UNIVERSIDAD POLITÉCNICA DE VALENCIA

DEPARTAMENTO DE SISTEMAS INFORMÁTICOS Y COMPUTACIÓN

MÁSTER UNIVERSITARIO EN INGENIERÍA Y TECNOLOGÍA DE SISTEMAS SOFTWARE

CURSO ACADÉMICO 2014-2015

TESIS DE MÁSTER

APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS

EN ACCIDENTES DE TRÁFICO

Para la obtención del

GRADO DE MÁSTER EN INGENIERÍA Y TECNOLOGÍA

DE SISTEMAS SOFTWARE

AUTOR:

MARK MIRKO HASSINGER RODRIGUEZ

TUTOR:

Dra. MARÍA JOSÉ RAMÍREZ QUINTANA.

Universidad Politécnica de Valencia.

COTUTOR:

Dr. CÉSAR FERRI RAMÍREZ.

Universidad Politécnica de Valencia.

VALENCIA, 2015

AGRADECIMIENTOS

Quiero agradecer a Dios, en primer lugar por ser mi guía en todo momento y por brindarme la fuerza y voluntad necesaria para seguir avanzando en los diversos aspectos que engloban estudiar un Máster en el extranjero.

A mis padres y familiares, por brindarme el apoyo desde tan lejos.

A mis asesores de Tesis, la Dra. María José Ramírez y el Dr. César Ferri, por permitirme desarrollar el presente trabajo, por el tiempo invertido y por instruirme para el desarrollo del mismo.

A PRONABEC, por brindarme la oportunidad de realizar los estudios del Máster en esta prestigiosa universidad.

RESUMEN

Actualmente los accidentes de tráfico representan un problema mundial, siendo la undécima causa de muerte en el mundo y se estima que en el año 2020 será la tercera. Todos los años mueren más de 1 millón de personas a causa de estos accidentes, entre conductores, peatones, ciclistas, motociclistas y usuarios del transporte público, además de miles de personas que resultan con discapacidades permanentes. Por tales motivos, reducir el impacto económico y los problemas de salud pública sigue siendo una de las prioridades estratégicas planteadas en los planes de seguridad vial. Uno de los factores clave para alcanzar este objetivo radica en el mejoramiento de la seguridad vial en las carreteras. Por dichas razones, en esta tesis de máster se analiza en profundidad la accidentalidad de estas carreteras.

Existen diferentes enfoques para llevar a cabo el estudio de los accidentes de tráfico. Este estudio se realiza en términos de la gravedad de sus consecuencias.

En la actualidad numerosos investigadores han comenzado a utilizar técnicas que se encuentran dentro del campo de la Minería de Datos. Estas técnicas permiten extraer conocimiento de datos previamente desconocidos, y normalmente, no parten de hipótesis ni requieren un previo conocimiento probabilístico del problema objeto de estudio.

Para realizar estos estudios, se han aplicado diferentes técnicas, pero, particularmente, los Árboles de Decisión son una técnica de Minería de Datos muy apropiada para el estudio de los accidentes de tráfico, por diferentes razones: son fácilmente interpretables, pueden trabajar con grandes volúmenes de datos, provee una estructura sumamente efectiva dentro la cual se puede estimar cuales son las opciones e investigar las posibles consecuencias, y la probabilidad de que suceda, de seleccionar cada una de ellas, y sobre todo, permiten descubrir interacciones entre los datos. Un aspecto relevante, es que esta técnica permite la extracción de Reglas de Decisión del tipo "si-entonces", que son utilizadas para descubrir patrones de comportamientos que ocurren dentro de un conjunto de datos. Estos patrones son de vital importancia, ya que nos ayudan a la comprensión del suceso de un accidente, así como a la identificación de las principales variables que determinan su gravedad. Por ello en esta tesis de máster se utilizan diferentes técnicas de minería de datos con el fin de obtener estos patrones.

Existen algunos estudios similares, donde se analiza la severidad de los accidentes, pero debido a la heterogeneidad de las características de las carreteras, medio ambiente y otros elementos que están relacionados con los accidentes, los estudios se tienen que realizar de forma independiente. En este caso, se utilizan datos de accidentes de tráfico ocurridos en el Reino Unido y en Queensland (Australia).

