 31, 40, 65, 66, 99, 131, 133, 190, 220, 244, 250), (ALCOY, ALICANTE, BENIDORM, CASTELLÓN DE LA PLANA, ELCHE, ELDA, GANDIA, ORIHUELA, PATERNA, SAGUNTO, TORRENT, TORREVIEJA, VALENCIA) que se corresponden con el código y nombre de cada municipio

0 valores nulos o desconocidos

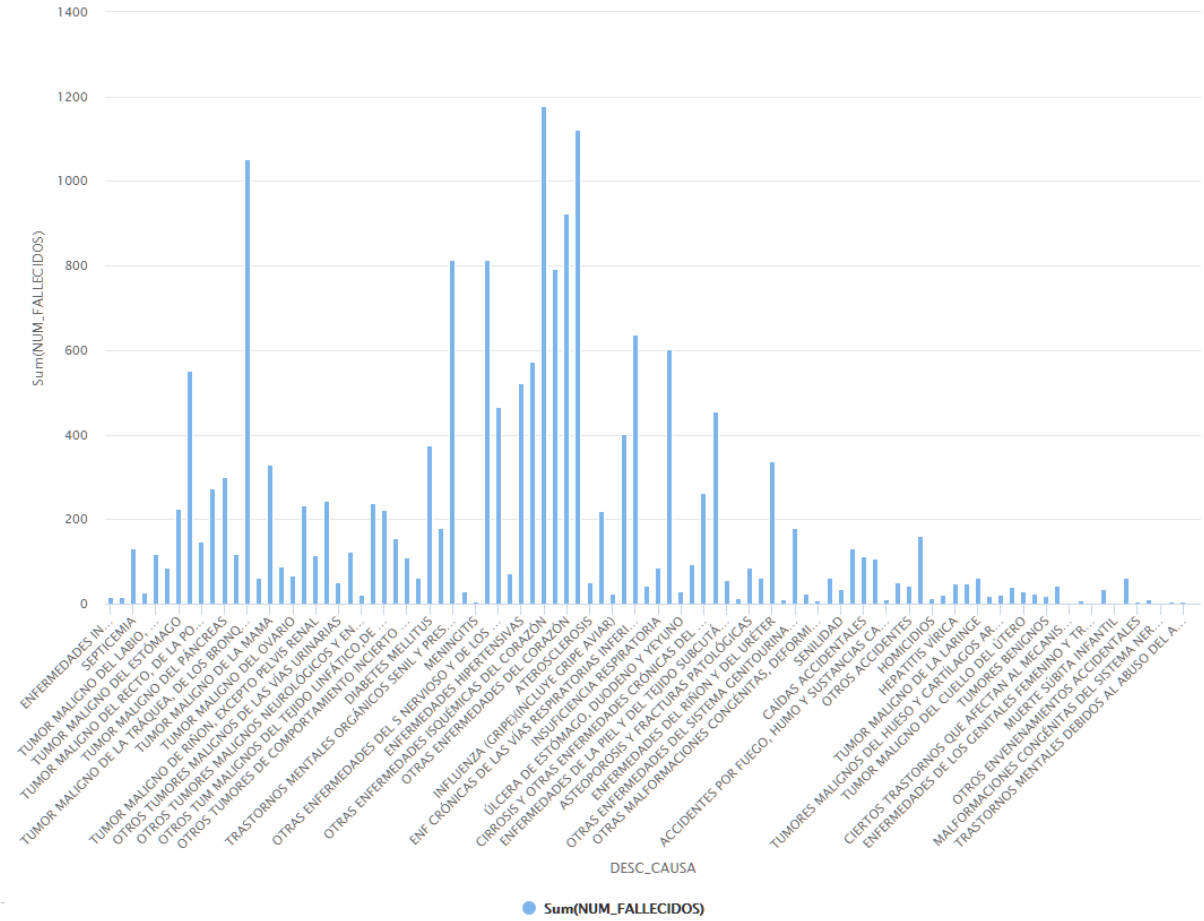


Distrubución de la suma del campo NUM_FALLECIDOS por COD_MUN o NOM_MUN

Campo COD_CAUSA, NOM_CAUSA

Todos los registros tienen un valor entre los 98 posibles (números del 1 al 101), (valor entre los 98 posibles. Descripción de la causa recogida por el INE) que se corresponden con el código de cada causa de defunción

0 valores nulos o desconocidos

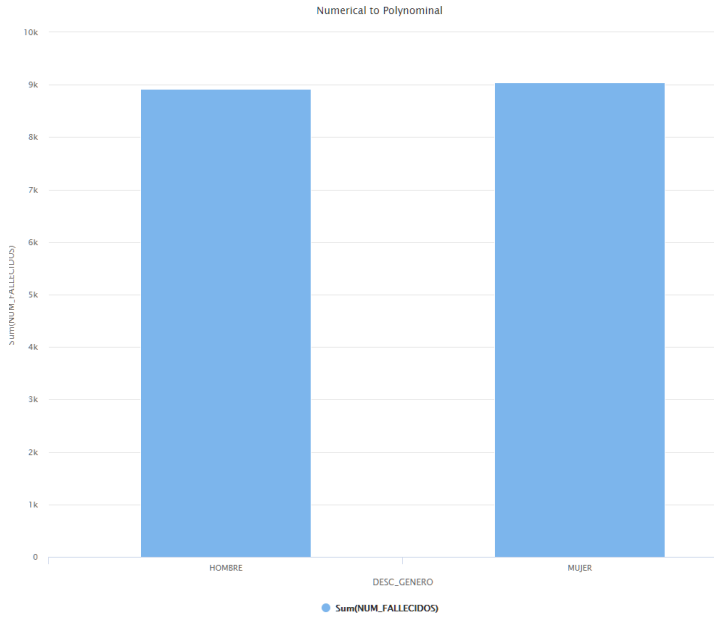


Distribución de la suma del campo NUM_FALLECIDOS por NOM_CAUSA

Campo COD_GENERO, DESC_GENERO

Todos los registros tienen un valor entre los 2 posibles (H, M), (HOMBRE, MUJER) que se corresponden con el código de cada genero

0 valores nulos o desconocidos

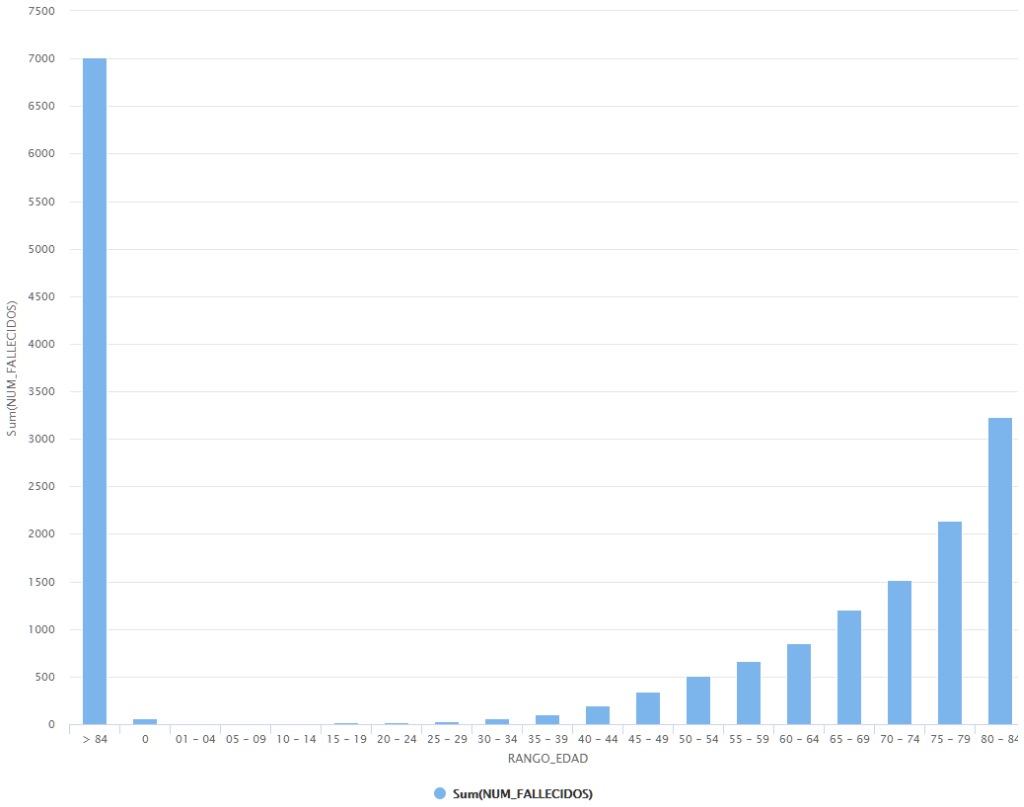


Distrubución de la suma del campo NUM_FALLECIDOS por DESC_GENERO

Campo RANGO_EDAD

Todos los registros tienen un valor entre los 17 posibles que se corresponden con la descripción de cada rango de edad.

0 valores nulos o desconocidos



Distrubución de la suma del campo NUM_FALLECIDOS por RANGO_EDAD

Campo NUM_FALLECIDOS

Valor de la medida sobre la que se basa el estudio. Todos los valores son numéricos, están comprendidos entre 0 y 210 y no existen nulos.

Datos no válidos

En el caso de que los datos no cumplan con los requisitos de calidad buscados como puede ser contener valores vacíos, valores incorrectos (por ejemplo, en COD_PROV un valor 'Valencia'), registros duplicados, etc., se debería de resolver estos errores en la fase de Preparación de Datos.