Con los resultados obtenidos, se puede concluir que los Árboles de Decisión son una herramienta idónea para analizar los accidentes de tráfico, ya que permiten identificar las variables con mayor relevancia en la gravedad del accidente, y que la extracción de las Reglas de Decisión, nos ayudan a descubrir patrones que son de vital interés para los analistas y gestores de seguridad vial, para que posteriormente se puedan realizar planes concretos con el fin de reducir el impacto socio-económico causado por los accidentes de tráfico.

ABSTRACT

Nowadays traffic accidents represent a global problem being the eleventh cause of death in the world and it is believed that it will take the third place in 2020. Every year more than 1 million of people die in accidents, including drivers, pedestrians, cyclists, motorcyclists and users of public transportation, besides of thousands of people who get permanently disability. For these, one of the main strategies set out in the road safety plans is the reduction of the economic impact and the public health problems. One of the key factors for achieving this goal is the improvement of road safety. For these reasons, this master thesis discusses in depth these road accidents.

There are different approaches to carry out the study of traffic accidents. This analysis is conducted in terms of the severity of their consequences.

Currently many researchers have begun to use techniques in the data mining field. These techniques allow to extract knowledge of previously unknown data, and usually are not based on assumptions not even require prior knowledge of the probabilistic problem under study.

Different techniques have been used to perform these studies. Particularly, Decision Trees Learning is a mining data technique very appropriate for the study of traffic accidents for different reasons: they can be easily interpreted, they can work with large amounts of data, they provide a highly effective structure which allows to estimate the options and investigate the possible consequences and the probability of occurrence and then select one of them, and above all, allow to discover interactions between data.

It is remarkable that this technique allows the extraction of "if-then" kind decision rules, which are used to discover patterns of behavior that occur within a data set. These patterns are vital because they help us to understand the event of an accident, as well as identifying the main variables that determine its severity. Therefore in this thesis master different data mining techniques are used to obtain these patterns.

There are some similar studies which analyze the severity of the accidents; however they have to be performed independently because of the heterogeneity of the roads´ characteristics, the environment and other elements that are related to accidents. In this work, data traffic accidents in the UK and Queensland (Australia) were used.

According to the results obtained, it can be concluded that decision tree are a useful tool to analyze traffic identifying the variables most relevant in the severity of the accident. Besides, the extraction of decision rules help us to discover patterns that are of vital interest to analysts and managers of road safety, and from these they can make concrete plans in order to reduce the socio-economic impact caused by traffic accidents.

ÍNDICE GENERAL

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

CAPÍTULO 1.
INTRODUCCIÓN

CAPÍTULO 1.

INTRODUCCIÓN

1.1 Motivación

Los accidentes de tráfico son un problema de primer orden en términos de salud pública, al punto de que son responsables de algo más del 2% de todas las muertes que se producen a nivel global. Actualmente son la undécima causa de muerte en el mundo, pero se estima que en el año 2020 será la tercera. Cada año mueren en el mundo 1,2 millones de personas en accidentes de tráfico, entre conductores, peatones, ciclistas, motociclistas y usuarios del transporte público; un tercio de esas muertes (400.000) corresponden a jóvenes menores de 25 años. A este goteo incesante de vidas truncadas día a día deben sumarse las discapacidades permanentes con las que afrontan el resto de su vida quienes sobreviven a los accidentes (Atlas Mundial de la Salud).

La Organización Mundial de la Salud consideró necesaria su intervención en esta problemática y comenzó a establecer programas de control en todos los países de influencia. Fue entonces cuando empezó a concebirse la accidentalidad como un problema prioritario de salud pública mundial y se reconoció la necesidad de implantar políticas sostenidas de investigación e intervención en los distintos países.

El tránsito es uno de los sistemas más peligrosos y complejos al que las personas se enfrentan cada día, y las lesiones causadas por el tránsito constituyen un importante problema de salud pública que no está recibiendo el tratamiento que merecería por su envergadura y cuya prevención exige una estrategia global.

Sin embargo, la evidencia de muchos países demuestra que se pueden lograr éxitos espectaculares en la prevención de accidentes, mediante la adopción de leyes integrales sobre los factores de riego fundamentales, como son: exceso de velocidad, conducción bajo los efectos del alcohol, o la no utilización del casco de motociclista, del cinturón de seguridad y de sistemas de retención para niños.

Tras esta visión general, es necesario seguir buscando medidas que permitan mejorar la seguridad vial de las carreteras, dirigiendo los esfuerzos a estudiar, analizar y comprender la complejidad de los accidentes de tráfico, **con el objeto de descubrir patrones de comportamiento que ayuden a su prevención**, como a la disminución de las consecuencias que se producen como resultado de los mismos.

Las técnicas de Minería de Datos permiten extraer conocimiento de los datos previamente desconocidos e indistinguibles, y normalmente no parten de hipótesis ni requieren un previo conocimiento probabilístico del problema objeto de estudio.

Además, el uso de técnicas de Minería de Datos en el estudio de los accidentes resulta muy apropiado por la naturaleza de estos fenómenos. Un accidente puede definirse como un evento raro, aleatorio y de múltiples factores siempre precedido por una situación por la que uno o más conductores no pueden hacer frente al entorno de la carretera. Así, cada accidente es el resultado de una cadena de eventos que es, en su totalidad único, pero algunos factores son comunes a varias circunstancias del accidente, y la identificación de estos factores y sus interdependencias puede llevarse a cabo mediante el uso de técnicas de Minería de Datos.

Dentro de los modelos de Minería de Datos existen diferentes técnicas, y cada una de ellas, posee sus ventajas y desventajas, y la elección de la misma dependerá del objetivo perseguido por el

analista; siendo las redes Neuronales Artificiales, las Redes Bayesianas y los Árboles de Decisión, las más utilizadas en el campo de la seguridad vial (López G., 2013).

Particularmente, los Árboles de Decisión son una técnica de Minería de Datos muy apropiada para el estudio de los accidentes de tráfico, ya que, teniendo en cuenta sus ventajas y limitaciones, constituyen uno de los modelos más utilizados en aprendizaje supervisado y en aplicaciones de Minería de Datos (Gehke et al., 1999).

Una de las ventajas más importantes que poseen los Árboles de Decisión, se puede destacar que permiten la extracción de reglas de decisión del tipo "SI-ENTONCES". Estas reglas son fácilmente comprensibles por los gestores de seguridad vial y pueden ser usadas para descubrir determinados patrones de comportamiento que ocurren dentro de un conjunto de datos. Estos patrones pueden ayudar a la comprensión del suceso de un accidente, a la identificación de las principales variables que determinan su gravedad, así como al establecimiento de actuaciones concretas por parte de las Administraciones encargadas con el fin de mejorar la seguridad vial de las carreteras analizadas (carreteras del Reino Unido y Queensland, Australia).

Existen algunos estudios similares, descritos en el epígrafe 1.3, donde se analiza la severidad de los accidentes, pero debido a la heterogeneidad de las características de las carreteras, medio ambiente y otros elementos que están relacionados con los accidentes, los estudios se tienen que realizar de forma independiente.

Finalmente, para realizar los estudios anteriormente mencionados, utilizaremos como herramienta principal "R" (The R Foundation, 2015), que es un lenguaje de programación especializado en el análisis de datos estadístico y minería de datos; y como herramienta secundaria, "Weka" (Waikato Environment for Knowledge Analysis (Witten & Frank, 2005), en español «entorno para análisis del conocimiento de la Universidad de Waikato»), que es una plataforma de software para el aprendizaje automático y la minería de datos desarrollado en la Universidad de Waikato, Nueva Zelanda.