Cada caso se debería de resolver como corresponda, eliminando registros incorrectos, duplicados, asignando valores por defecto, etc.

5.5. Datos Externos

En este estudio se van a añadir datos externos a los obtenidos inicialmente. Estos datos externos se han obtenido por diferentes medios, desde Wikipedia, Google Maps, INE, periódicos.

Pensamos que aportan valor añadido a los datos iniciales y pueden ayudar a los modelos a mejorar sus resultados dando una visión ampliada de la realidad a la cual no llegan con una simple descripción o un nombre de municipio.

En la creación de una vista minable toda información que se pueda añadir, que cumpla que son datos correctos y no alteran la realidad, va a ayudar en la búsqueda de conocimiento y en la creación de modelos predictivos.

La información externa que se va a añadir es la siguiente:

Población del municipio

Se incluye el número correspondiente con la población actual de los municipios que participan en el estudio [18].

Renta media del municipio

Se incluye el número correspondiente con la renta media de la población actual de los municipios que participan en el estudio [18].

Con esta información [18] hemos importado al estudio un fichero que será utilizado en la fase de preparación de los datos.

Municipio Costa / Interior

Se incluye el valor “Si” / “No” según corresponda con la localización de los municipios que participan en el estudio [19].

Sector Principal

Se incluye el valor “Agricultura” / “Industria” / “Servicios” según corresponda con la actividad económica principal de los municipios que participan en el estudio [20].

Teniendo en cuenta estos 2 últimos puntos en cuanto a datos externos, hemos generado un fichero de texto donde se recoge esta información. En él se muestra el nombre del municipio, si es de costa o no, y el sector económico principal. Este fichero también será utilizado en la fase de preparación de los datos.

5.6. Comparación con otros tipos de almacenamiento

En este apartado vamos a tomar como referencia los conocimientos impartidos en las asignaturas de “Sistemas de información estratégicos” y “Almacenes de datos y minería de datos” y tomarlos como referencia para comparar algunas características de los almacenes de datos con las características de nuestros datos. Hemos elegido los almacenes de datos porque es la alternativa más comúnmente usada para almacenar los datos de una empresa para su análisis posterior.

Datos del estudio vs Almacenes de datos

En la mayoría de trabajos o estudios de minería de datos, se tiene como fuente un almacén de datos. La información con la que vamos a trabajar no se rige por esa estructura, puesto que los datos están en formato de texto plano y no almacenados en una base de datos al uso.

Sabiendo eso, vamos a comentar algunos aspectos y características de un almacén de datos que son comparables con el formato que encontramos en el conjunto de datos que hemos obtenido.

Viendo ejemplos de los datos y sus correspondientes metadatos, podemos pasar a describirlos desde un punto más teórico, describiendo qué características cumplen y cuáles no, tratándolos como si fuera un almacén de datos.

Extracción, Transformación y Carga

En un proceso de Extracción, Transformación y Carga (ETL) realizado en la construcción de un almacén de datos, se realizan una serie de acciones para pasar del dato original al dato correctamente almacenado en cuanto a consistencia y formato. En nuestro caso estas acciones no son necesarias al obtener los datos ya construidos correctamente. Esto no quiere decir que a posteriori se realicen transformaciones sobre ellos, pero ya será para completarlos o adaptarlos a una tarea de minería de datos.

Dimensión Tiempo

En todos los almacenes de datos existe la dimensión Tiempo, esta es básica para realizar búsquedas, agregaciones y conocer el histórico de los datos almacenados. El contenido de esta dimensión y su detalle depende de las características del dato almacenado. En nuestro caso la dimensión tiempo solo viene representada por el valor del año.

Jerarquía en datos

Una de las características principales en un almacén de datos. En nuestro caso solo se daría entre los campos [Provincia <-> Municipio] dando lugar a un nivel de navegación. La operación de navegar por distintos niveles de una jerarquía también se llama “Drill Down/up”.

Hechos y dimensiones

Tomando como referencia estos dos aspectos claves en la definición de los almacenes de datos, viendo los datos del estudio podemos observar que solo hay un esquema de datos, lo que sería equivalente a una tabla, por lo cual no existirán tablas de hechos y tablas de dimensiones.

Los valores de los hechos se utilizan para crear medidas y estas medidas pueden ser de 3 tipos según su capacidad de agregarse con el conjunto de las dimensiones.

Viendo los datos almacenados en al campo NUM_FALLECIDOS podemos decir que la medida resultante será de tipo “Aditiva”. Esto quiere decir que se pueden agregar a todas las dimensiones de la tabla de hechos.

En el caso de no ser así pasarían a ser “Semiaditivas”, cuando solo son agregables a algunas dimensiones.

Y, por último, las “No aditivas” son aquellas medidas en las que no se puede realizar una agregación con ninguna de las dimensiones.

Información Extensional/intensional

Partimos de una información extensional (datos) obtenida a partir de los ficheros descargados, y se busca llegar a una información intensional (conocimiento).

6. Análisis predictivo y descriptivo – Fase de preparación

Una vez tenemos los datos, la herramienta software y la metodología seleccionadas, es hora de pasar de un planteamiento teórico, a una aplicación técnica.

Para ello vamos a tomar los ficheros correspondientes al conjunto “datos-mortalidad---municipios---causas-ine-ccaa” que consta de 10 ficheros, que van desde el año 2007 al año 2016. El motivo de seleccionar este conjunto de ficheros y sus características están descritos en el capítulo 5.

Estos datos van a ser importados en la herramienta de minería de datos RapidMiner. Las diferentes versiones existentes, los motivos y las características de este están descritas en el capítulo 4.

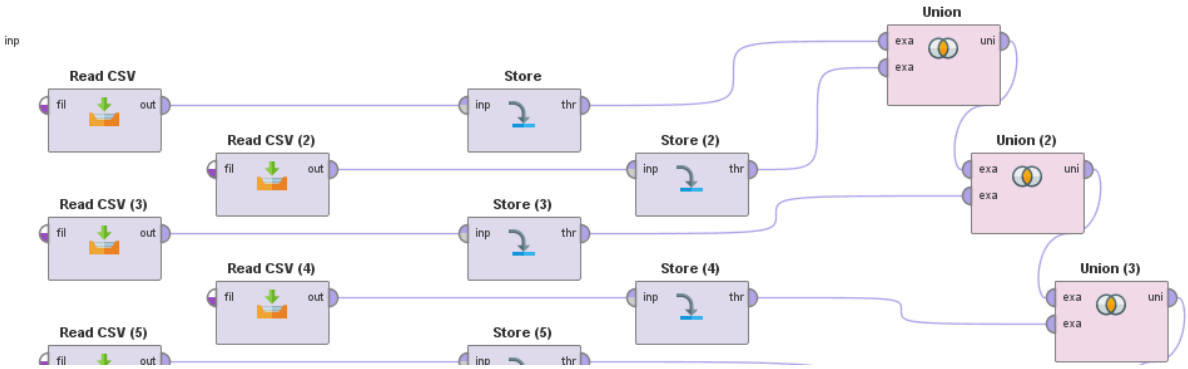
Una vez ya tenemos los datos cargados se va a pasar a realizar las 3 fases clásicas en un proceso de minería de datos, estas son “Preparación de los datos”, “Modelado” y “Evaluación”. Dentro de cada una se realizarán varias tareas las cuales se describirán en este capítulo. Se tomará como referencia la metodología CRISP-DM, pero no se realizarán todos los pasos que en ella se describen, para eso sería necesario trabajar sobre un proyecto o estudio más amplio en que tenga cabida o se le pueda sacar provecho en toda su amplitud.

6.1. Preparación de los datos

En esta fase explicaremos los pasos realizados en RapidMiner para, partiendo de los ficheros originales, llegar a obtener un conjunto de datos que se tomara como vista minable. Es muy importante realizar una buena preparación, tanto en limpieza como en enriquecimiento de los datos para poder sacar todo el potencial a estos. Tanto es así que en muchos de los textos que hablan sobre cómo plantear esta parte del proceso de minería de datos, recalcan que puede llegar ocupar el 50% del tiempo sobre el total del estudio/proyecto.

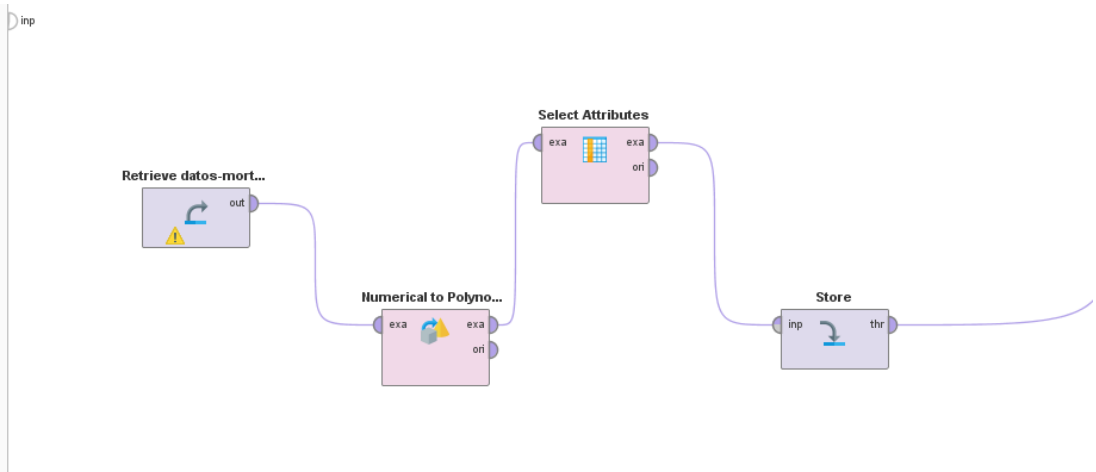
A continuación, se describen las 6 tareas realizadas, dentro de las cuales se pueden llevar a cabo varias operaciones, mostrando una vista del flujo de trabajo, conexión entre las operaciones y comportamiento de estas.