1.2 Trabajos relacionados

En el 2001, Abdelwahab, H. y Abdel-Aty, M., presentaron un estudio titulado "DEVELOPMENT OF ARTIFICIAL NEURONAL NETWORK MODELS TO PREDICT DRIVER INJURY SEVERITY IN TRAFFIC ACCIDENTS AT SIGNALIZED INTERSECTIONS" (Abdelwahab & Abdel-Aty, 2001), en donde se examinaron la relación existente entre la gravedad de las lesiones del conductor y las características del vehículo, carretera y entorno. Se utilizaron dos paradigmas conocidos de redes neuronales, el Perceptrón Multicapa (MLP) y la Teoría de la Resonancia Adaptativa Difusa (ART) (Flores López & Fernández Fernández, 2008). Se utilizaron datos de accidentes de 1997 ocurridos en el área de la Florida Central, que consta de Orange, Osceola y Seminole. El análisis de centró en los accidentes ocurridos entre dos vehículos en las intersecciones señalizadas. La red neuronal MLP obtuvo un mejor rendimiento de 65.6% y 60.4% para las fases de entrenamiento y prueba, respectivamente. Los resultados mostraron que las intersecciones rurales son más peligrosas en términos de lesiones de gravedad que las lesiones controladas por intersecciones urbanas. Además los conductores de sexo femenino son más probables de experimentar una lesión grave que los conductores masculinos. La velocidad aumenta la probabilidad de gravedad de la lesión. Los conductores que poseen faltas son más propensos a experimentar lesiones graves que los que no lo tienen. El uso de un cinturón de seguridad reduce el riesgo de sufrir lesiones graves. El tipo de vehículo juega un papel importante en la gravedad de la lesión. Los conductores de turismo son más propensos a experimentar un mayor nivel de gravedad de las lesiones que los conductores de furgonetas o camionetas. Finalmente, los conductores que son impactados por el lado del vehículo en donde ellos se encuentran, la severidad es mayor.

En el 2011, De Oña, J., Oqab, R. y Calvo, F., presentaron un estudio titulado "Analysis of traffic accident injury severity on Spanish rural highways using Bayesian networks" (de Oña, Oqab Mujalli, & Calvo, 2011). Se utilizaron redes Bayesianas (BNs) para clasificar los accidentes de tráfico en función a la gravedad de la lesión. Las redes Bayesianas son capaces de realizar predicciones sin necesidad de suposiciones previas y se utilizan para hacer representaciones gráficas de sistemas complejos con componentes interrelacionados. En el estudio se analizaron 1536 accidentes ocurridos en las carreteras rurales de España, donde se utilizaron 18 variables que representan los factores que contribuyen a la ocurrencia del accidente y se construyeron 3 diferentes BNs que clasificaban los accidentes en levemente y gravemente heridos o muertos. Las variables que mejor identificaban los factores que se asocian con un accidente de muerte o lesiones graves (tipo de accidente, edad del conductor, iluminación y el número de lesiones) fueron identificadas por inferencia.

En el año 2013, Griselda López, presentó una tesis doctoral titulada "Análisis de la Severidad de Accidentes de Tráfico, utilizando Técnicas de Minería de Datos" (López Maldonado, 2013); en donde se propone un estudio en profundidad de la accidentabilidad de las carreteras de Granada, España. Existen diferentes enfoques para llevar a cabo el estudio de los accidentes de tráfico. La investigación se realizó en términos de la gravedad de sus consecuencias. Inicialmente se utilizan árboles de decisión para extraer patrones de accidentes cuya severidad es Grave usando las reglas de decisión. No obstante, la principal limitación de las reglas de decisión que se extraen es que son dependientes de la estructura del árbol, de modo que pueden existir ciertos patrones de accidentes que no sean detectados, por lo que no se estaría obteniendo todo el conocimiento posible de la base de datos de accidentes analizada. Por tal motivo, en la investigación se plantea además utilizar un nuevo método para la extracción de reglas de decisión, llamado *Information Root Node Variation*, que mejora el desempeño de los árboles de decisión, permitiendo extraer todo el conocimiento existente de la base de datos analizada. Se obtuvo una precisión del 55.57 %. Con los resultados obtenidos en la investigación se demuestra que los árboles de decisión son

una herramienta adecuada para analizar los accidentes de tráfico de un modo sencillo y fácilmente comprensible para los analistas de la seguridad vial.