Tarea 1 – Read & Store



1. Se utiliza un operador “Read CSV” para cargar un fichero con los datos.
2. Se utiliza un operador “Store” para almacenarlos dentro del repositorio del RapidMiner creado para el estudio generando un DataSet.
3. Este paso se realiza para los 10 ficheros que se van a cargar.
4. Se utiliza el operador “Union” con cada fichero y el resultado del anterior “Union” hasta obtener un conjunto de datos con los 10 ficheros.
5. Se utiliza un operador “Store” para almacenar el resultado generando un DataSet y poder utilizarlo en próximas tareas.

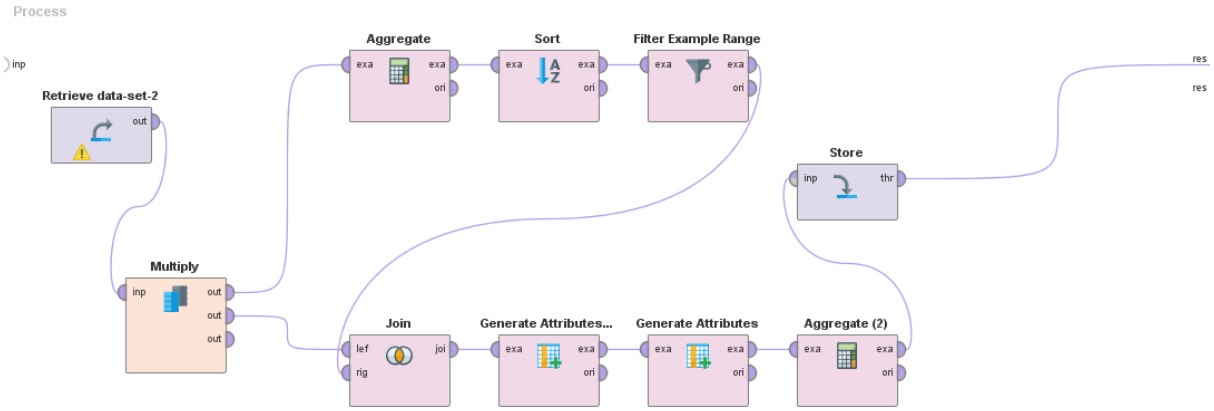
Tarea 2 – Filtrado



1. Se utiliza un operador “Retrieve” para importar el DataSet de la tarea 1.

2. Se transforma el tipo del campo ANYO de numérico a polinómico mediante un operador de cambio de tipo.
3. Se utiliza un operador “Select Atributes” para desechar los campos COD_CAUSA, COD_GENERO, COD_MUN, COD_PROV, NOM_PROV. Los valores d estos campos ya están representados en sus versiones de NOM o no se van a tener en cuanta como es el caso de los campos correspondientes a la información de la provincia.
4. Se utiliza un operador “Store” para almacenar el resultado generando un DataSet y poder utilizarlo en próximas tareas.

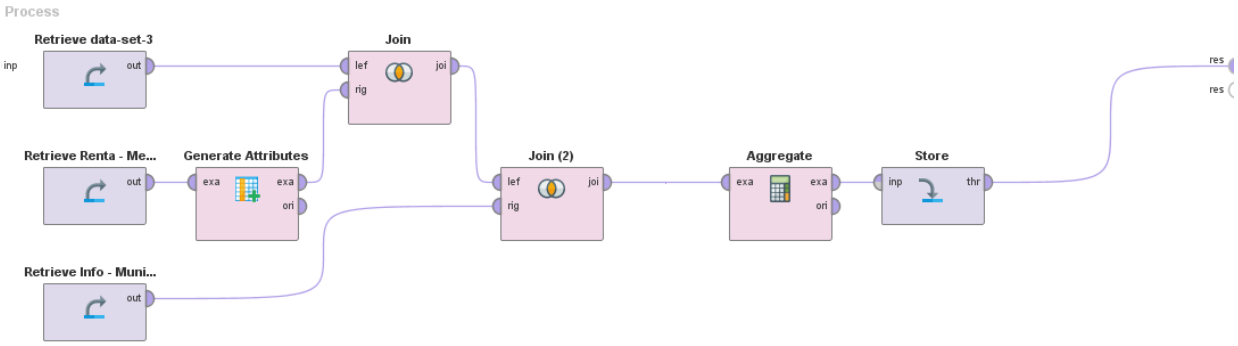
Tarea 3 – Edad y Top Causas



1. Se utiliza un operador “Retrieve” para importar el DataSet de la tarea 2.
2. Rama Top Causas
    a. Se utiliza un operador “Aggregate” para agregar los datos del campo NUM_FALLECIDOS por el campo DESC_CAUSA.
    b. Se utiliza un operador “Sort” para ordenar en orden descendente los valores del campo SUM(NUM_FALLECIDOS).
    c. Se utiliza un operador “Filter” para recoger los 25 valores de DESC_CAUSA con mayor resultado en SUM(NUM_FALLECIDOS).
3. Una vez obtenidas las 25 causas con mayor resultado, se utiliza un operador “Join” para cruzar con el DataSet importado en el paso 1. Esto da como resultado un DataSet donde solo se recogen los registros de las 25 causas comentadas.
4. Se utiliza un operador “Generate Attributes” para corregir un error en los datos originales, en el campo RANGO_EDAD, el valor ‘55 - 59’ viene con un espacio al final. Transformamos todos los valores ‘55 - 59 ‘ a ‘55 – 59’.

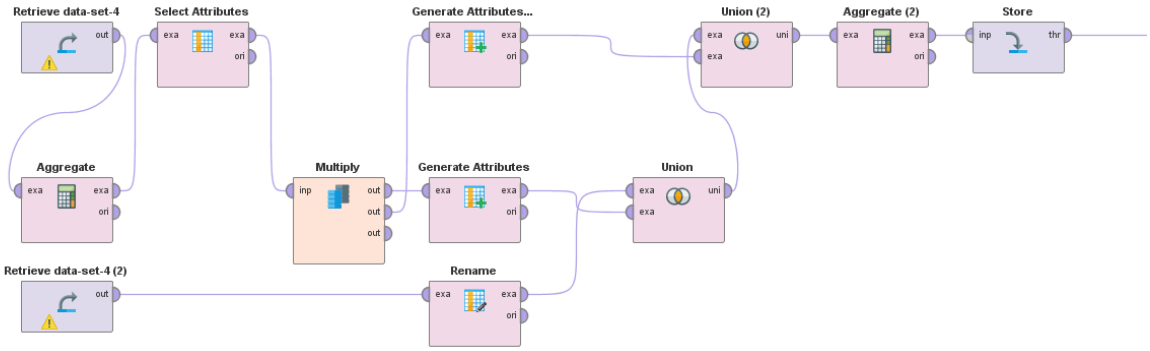
5. Se vuelven a tratar los valores del campo RANGO_EDAD, esta vez para pasar de 16 valores distintos a 7 creando un nuevo campo
6. Se agregan los datos en torno a los nuevos valores de RANGO_EDAD y se desecha la columna anterior.
7. Se utiliza un operador “Store” para almacenar el resultado generando un DataSet y poder utilizarlo en próximas tareas

Tarea 4 - Población



1. Se importa el fichero “Renta - Media - Municipios”
2. Se crean nuevos campos que se corresponden con:
    a. Gran_Ciudad = if(Habitantes > 100000, "Si","No")
    b. Renta = if([Renta bruta media] > 22500, "Alta", "Baja")
3. Se importa el fichero “Info - Municipios”
    a. Se incluyen nuevos campos “Costa” y “Sector”
4. Se importa el DataSet creado en la tarea 3
5. Se realizan dos Join con los dos DataSets anteriores que aportan nuevos campos y el creado en el paso 4 de esta tarea
6. Realizamos una agregación sobre los datos desechando los campos intermedios o que no aportan información
7. Se utiliza un operador “Store” para almacenar el resultado generando un DataSet y poder utilizarlo en próximas tareas

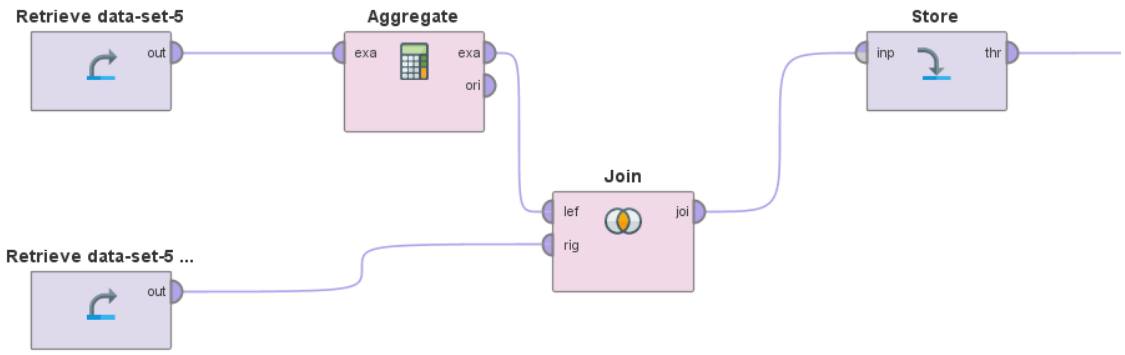
Tarea 5 – Completar con 0



Aclaración previa: Para igualar el número de registros entre los datos de cada año, a los datos correspondiente a los años 2015 y 2016 se les va a añadir registros con el valor 0 en el campo NUM_FALLECIDOS para las combinaciones del resto de valores que no tengan ningún valor.