El mismo año, el 2013, De Oña, J., López, G., Mujalli, R. y Calvo, F., presentaron un nuevo estudio titulado "Analysis of traffic accidents on rural highways using Latent Class Clustering and Bayesian Networks" (de Oña, López, Mujalli, & Calvo, 2013). Utilizaron *Latent Class Clustering* (LCC) (Depaire, Wets, & Vanhoof, 2008), como herramienta preliminar para la segmentación de 3229 accidentes ocurridos en las carreteras rurales de Granada (España), entre los años 2005 y 2008. Luego se utilizaron las redes Bayesianas para identificar los principales factores que intervienen en la gravedad de los accidentes en toda la base de datos y en los clusters obtenidos mediante LCC. Los resultados mostraron que el uso combinado de ambas técnicas es muy interesante, ya que revelaron información adicional que no se hubieran obtenido utilizando solamente las redes Bayesianas. Las redes Bayesianas fueron utilizadas para obtener las variables relacionadas con los accidentes con heridos graves o muertes. Las variables tipo de accidente y la distancia de visibilidad han sido identificadas relacionadas en todos los casos analizados; otras variables tales como el tiempo, ocupantes involucrados o la edad son identificados en toda la base de datos y sólo en un clúster, mientras que las variables vehículos implicados, número de lesiones, factores atmosféricos, marcas en el pavimento y el ancho del pavimento se identifican sólo en un clúster.

1.3 Objetivos

El principal objetivo de esta tesis de Máster consiste en la identificación de patrones de accidentes cuya severidad como consecuencia del mismo es de carácter "Grave", mediante el uso de reglas de decisión extraídas de Árboles de Decisión. Para poder alcanzar este objetivo, se han desarrollado varios objetivos específicos, tales como: la unión de data sets para tener toda la información unificada, la limpieza de los datos, el filtro de los datos que pertenecen a aquellas circunstancias en donde los accidentes mayormente son graves, la identificación de las variables trascendentes que afectan a la gravedad de los accidentes, el balanceo de las clases (variable a predecir, en este caso, severidad del accidente) y, finalmente, la aplicación de varias técnicas de aprendizaje supervisado, entre ellas, Árboles de Decisión para extraer conocimiento existente en la base de datos, en forma de reglas de decisión. También uno de los objetivos consiste en el análisis estadístico de los accidentes ocurridos, que nos dará información relevante sobre las tendencias que éstos poseen y bajo qué condiciones ocurren con más frecuencia.

Los patrones identificados deben ser comprensibles por aquellas personas que están encargadas de velar por la seguridad vial de las carreteras, de tal forma, que se puedan realizar actividades concretas para prevenir la severidad de los accidentes, motivo por el cual, aun aplicando diversas técnicas de minería de datos nos centramos en el análisis en los árboles de decisión.

1.4 Metodología

Para la idónea extracción de conocimiento, utilizaremos las siguientes fases (Hernández J., Ramírez M. y Ferri C., 2004):

Fase de integración y recopilación

Se determinan las fuentes de información que pueden ser útiles y dónde conseguirlas. Lo común es que los datos necesarios para realizar el estudio pertenezcan a diferentes organizaciones o a distintos departamentos de una misma entidad. Las fuentes son diversas, organizaciones privadas y públicas. Esto representa un reto, ya que cada fuente de datos usa diferentes formatos de registro, diferentes grados de agregación de los datos, diferentes claves primarias, diferentes tipos de error, etc. Lo primero, por lo tanto es integrar todos estos datos.

Fase de selección, limpieza y transformación

Dado que los datos provienen de diversas fuentes, pueden contener valores erróneos o faltantes. En esta fase se eliminan o corrigen los datos incorrectos y se decide la estrategia a seguir con los datos incompletos. Además, se proyectan los datos para considerar únicamente aquellas variables o atributos que van a ser relevantes, con el objetivo de hacer más fácil la tarea propia de minería y para que los resultados de la misma sean más útiles. La selección incluye tanto una criba o fusión horizontal (filas / registros) como vertical (columnas / atributos).

Fase de minería de datos

La fase de minería de datos es la más característica del proceso de descubrimiento de conocimiento, y por esta razón, muchas veces se utiliza esta fase para nombrar todo el proceso. El objetivo de esta fase es producir nuevo conocimiento que pueda utilizar el usuario. Esto se realiza construyendo un modelo basado en los datos recopilados para este efecto. El modelo es una descripción de los patrones y relaciones entre los datos que pueden usarse para hacer predicciones, para entender mejor los datos o para explicar situaciones pasadas.

Fase de evaluación e interpretación

Se mide la calidad de los patrones descubiertos por un algoritmo de minería de datos, y si es necesario se vuelve a las fases anteriores para una nueva iteración. Esto incluye resolver posibles conflictos con el conocimiento que se disponía anteriormente. Idealmente, los patrones descubiertos deben tener al menos las siguientes cualidades: ser precisos e interesantes (útiles y novedosos).