1. Se importa el DataSet creado en la tarea 4.
2. Se agrupan los datos por los valores de todos los campos menos de ANYO.
3. Se elimina el campo SUM(NUM_FALLECIDOS) y se deja un DataSet con todos los valores posibles del resto de campos.
4. Se añade un campo ANYO con los valores 2015 y 2016 y otro campo NUM_FALLECIDOS con valor 0 en todos los registros.
5. Se importa el DataSet creado en la tarea 4 en otra rama.
6. Por medio de Union se unifican los 3 DataSets.
7. Se vuelve a agregar por todos los campos para eliminar registros duplicados con 0.
8. Se utiliza un operador “Store” para almacenar el resultado generando un DataSet y poder utilizarlo en próximas tareas.

Tarea 6 - Media



1. Se importa el DataSet creado en la tarea 5.
2. Se agregan los datos juntando todos los años recogidos generando una media.
3. Se realiza una Join con el DataSet creado en la tarea 5 para añadir el nuevo campo AVERAGE(NUM_FALLECIDOS).
4. Se utiliza un operador “Store” para almacenar el resultado generando un DataSet y poder utilizarlo en próximas tareas.

6.2. Vista minable

Una vez completados todos estos pasos el DataSet obtenido será tomado como vista minable. Una vista minable consiste en un conjunto de datos a los cuales se les ha aplicado una seria de transformaciones (conversión, filtrado, etc.) y se toma como punto de partida para aplicarle las distintas tareas de modelado.

Row ...	DESC_CAUSA	DESC_GENERO	Edad	Gran_Ciudad	Renta	Costa	Sector	average(...	ANYO	sum(NUM_FALLECIDOS)
1	CIRROSIS Y OTRAS ENFE...	HOMBRE	0	No	Alta	No	Industria	0	2007	0
2	CIRROSIS Y OTRAS ENFE...	HOMBRE	0	No	Alta	No	Industria	0	2008	0
3	CIRROSIS Y OTRAS ENFE...	HOMBRE	0	No	Alta	No	Industria	0	2009	0
4	CIRROSIS Y OTRAS ENFE...	HOMBRE	0	No	Alta	No	Industria	0	2010	0
5	CIRROSIS Y OTRAS ENFE...	HOMBRE	0	No	Alta	No	Industria	0	2011	0
6	CIRROSIS Y OTRAS ENFE...	HOMBRE	0	No	Alta	No	Industria	0	2012	0
7	CIRROSIS Y OTRAS ENFE...	HOMBRE	0	No	Alta	No	Industria	0	2013	0
8	CIRROSIS Y OTRAS ENFE...	HOMBRE	0	No	Alta	No	Industria	0	2014	0
9	CIRROSIS Y OTRAS ENFE...	HOMBRE	0	No	Alta	No	Industria	0	2015	0
10	CIRROSIS Y OTRAS ENFE...	HOMBRE	0	No	Alta	No	Industria	0	2016	0
11	CIRROSIS Y OTRAS ENFE...	HOMBRE	0	No	Alta	No	Servicios	0	2007	0
12	CIRROSIS Y OTRAS ENFE...	HOMBRE	0	No	Alta	No	Servicios	0	2008	0
13	CIRROSIS Y OTRAS ENFE...	HOMBRE	0	No	Alta	No	Servicios	0	2009	0
14	CIRROSIS Y OTRAS ENFE...	HOMBRE	0	No	Alta	No	Servicios	0	2010	0
15	CIRROSIS Y OTRAS ENFE...	HOMBRE	0	No	Alta	No	Servicios	0	2011	0
16	CIRROSIS Y OTRAS ENFE...	HOMBRE	0	No	Alta	No	Servicios	0	2012	0
17	CIRROSIS Y OTRAS ENFE...	HOMBRE	0	No	Alta	No	Servicios	0	2013	0

Imagen tomada desde RapidMiner de estado final de la vista minable, la cual tiene un total de 21.618 registros.

7. Análisis predictivo y descriptivo – Fase de modelado y evaluación

En esta parte explicaremos los pasos realizados en RapidMiner para, partiendo del DataSet generado en el apartado de **Preparación de los datos**, aplicar los algoritmos predictivos seleccionados para obtener varios modelos de predicción. Los cuales se utilizarán para medir su acierto en el apartado de **Evaluación**.

Señalar que en todos los modelos y pruebas se van a realizar tareas de **Regresión**. Este es el tipo de tarea que mejor se adapta a los datos con los que se ha realizado el estudio y con los valores esperados.

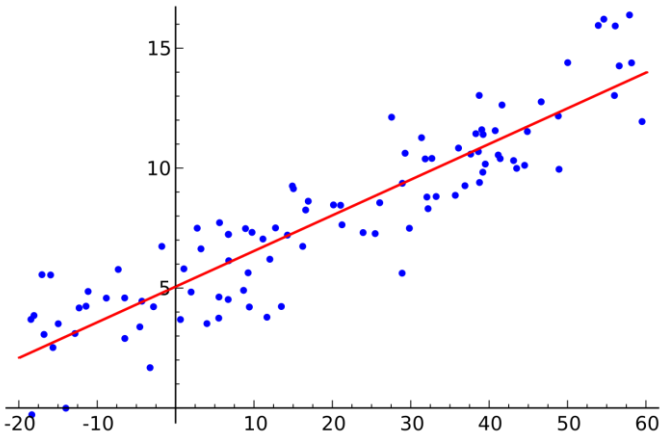
7.1. Tareas utilizadas para el modelado

Entre las tareas de regresión conocidas y/o utilizadas en las asignaturas ya comentadas, vamos a seleccionar 3 para generar los modelos. Estas serán Regresión Lineal, Árbol de decisión y Red Neuronal. Estas son sus principales características y como se utilizarán en el estudio.

Regresión Lineal

“En estadística la regresión lineal o ajuste lineal es un modelo matemático usado para aproximar la relación de dependencia entre una variable dependiente Y, las variables independientes Xi y un término aleatorio ε.” [21]

En la imagen podemos observar la representación gráfica de una regresión lineal con una variable dependiente y una variable independiente.

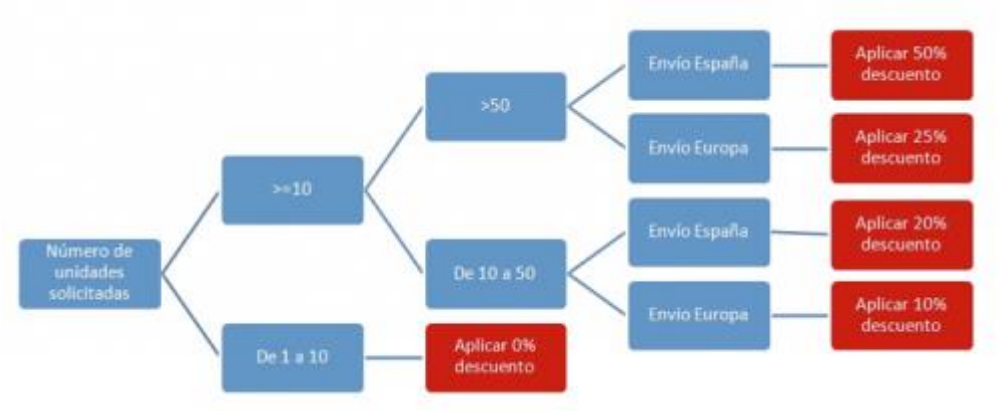


Fuente [21]

Árbol de decisión

"Los árboles de decisión son estructuras de datos basadas en toma de decisiones en forma de árbol. Cada árbol representa un conjunto de decisiones y van guiando el proceso de aprendizaje en función de la decisión tomada; las decisiones generan reglas para la clasificación de los datos." [22]

A continuación, puede verse un ejemplo de árbol de decisión:

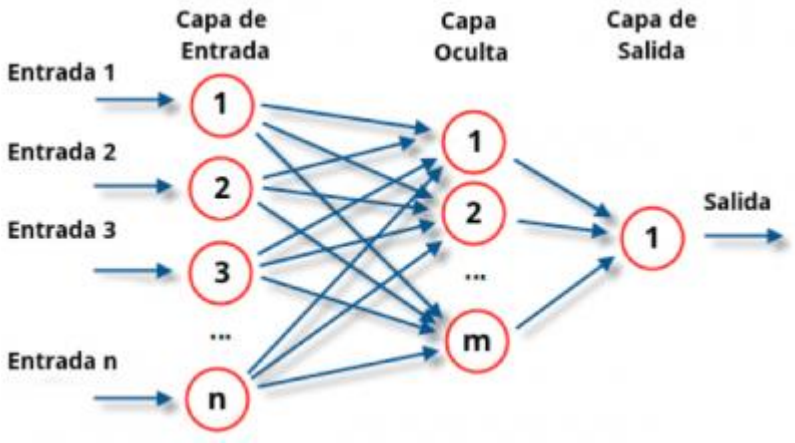


Fuente: [22]

Red Neuronal

"Las redes neuronales son una herramienta utilizada para análisis estadísticos que permite la construcción de un modelo de comportamiento a partir de unos datos de entrada. La red neuronal va aprendiendo desde cero a partir de los datos de entrada, para después transformarse en un modelo capaz de rendir cuenta del comportamiento observado en función de los datos de entrada. Es decir, la red neuronal, una vez construida, constituye un verdadero modelo que actúa en función de lo que percibe." [22]

En la imagen siguiente podemos observar el funcionamiento de una red neuronal:



Fuente: [22]

7.2. Versiones a utilizar de la vista minable

Estas tres técnicas para resolver tareas de regresión se van a aplicar sobre diferentes versiones de la vista minable generada en la fase de Preparación de los datos. Estas versiones nos servirán para comprobar si se mejoran los resultados obtenidos una vez partimos de la vista minable.

Estas versiones serán:

1 - Vista minable inicial

Esta versión de la vista minable parte del dataset creado en el último paso de la fase de Preparación de los datos, pero excluyendo la columna AVERAGE(NUM_FALLECIDOS).

Estos datos se tomarán como referencia para crear el resto de versiones (excluyendo la que incluye la media).

2 - Vista minable (con media)

Solo vamos a incluir el campo AVERAGE(NUM_FALLECIDOS) en esta versión de la vista minable. Esto es debido a que el campo AVERAGE(NUM_FALLECIDOS) lo vamos a tratar aparte puesto que un campo de estas características puede orientar mucho la predicción sobre él, dejando el resto de campos con menor peso.

La media que se representa en este campo, **no es la media de todo el conjunto** de datos si no la media que resulta de eliminar el campo ANYO y sacar la media a todos los registros con valores iguales. Por ejemplo, para cada DESC_CAUSA, DESC_GENERO, EDAD, Gran_Ciudad, Renta, Costa y Sector resultante, se saca la media de sus registros.

Al incluir un campo que refleja la media de los 10 años de que se compone el estudio la predicción siempre se va a parecer a este valor, y lo que queremos es no condicionar los resultados a un campo que ya le estamos dando, si no aportar valor al resto de campos que son los que conforman las características reales y futuras de los datos. La media siempre se puede considerar un valor de referencia con el que comparar nuestras predicciones las cuales aportarán conocimiento si son mejores que la media.

3 - Sin Datos Externos

Durante la fase de la fase de Preparación de los datos se han incluido una serie de campos que reflejan información sobre los municipios de los que se disponen datos. Estos campos eran “Gran Ciudad”, “Renta Media”, “Costa” y “Sector”.

Se han incluido para buscar características de los datos que puedan aportar valor añadido y para poder exportar los modelos de predicción a otras comunidades o a todo

el territorio español. Pensamos que trabajar con el nombre de un municipio no aporta nada, pero si las características de este que se repetirán a lo largo de la geográfica española.

Por lo que vamos a realizar una versión sin estos campos para comprobar si se ha alcanzado el objetivo de mejorar la predicción incluyéndolos.

4 - Por Sexo (Hombre | Mujer)

Pensamos que esta separación puede ser muy importante, las causas de mortalidad siempre han marcado una tendencia distinta en cuanto al sexo de la persona. Viendo las estadísticas globales, no solo de la comunidad valencia, las causas pueden ser muy distintas, entre las que destacan distintos tipos de cáncer, por ejemplo.

Al aplicar las tareas de modelado, vamos a filtrar antes por el campo DESC_GENERO en sus dos valores recogidos HOMBRE y MUJER y comparar sus resultados

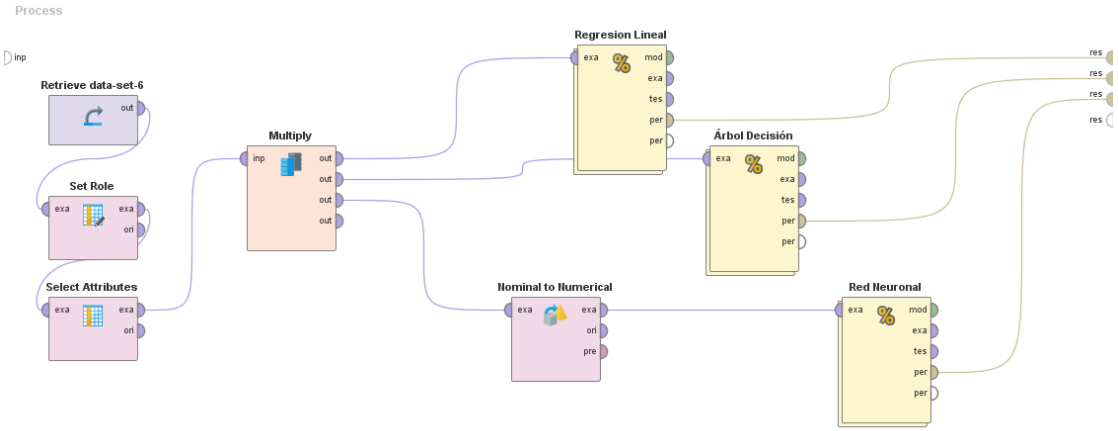
5 - Por Edad (Menor de 65 | Mayor de 65)

Del mismo modo que el apartado anterior, las causas de mortalidad difieren en gran medida si se tiene en cuenta la edad a partir de los 65 años. Para comprobar este factor vamos a filtrar los datos por el campo RANGO_EDAD separando la población en menor de 65 y mayor de 65 años. Después les aplicaremos las tareas de modelado y compararemos sus resultados.

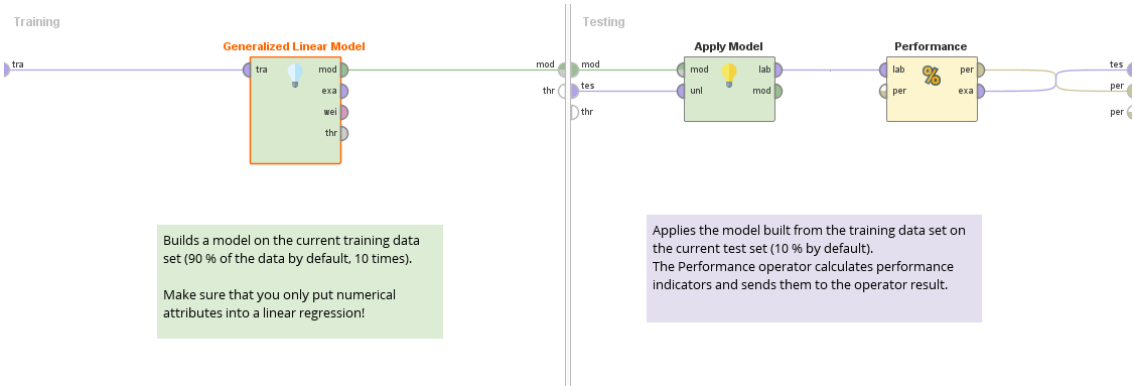
7.3. Aplicación de las técnicas de minería sobre cada vista minable

Una vez tenemos seleccionas las tareas de modelado y las vistas minables sobres las cuales las vamos a aplicar, vamos a mostrar cómo es su aplicación desde RapidMiner.

A continuación, se describen los pasos realizados en cada una de los 5 procesos (uno para cada versión de la vista minable utilizada) generados, en las cuales dentro se pueden llevar a cabo varias operaciones, mostrando una vista del flujo de trabajo, conexión entre las operaciones y comportamiento de éstas.



Este es el proceso generado para trabajar con la vista “1 - Vista minable inicial” y que será tomada como base para el resto.



Aquí se muestra el interior del proceso “Regresión lineal” de la imagen anterior.

A continuación, se explica el funcionamiento de cada proceso y sus características.

Proceso para **1 - Vista minable inicial**

1. Se importa el DataSet final creado en la fase de Preparación de datos.
2. Se utiliza un operador “Set Role” para indicar que el campo NUM_FALLECIDOS es el atributo de salida a tener en cuenta en las tareas de modelado.
3. Se excluye el campo AVERAGE(NUM_FALLECIDOS).
4. Se utiliza un operador “Multiply” para poder utilizar la salida del paso anterior en varios operadores.
5. Se utiliza un “Building Block” del tipo “Numerical Cross Validation” que se compone de tres operadores:
    a. Generación del modelo, en este caso “Regresión Lineal”
    b. Aplicación del modelo
    c. Evaluación del rendimiento
6. El paso 5 se repite para generar el modelo con “Árbol de decisión” y “Red Neuronal”.

7. La salida con el valor de la evaluación de rendimiento de cada uno de los 3 “Building Block” se utiliza para mostrar sus resultados.

Dentro de cada “Building Block” se realiza el proceso de “**Validación cruzada**”, que se describirá en el apartado **7.4 Evaluación.**

Proceso para **2 - Vista minable (con media)**

1. Se importa el DataSet final creado en la fase de Preparación de datos.
2. Se utiliza un operador “Set Role” para indicar que el campo NUM_FALLECIDOS es el atributo de salida a tener en cuenta en las tareas de modelado.
3. Se utiliza un operador “Multiply” para poder utilizar la salida del paso anterior en varios operadores.
4. Se utiliza un “Building Block” del tipo “Numerical Cross Validation” que se compone de tres operadores:
    a. Generación del modelo, en este caso “Regresión Lineal”
    b. Aplicación del modelo
    c. Evaluación del rendimiento
5. El paso 5 se repite para generar el modelo con “Árbol de decisión” y “Red Neuronal”.
6. La salida con el valor de la evaluación de rendimiento de cada uno de los 3 “Building Block” se utiliza para mostrar sus resultados.