Fase de difusión, uso y monitorización

Una vez construido y validado, el modelo puede usarse principalmente con dos finalidades: para que un analista recomiende acciones basándose en el modelo y en sus resultados, o bien para aplicar el modelo a diferentes conjuntos de datos.

CAPÍTULO 2.
CONCEPTOS PREVIOS

CAPÍTULO 2.

CONCEPTOS PREVIOS

2.1 Minería de Datos

La minería de datos es un término relativamente moderno que integra numerosas técnicas de análisis de datos y extracción de modelos. Aunque se basa en varias disciplinas, algunas de ellas más tradicionales, se distingue de ellas en la orientación más hacia el fin que hacia el medio, hecho que permite nutrirse de todas ellas sin prejuicios. Y el fin lo merece: ser capaces de extraer patrones, de describir tendencias y regularidades, de predecir comportamientos y, en general, de sacar partido a la información computarizada que nos rodea hoy en día, generalmente heterogénea y en grandes cantidades, permite a los individuos y a las organizaciones comprender y modelar de una manera más eficiente y precisa el contexto en el que deben actuar y tomar decisiones.

Pese a la popularidad del término, la minería de datos es sólo una etapa, si bien la más importante, de lo que se ha venido llamando el proceso de extracción del conocimiento a partir de datos. Este proceso consta de varias frases e incorpora muy diferentes técnicas de los campos del aprendizaje automático, la estadística, las bases de datos, los sistemas de toma de decisión, la inteligencia artificial y otras áreas de la informática y de la gestión de la información.

Witten & Frank (Witten & Frank, 2005), definen la minería de datos como el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos. Es decir, la tarea fundamental de la minería de datos es encontrar modelos inteligibles a partir de los datos. Para que este proceso sea efectivo debería ser automático o semi-automático (asistido) y el uso de los patrones descubiertos debería ayudar a tomar decisiones más seguras que reporten, por tanto, algún beneficio a la organización.

2.2 Técnicas de Minería de Datos

La minería de datos ha dado lugar a una paulatina sustitución del análisis de datos *dirigido a la verificación* por un enfoque de análisis de datos *dirigido al descubrimiento del conocimiento*. La principal diferencia entre ambos se encuentra en que en el último se descubre información sin necesidad de formular previamente una hipótesis. La aplicación automatizada de algoritmos de minería de datos permite detectar fácilmente patrones en los datos, razón por la cual esta técnica es mucho más eficiente que el análisis dirigido a la verificación cuando se intenta explorar datos procedentes de repositorios de gran tamaño y complejidad elevada. Dichas técnicas emergentes se encuentran en continua evolución como resultado de la colaboración entre campos de investigación tales como bases de datos, reconocimiento de patrones, inteligencia artificial, sistemas expertos, estadística, visualización, recuperación de información, y computación de altas prestaciones.

Los algoritmos de minería de datos se clasifican en dos grandes categorías: supervisados o predictivos y no supervisados o de descubrimiento del conocimiento (Weiss & Indurkhya, 1998).

Los algoritmos **supervisados o predictivos** predicen el valor de un atributo (*etiqueta*) de un conjunto de datos, conocidos otros atributos (*atributos descriptivos*). A partir de datos cuya etiqueta se conoce se induce una relación entre dicha etiqueta y otra serie de atributos. Esas relaciones sirven para realizar la predicción en datos cuya etiqueta es desconocida. Esta forma de trabajar se conoce como *aprendizaje supervisado* y se desarrolla en dos fases: Entrenamiento

(construcción de un modelo usando un subconjunto de datos con etiqueta conocida) y prueba (prueba del modelo sobre el resto de los datos).