Proceso para **3 - Sin Datos Externos**

1. Se importa el DataSet final creado en la fase de Preparación de datos.
2. Se utiliza un operador “Set Role” para indicar que el campo NUM_FALLECIDOS es el atributo de salida a tener en cuenta en las tareas de modelado.
3. Se excluyen los campos AVERAGE(NUM_FALLECIDOS), Gran Ciudad, Renta Media, Costa y Sector.
4. Se utiliza un operador “Multiply” para poder utilizar la salida del paso anterior en varios operadores.
5. Se utiliza un “Building Block” del tipo “Numerical Cross Validation” que se compone de tres operadores:
    a. Generación del modelo, en este caso “Regresión Lineal”
    b. Aplicación del modelo
    c. Evaluación del rendimiento
6. El paso 5 se repite para generar el modelo con “Árbol de decisión” y “Red Neuronal”.
7. La salida con el valor de la evaluación de rendimiento de cada uno de los 3 “Building Block” se utiliza para mostrar sus resultados.

Proceso para **4 - Por Sexo (Hombre | Mujer) (2 ejecuciones)**

1. Se importa el DataSet final creado en la fase de Preparación de datos.
2. Se utiliza un operador “Set Role” para indicar que el campo NUM_FALLECIDOS es el atributo de salida a tener en cuenta en las tareas de modelado.
3. Se excluyen los campos AVERAGE(NUM_FALLECIDOS).
4. Se utiliza un operador “Filter” para filtrar los datos por el campo DESC_GENERO (en una ejecución por el valor HOMBRE y en la siguiente por el valor MUJER).
5. Se utiliza un operador “Multiply” para poder utilizar la salida del paso anterior en varios operadores.
6. Se utiliza un “Building Block” del tipo “Numerical Cross Validation” que se compone de tres operadores:
    a. Generación del modelo, en este caso “Regresión Lineal”
    b. Aplicación del modelo
    c. Evaluación del rendimiento
7. El paso 5 se repite para generar el modelo con “Árbol de decisión” y “Red Neuronal”.
8. La salida con el valor de la evaluación de rendimiento de cada uno de los 3 “Building Block” se utiliza para mostrar sus resultados.

Proceso para **5 - Por Edad (Menor de 65 | Mayor de 65) (2 ejecuciones)**

1. Se importa el DataSet final creado en la fase de Preparación de datos.
2. Se utiliza un operador “Set Role” para indicar que el campo NUM_FALLECIDOS es el atributo de salida a tener en cuenta en las tareas de modelado.
3. Se excluyen los campos AVERAGE(NUM_FALLECIDOS).
4. Se utiliza un operador “Filter” para filtrar los datos por el campo RANGO_EDAD (en una ejecución por el valor ‘< 65’ y en la siguiente por el valor ‘>= 65’).
5. Se utiliza un operador “Multiply” para poder utilizar la salida del paso anterior en varios operadores.
6. Se utiliza un “Building Block” del tipo “Numerical Cross Validation” que se compone de tres operadores:
    a. Generación del modelo, en este caso “Regresión Lineal”
    b. Aplicación del modelo
    c. Evaluación del rendimiento
7. El paso 5 se repite para generar el modelo con “Árbol de decisión” y “Red Neuronal”.

La salida con el valor de la evaluación de rendimiento de cada uno de los 3 “Building Block” se utiliza para mostrar sus resultados.

7.4. Evaluación

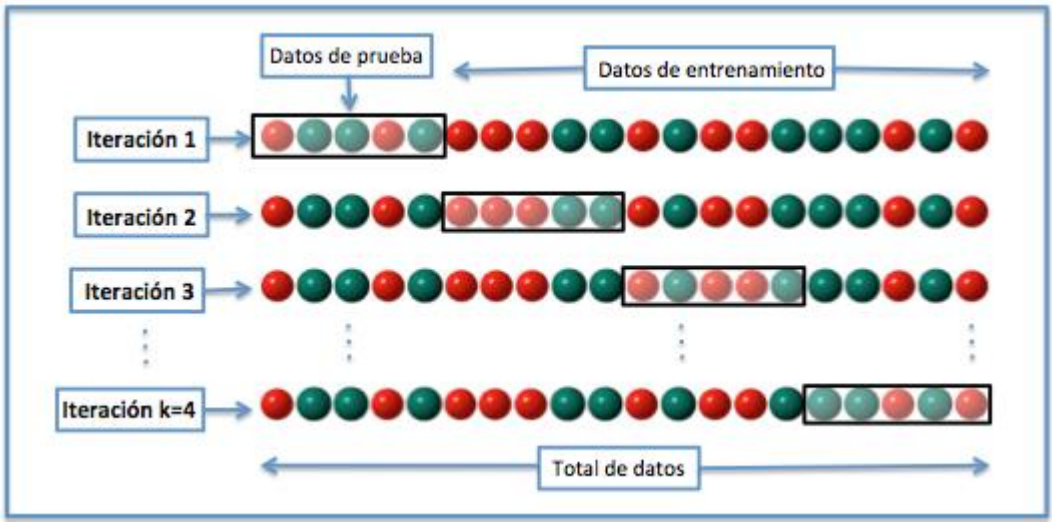
Una vez ya hemos desarrollado y ejecutado los procesos descritos en el apartado anterior, vamos a pasar a mostrar los resultados obtenidos y realiza una evaluación de su valor.

Método de validación

En el apartado anterior se ha indicado que se utiliza "**Validación cruzada**" para aplicar cada tarea de modelado y su validación, que consiste en lo siguiente [23]:

"Los datos se parten en N pliegues, de los cuales el primero se usa para test y el resto para entrenamiento. A continuación, el segundo se usa para test y el resto para entrenamiento. Y así sucesivamente hasta completar los N pliegues. Esto se realiza para evitar entrenar y evaluar con los mismos datos, si no se corre el riesgo de caer en el sobreajuste (*overfitting*)".

El caso habitual y el que utilizaremos en este estudio es realizar la validación cruzada con 10 pliegues.



Fuente: [23]

Medida de evaluación

Para comparar resultados y validar su comportamiento, vamos a tomar como referencia una medida de evaluación de la regresión como es la "**Raíz del error cuadrático medio**" (Root Mean Squared Error). Con este valor podemos observar la diferencia entre los valores predichos que hemos obtenidos y los reales (es decir, el

error en las predicciones), las diferencias en cuanto al error entre utilizar una tarea de modelado u otra, y si las distintas versiones de la vista minable cumplen con lo esperado.

La definición de Raíz del error cuadrático medio es [24]:

"Medida de uso frecuente de las diferencias entre los valores (valores de muestra o de población) predichos por un modelo o un estimador y los valores observados. La RECM representa la raíz cuadrada del segundo momento de la muestra de las diferencias entre los valores previstos y los valores observados o la media cuadrática de estas diferencias."

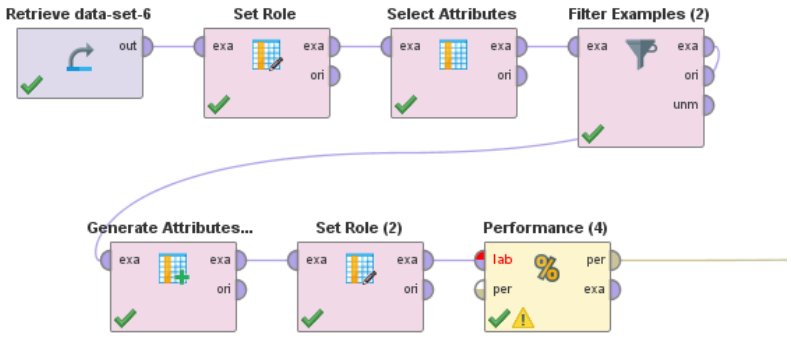
Modelo de referencia (media)

En las tareas de regresión, siempre es interesante comprar los resultados con el modelo de referencia (media). Este modelo no es necesario construirlo basta con ejecutar un proceso que tome como valor de predicción la media. Este paso se realizará para todas las versiones de la vista minable y se incluirá en la tabla de resultados junto al resto de resultados obtenidos por las técnicas de minería ya comentadas.

Estadísticas del campo NUM_FALLECIDOS

- Min 0
- Max 354
- Media 5,364
- Desviación 20,128

Y esta su implementación en RapidMiner:



Proceso para **1 - Vista minable inicial (Resultado con media)**

1. Se importa el DataSet final creado en la fase de Preparación de datos.

2. Se utiliza un operador “Set Role” para indicar que el campo NUM_FALLECIDOS es el atributo de salida a tener en cuenta en las tareas de modelado.
3. Se excluye el campo AVERAGE(NUM_FALLECIDOS).
4. Se utiliza un operador “Filter” para filtrar los registros (en las pruebas que hiciera falta).
5. Se utiliza un operador “Generate Attributes” con el que se crea el campo MEDIA EVALUACION con valor = 5,364
6. Se indica que el campo MEDIA EVALUACION es el resultado de la predicción.
7. Se evalúa el rendimiento del modelo ficticio creado y se recoge el “root_mean_squared_error”

Resumen de resultados

En esta tabla recogemos el valor del error cuadrático medio (Root Mean Squared Error) de los 4 modelos propuestos:

Root Mean Squared Error	Regresión Logística	Árbol de Decisión	Red Neuronal	Media
1 - Vista minable inicial	17,17	6,04	4,69	20,13
2 - Vista minable (con media)	5,06	6,41	3,48	20,13
3 - Sin Datos Externos	18,42	19,67	18,43	20,13
4a - Por Sexo (Hombre)	13,79	4,82	3,79	23,74
4b - Por Sexo (Mujer)	19,67	7,73	5,26	28,33
5a - Por Edad (Menor de 65)	4,44	1,34	1,48	21,57
5b - Por Edad (Mayor de 65)	25,90	12,63	7,84	34,97

7.5. Resultados comentados

Después de mostrar un resumen de los resultados obtenidos tras aplicar las tareas de modelados, vamos a pasar a comentar e intentar explicar cómo se han llegado a dar.

Resultados de 1 - Vista minable inicial

La versión estándar, sin excluir campos o filtrar los datos. Pobre resultado en regresión lineal, pero buenos resultados en árbol de decisión y red neuronal. Aparte de obtener estos resultados se puede tomar como referencia para los siguientes apartados.

Resultados de 2 - Vista minable (con media)

En esta iteración se ha incluido un valor que representa la media de los 10 años para combinación de elementos. Esto hace que las tareas de modelado tomen este valor con mucho peso, puesto que los datos se comportan de una forma bastante lineal a lo largo de los 10 años.

Teniendo en cuenta esta, podemos ver que los resultados son bastante buenos en los 3 casos, pero no aconsejamos tomar este modelo por lo explicado anteriormente.

Resultados de 3 - Sin Datos Externos

Utilizando la versión de la vista minable sin los datos externos que han sido añadidos a posteriori, vemos que la precisión de los modelos empeora mucho acercándose al valor obtenido con la media. Esto nos ayuda a comprobar que el haberlos incluido aporta valor añadido a los mismos datos y a cualquier modelo que se quiera utilizar ya sea con datos de la comunidad valencia o fuera de esta.

Resultados de 4 - Por Sexo

Aquí se han diferenciado los datos por el valor que toma el campo DESC_GENERO y podemos observar que los resultados para los modelos creados en cada caso difieren bastante. Por un lado, para los modelos basados en el valor HOMBRE mejora ligeramente respecto a los resultados obtenidos con “1 - Vista minable inicial” pero cuando comprobamos los resultados tomando el valor MUJER estos empeoran respecto a la referencia tomada, acercándose al valor de la media en la regresión logística.

Se esperaba que mejoraran mucho los resultados en las dos situaciones al diferenciar la muestra y ser conocedores que hay causas de mortalidad que afectan más a un género que a otro.

Resultados de 5 - Por Edad

Aquí se han diferenciado los datos por el valor que toma el campo RANGO_EDAD y podemos observar que los resultados para los modelos creados en cada caso difieren bastante. Por un lado, para los modelos basados en el valor “Menor de 65” mejoran mucho respecto al modelo de referencia, incluso en regresión lineal se consiguen valores por debajo de 5, mejorando el modelo que utilizaba la media, y con valores inferiores a 1,5 en los otros dos.

A pesar de estos resultados resaltamos que este conjunto de datos es inferior en número de registros al resulta con el resto.

Para los modelos generados tomando “Mayor de 65”, los resultados empeoran mucho, dando lugar al peor resultado de todas las pruebas en el caso de regresión lineal. Solo la red neuronal arroja resultados aceptables.

Resultados de Regresión Lineal

Si observamos los resultados obtenidos con la tarea de modelado regresión lineal, podemos ver que es el modelo que aporta los peores resultados en cada ejecución

incluyendo el resultado más bajo de toda la batería de pruebas. Solo se podrían dar por buenos los casos de “con media” y “menor de 65”, el resto no aporta grandes mejoras al modelo de referencia (media).

Resultados de Árbol de Decisión

Al utilizar la tarea de modelado “Árbol de Decisión”, obtenemos unos resultados más robustos que con la tarea anterior. En general mejora en gran medida al modelo de referencia, exceptuando los casos de “Sin Datos Externos” y “Mayor de 65” que ya se han comentado en su apartado. Hay que destacar que en el caso de “Menor de 65” se ha conseguido el mejor resultado de toda la batería de pruebas.

Resultados de Red Neuronal

Los resultados obtenidos con la tarea de modelado “Red Neuronal” han sido los mejores en comparación con los obtenidos de las otras dos. En el caso de “Sin datos externos” sí que pierde casi toda su capacidad de predicción acercándose mucho al modelo de referencia, pero en el caso de “Mayor de 65” a pesar de empeorar su resultado, sigue manteniendo un valor aceptable.

Destacar que el coste de cálculo computacional ha sido muy elevado respecto a las otras dos tareas, pasando de menos de 10 segundos en los casos de regresión lineal y árbol de decisión a 270 segundos

7.6. Despliegue y exportación de modelos

Una vez ya tenemos generados los modelos vamos a explicar cómo exportarlos y representarlos para así poderlos aplicar en otro entorno o en otro programa.

El modelo que vamos a tomar como ejemplo es el resultante de aplicar la tarea de árbol de decisión sobre la versión de la vista minable “Menores de 65”. Hemos elegido este por dos motivos, el primero porque es el modelo que ha obtenido mejor resultado en la batería de pruebas que hemos realizado, y el segundo es que un modelo basado en un árbol de decisión se puede representar gráficamente y líneas de seudocódigo.

Comentar que la regresión lineal también se puede representar con líneas de seudocódigo, pero en cambio una red neuronal no es posible esto, ni una representación gráfica.

Exportar modelo a XML y PMML

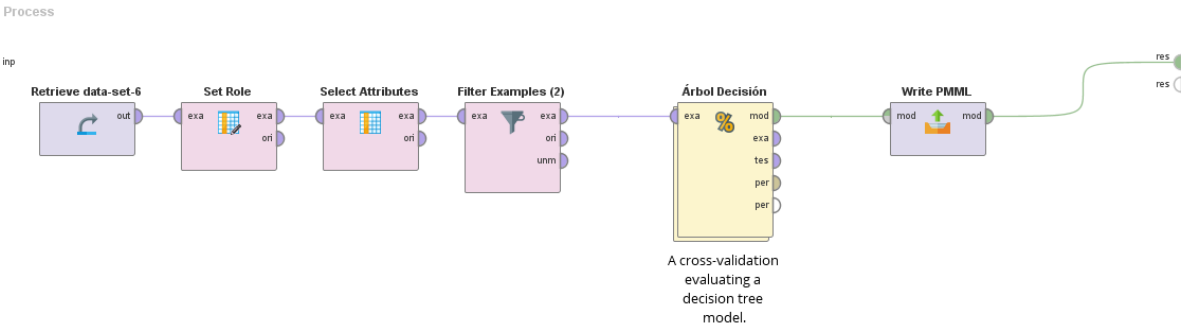
Los modelos creados en RapidMiner se pueden exportar en formato XML, o mejor dicho en PMML que es la variante de XML utilizada para modelos de minería de datos.

Definición PMML [25]: “El Predictive Model Markup Language (PMML) es un lenguaje de marcado de texto XML desarrollado por el Data Mining Group (DMG) para

proveer a las aplicaciones una manera de definir modelos relacionados con los análisis predictivos y la minería de datos para compartir estos modelos entre las aplicaciones PMML."

Fuente: [25]

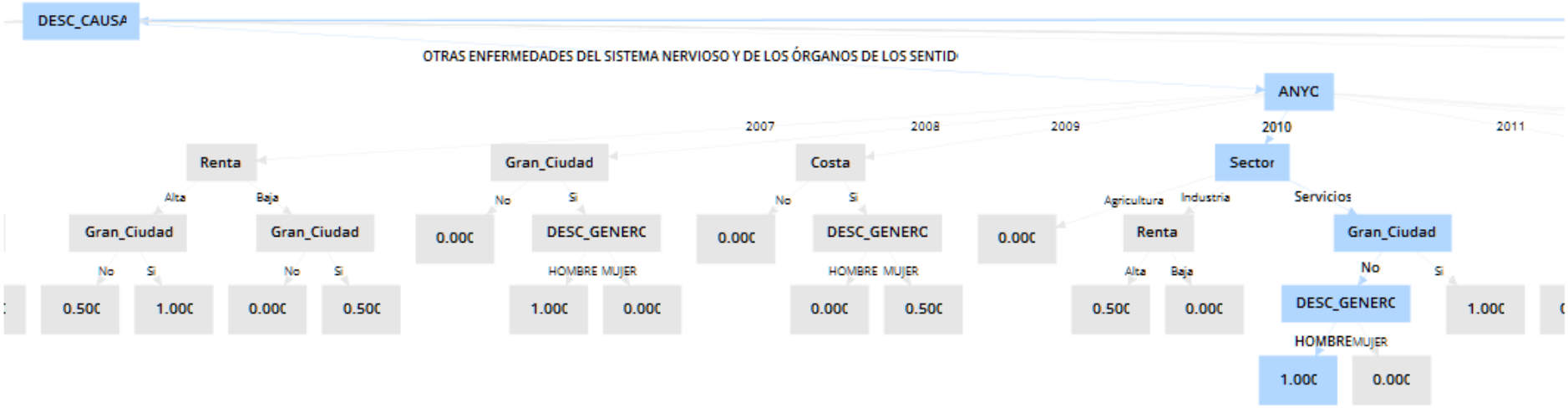
El proceso desde RapidMiner se realiza desde un operador "Write PMML" al cual le llega la salida "Model" de un operador anterior que lo genera y lo escribe en local en formato PMML.



Modificación del proceso ya utilizado para la tarea de árbol de decisión para "Menores de 65" incluyendo el operador "Write PMML"

El propio RapidMiner ofrece varias formas para representar un modelo (solo para los que esta opción es posible como se ha comentado antes) y para el caso de un árbol de decisión éstas son gráficamente y en seudocódigo.