Dentro del aprendizaje supervisado existen dos tipos de modelos:

1. **Modelo de Clasificación**. Cuando el atributo o variable de respuesta a predecir es cualitativa; que es aquí en donde nos centraremos para desarrollar el estudio, ya que nuestra finalidad radicará en predecir la gravedad del accidente: Fatal o No Fatal.
2. **Modelo de Regresión**. Cuando la variable de respuesta y las variables explicativas son todas ellas cuantitativas. Si sólo disponemos de una variable explicativa hablamos de *regresión simple*, mientras que si disponemos de varias variables explicativas se trata de un problema de *regresión múltiple.*

Cuando una aplicación no es lo suficientemente madura no tiene el potencial necesario para una solución predictiva, en ese caso hay que recurrir a los métodos **no supervisados o de descubrimiento del conocimiento** que descubren patrones y tendencias en los datos actuales (no utilizan datos históricos). El descubrimiento de esa información sirve para llevar a cabo acciones y obtener un beneficio (científico o de negocio) de ellas. En la tabla siguiente se muestran algunas de las técnicas de minería de ambas categorías.

Supervisados	**No supervisados**
Árboles de decisión	Detección de desviaciones
K vecinos más cercanos	Segmentación
Naive Bayes	Agrupamiento (“clustering”)
Máquinas de soporte vectorial	Reglas de asociación
Bosques aleatorios	Patrones secuenciales
Potenciación	K-means

Tabla 1. Clasificación de las técnicas de minería de datos

2.2.1 Técnicas de Clasificación

2.2.1.1 Árboles de Decisión (ADD)

Los sistemas de aprendizaje basados en árboles de decisión son quizás el método más fácil de utilizar y de entender. Un árbol de decisión es un conjunto de condiciones organizadas en una estructura jerárquica, de tal manera que la decisión final a tomar se puede determinar siguiendo las condiciones que se cumplen desde la raíz del árbol hasta alguna de sus hojas. Simplemente, el algoritmo va construyendo el árbol (desde el árbol que sólo contiene la raíz) añadiendo particiones (tests sobre los valores de una atributo seleccionado de acuerdo a un criterio de partición) y los hijos resultantes de cada partición. Lógicamente, en cada partición, los ejemplos se van dividiendo entre los hijos. Finalmente, se llega a la situación en la que todos los ejemplos que caen en los nodos inferiores son de la misma clase y esa rama ya no sigue creciendo. La única condición que hay que exigir es que las particiones al menos separen ejemplos en distintos hijos, con lo que la cardinalidad de los nodos irá disminuyendo a medida que se desciende el árbol.
Los dos puntos más importantes para que el algoritmo anterior funciones bien son: Particiones a considerar y Criterio de selección de particiones.

2.2.1.2 K vecino más cercano (k-NN)

Una forma práctica y de fácil aplicación para predecir o clasificar un nuevo dato, basado en observaciones conocidas o pasadas, es la técnica del vecino más cercano. A manera de ejemplo, el caso de un médico que está tratando de predecir el resultado de un procedimiento quirúrgico puede predecir que el resultado de la cirugía del paciente será aquel del paciente más parecido que conoce, que haya sido sometido al mismo procedimiento. Esto puede resultar un tanto extremo, ya que un solo caso similar en el cual la cirugía falló puede influir de manera excesiva sobre otros casos, ligeramente menos similares, en los cuales la cirugía fue un éxito. Por esta razón el método del vecino más cercano se generaliza a uso de los k vecinos más cercanos.
Esta técnica se basa, simplemente, en "recordar" todos los ejemplos que se vieron en la etapa de entrenamiento. Cuando un nuevo dato se presenta al sistema de aprendizaje, este se clasifica según el comportamiento de los datos más cercanos (Aha et al., 1991; Moreno, 2004).

2.2.1.3 Redes Bayesianas (RBs)

Una *Red Bayesiana* es un modelo probabilístico que relaciona un conjunto de variables aleatorias mediante un grafo dirigido acíclico en el que se representan las variables aleatorias y las relaciones de probabilidad que existan entre ellas, lo que permite conseguir soluciones a problemas de decisión en casos de incertidumbre.

Una red bayesiana es una representación ilustrada de dependencias para razonamiento probabilístico, en la cual los nodos representan variables aleatorias y los arcos simbolizan relaciones de dependencia directa entre las variables.

Las redes bayesianas organizan un caso problema mediante un conjunto de variables y las relaciones de dependencia entre ellas. Dado este modelo, se puede hacer inferencia bayesiana; es decir, estimar la probabilidad posterior de las variables no conocidas, en base a las variables conocidas. Estos modelos bayesianos poseen diferentes aplicaciones para diagnóstico, clasificación y decisión que brinde información importante en cuanto a cómo se relacionan las variables, las cuales pueden ser interpretadas como relaciones de causa efecto.