En la siguiente página se muestra una representación gráfica del árbol de decisión generado:


DESC_CAUSA
OTRAS ENFERMEDADES DEL SISTEMA NERVIOSO Y DE LOS ÓRGANOS DE LOS SENTID
ANYC
2007
2008
2009
2010
2011
Renta
Gran_Ciudad
Costa
Sector
Alta
Baja
No
Si
Agricultura
Industria
Servicios
Gran_Ciudad
Gran_Ciudad
0.00C
DESC_GENERC
0.00C
DESC_GENERC
0.00C
Renta
Gran_Ciudad
No
Si
No
Si
HOMBRE
MUJER
HOMBRE
MUJER
Alta
Baja
No
Si
0.50C
1.00C
0.00C
0.50C
1.00C
0.00C
0.00C
0.50C
0.50C
0.00C
DESC_GENERC
1.00C
HOMBRE
MUJER
1.00C
0.00C

Tras ver la imagen anterior vamos a mostrar el seudocódigo que se correspondería con la rama mostrada con las selecciones marcadas.

```
|   DESC_CAUSA = OTRAS ENFERMEDADES DEL SISTEMA NERVIOSO Y DE LOS ÓRGANOS DE LOS SENTIDOS
|   |   ANYO = 2007
|   |   |   Renta = Alta
|   |   |   |   Gran_Ciudad = No: 0.500 {count=4}
|   |   |   |   Gran_Ciudad = Si: 1.000 {count=2}
|   |   |   Renta = Baja
|   |   |   |   Gran_Ciudad = No: 0.000 {count=6}
|   |   |   |   Gran_Ciudad = Si: 0.500 {count=2}
|   |   ANYO = 2008
|   |   |   Gran_Ciudad = No: 0.000 {count=10}
|   |   |   Gran_Ciudad = Si
|   |   |   |   DESC_GENERO = HOMBRE: 1.000 {count=2}
|   |   |   |   DESC_GENERO = MUJER: 0.000 {count=2}
|   |   ANYO = 2009
|   |   |   Costa = No: 0.000 {count=10}
|   |   |   Costa = Si
|   |   |   |   DESC_GENERO = HOMBRE: 0.000 {count=2}
|   |   |   |   DESC_GENERO = MUJER: 0.500 {count=2}
|   |   ANYO = 2010
|   |   |   Sector = Agricultura: 0.000 {count=2}
|   |   |   Sector = Industria
|   |   |   |   Renta = Alta: 0.500 {count=2}
|   |   |   |   Renta = Baja: 0.000 {count=4}
|   |   |   Sector = Servicios
|   |   |   |   Gran_Ciudad = No
|   |   |   |   |   DESC_GENERO = HOMBRE: 1.000 {count=2}
|   |   |   |   |   DESC_GENERO = MUJER: 0.000 {count=2}
|   |   |   |   Gran_Ciudad = Si: 1.000 {count=2}
```

8. Conclusiones

El objetivo primigenio de este estudio, que se basaba en predecir el número de fallecidos en un año para cada causa de mortalidad dadas las características de un municipio de la comunidad valenciana y de su población, se ha alcanzado con la **obtención de una serie de modelos predictivos** que realizan esta tarea. Estos modelos pueden ser aplicados sobre datos recogidos de poblaciones con el fin de mejorar la salud de las personas.

Sobre las preguntas que se hacían en el apartado de **objetivos inicial**, vamos a tratar de responderlas una vez hemos finalizado el estudio.

1 - ¿Existe un cambio en las causas de la mortalidad en los últimos 10 años?

Según los datos observados y los resultados de los modelos **las causas de mortalidad se han mantenido** cuales son las más destacadas y en número de afectados. Los datos han mostrado una linealidad a lo largo del periodo estudiado.

2 - ¿Qué características son más influyentes?

Tras consultar los resultados obtenidos, podemos decir que los aspectos más influyentes son (por este orden):

I. **Edad**. La barrera de los 65 años provoca diferencias en cuáles son las causas más comunes y en su número.
II. **Genero**. El sexo determina que causas afectan más a ese ámbito de la población
III. **Características del municipio de residencia**. Todos los resultados obtenidos han mejorado al sustituir el nombre del municipio por una serie de características comunes al resto.

3 - ¿Es posible a día de hoy predecir cuál será el escenario en los próximos 5 años?

Como hemos comentado en el punto 1, la linealidad de los datos en los 10 años que abarca el estudio, nos dice que, si miramos a un horizonte de 5 años en el futuro, **los modelos construidos seguirán siendo válidos**.

Un riesgo que se ha anunciado y ha estado presente durante todo el estudio es que **los datos no tienen el nivel de detalle deseado para obtener modelos de predicción más incisivos** y que se hayan quedado en aspectos a tener en cuenta más generales. Esto es debido a la naturaleza de los datos, recordemos que estos son de

ámbito medico sobre los cuales siempre existe un límite en cuanto al detalle y se suelen mostrar agrupados para no poder identificar a los pacientes. Aun así, en la fase de transformación de los datos se ha enriquecido a estos para equilibrar este hándicap inicial.

Y en cuanto a la **solución técnica**, se ha desarrollado un proceso que recoge y transforma un conjunto de datos, valida su contenido y construye una vista minable, y por ultimo construye una serie de modelos de predicción y los valida.

La **utilización de RapidMiner para el desarrollo** pensamos que ha sido acertada, la herramienta aporta todo lo necesario y su utilización es fácil e intuitiva. No tiene una curva de aprendizaje elevada como puede tener R.

Este proceso ha sido orientado por la metodología CRISP-DM, aunque por la naturaleza y tamaño del estudio no exprime todo su potencial.

Todo esto no sería posible si no hubieran estado disponibles los **datos utilizados en el portal de transparencia de la Generalitat Valenciana**. El obtener los datos de un portal público nos ha permitido que el proceso de evaluación y limpieza haya sido mínimo y sin desarrollos adicionales para su utilización. A pesar de eso, hemos incluido datos externos para aportarle un valor añadido a los datos.

Una vez obtenidos los modelos predictivos y comparados sus resultados, podemos decir que el obtenido con **redes neuronales** (a pesar de su alto coste computacional) es el que ha arrojado **mejores resultados**. No obstante, los árboles de decisión también muestran un comportamiento bueno en términos de precisión siendo mejores que la media y con la ventaja adicional frente a las redes neuronales de que son comprensibles. Y en cuanto a las variaciones aplicadas a la vista minable, podemos resaltar dos aspectos: el primero que el incluir datos externos a los iniciales ha mejorado la precisión en todos los modelos, y el segundo es que se esperaban unos mejores resultados en todos los resultados en los que se utilizaban subconjuntos de datos (por género, por edad) sobre todo en el caso de los registros para mayores de 65 años, una porción de la población a la cual van muy dirigidos este tipo de estudios.

Y por último comentaremos que los modelos construidos **pueden ser exportables** para ser utilizados con otros conjuntos de datos (sería necesario revisar su estructura) ya sea con datos de **próximos años, o con poblaciones distintas** a la comunidad valenciana.

9. Bibliografía

[1] - https://roulive.com/category/medical/

[2] - https://www.phmk.es/oms-ranking-10-principales-causas-de-muerte/

[3] - https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Causes_of_death_statistics/es

[4] - https://www.ine.es/prensa/edcm_2016.pdf

[5] - https://www.archbronconeumol.org/es-causas-muerte-prediccion-mortalidad-epoc-articulo-S030028961000089X

[6] - https://www.abc.es/sociedad/abci-enfermedades-moriremos-2030-201609141844_noticia.html

[7] - *http://timeofsoftware.com/descubriendo-informacion/*

[8] - http://sedici.unlp.edu.ar/bitstream/handle/10915/29582/Documento_completo.pdf?sequence=1

[9] - http://www.oldemarrodriguez.com/yahoo_site_admin/assets/docs/Documento_CRISP-DM.2385037.pdf

[10] - https://barnraisersllc.com/2018/10/data-mining-process-essential-steps/

[11] - https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

[12] – "Sistemas de información estratégicos" – Parte Data Mining – TEMA I: INTRODUCCIÓN A LA MINERÍA DE DATOS – Autora: María José Ramírez (Departamento de Sistemas Informáticos y Computación Universitat Politècnica de València).

[13] - https://rapidminer.com/resource/gartner-magic-quadrant-data-science-platforms/

[14] - https://es.wikipedia.org/wiki/RapidMiner

[15] - http://www.dadesobertes.gva.es/es/about

[16] - https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176780&menu=resultados&secc=1254736194710&idp=1254735573175

[17] - https://es.wikipedia.org/wiki/Valores_separados_por_comas

[18] - https://www.lasprovincias.es/economia/municipios-renta-bruta-media-habitante-20181016195007-nt.html

[19] - https://www.google.es/maps/?hl=es

[20] - https://es.wikipedia.org/

[21] - https://es.wikipedia.org/wiki/Regresi%C3%B3n_lineal

[22] - http://timeofsoftware.com/descubriendo-informacion/

[23] - *https://es.wikipedia.org/wiki/Validaci%C3%B3n_cruzada*

[24] - https://es.wikipedia.org/wiki/Ra%C3%ADz_del_error_cuadr%C3%A1tico_medio

[25] - https://es.wikipedia.org/wiki/Predictive_Model_Markup_Language

[26] - https://en.wikipedia.org/wiki/SEMMA

[27] - https://www.sngular.com/es/data-science-crisp-dm-metodologia/

[28] - https://www.cs.waikato.ac.nz/ml/weka/