2.2.1.4 Máquinas de Soporte vectorial (SVM)

Según Martínez, Tolmos & Hernández-March (Heras Martínez, Tolmos, & Hernández-March, 2010), se define a las SVM como una técnica de clasificación que ha demostrado sobradamente su capacidad de resolución frente a problemas de elevado grado de complejidad. Diseñada en principio para tratar problemas de clasificación binarios (en dos grupos), se trata de una máquina de aprendizaje que implementa la siguiente idea: cuando no sea posible separar los datos en el espacio de entrada con un hiperplano lineal, trasladar, mediante una aplicación no lineal, los vectores de entrada a un nuevo espacio de dimensión más alta. En este nuevo espacio se construirá una superficie de decisión lineal. Las especiales propiedades que poseerá esta superficie garantizarán que la capacidad de generalización de la máquina de aprendizaje sea alta. Aunque esta idea se empleó en los primeros experimentos para datos que podían separarse sin errores, se puede extender para el caso no separable con notable éxito. La parte conceptual del problema la resolvió Vapnik (uno de los autores de la Teoría del Aprendizaje) para el caso de *hiperplanos óptimos* para clases separables. En este contexto, Vapnik definió un hiperplano óptimo como una función de decisión lineal con el margen de separación máximo entre los vectores de las dos clases. Se observó entonces que para construir tal hiperplano, uno sólo debía tener en cuenta una cantidad pequeña de los datos de entrenamiento, los llamados *vectores soporte*, quienes determinaban ese margen.

2.2.1.5 Bosques Aleatorios (RF, Random Forest)

Orozco y otros (Orozco Guillén, y otros, 2010), definen a *Random Forest* como un algoritmo compuesto por numerosos árboles de clasificación, en donde se definen una cantidad de árboles a desarrollar y una cantidad de atributos *m* tal que sea menor a la cantidad total de atributos. Entre los árboles se reparten *k* patrones con reemplazo y se desarrollan los árboles, el resto de los patrones son usados para la prueba. Al desarrollar cada nodo se eligen *m* atributos y se determina el mejor atributo para desarrollar el nodo. Para el entrenamiento los patrones son repartidos aleatoriamente con repetición entre cada árbol.

El método RF está siendo utilizado de una manera extensiva en multitud de campos de investigación, tanto para seleccionar aquellas variables con mayor poder clasificador de entre un conjunto, como para clasificar conjuntos de datos. En RF cada árbol depende de los valores de un vector aleatorio probado independientemente y con la misma distribución para cada uno de estos.

En RF cada árbol individual se explora de una manera particular:

1. Dado un conjunto de datos de entrenamiento N, se toman N muestras aleatorias con repetición como conjunto de entrenamiento.
2. Para cada nodo del árbol, se determinan M variables de entrada, y se determina "m" << M, para cada nodo, seleccionando *m* variables aleatorias. La variable más relevante elegida al azar se usa en el nodo. El valor de m se mantiene constante durante la expansión del bosque
3. Cada árbol es desarrollado hasta su expansión máxima, nunca se poda.

2.2.1.6 Potenciación (AdaBoost)

El algoritmo AdaBoost propone entrenar iterativamente una serie de clasificadores base, de tal modo que cada nuevo clasificador preste mayor atención a los datos clasificados erróneamente por los clasificadores anteriores, y combinarlos de tal modo que se obtenga un clasificador con elevadas prestaciones. Para ello, durante una serie de iteraciones entrena un clasificador que implementa una función asignándole un peso de salida, y lo añade al conjunto de modo que la salida global del sistema se obtenga como combinación lineal ponderada de todos los clasificadores base. Para conseguir que cada nuevo clasificador preste mayor atención a los datos más erróneos se emplea una función de énfasis que pondera la importancia de cada dato durante el entrenamiento del clasificador.

2.2.1.7 Redes Neuronales Artificiales (RNA)

Son generalizaciones de modelos estadísticos clásicos cuya estructura y operación está inspirada en las redes neuronales biológicas. Una RNA puede verse como un grafo dirigido formado por un conjunto interconectado de elementos simples de procesamiento, unidades o nodos. La capacidad de procesamiento de la red se almacena en las fuerzas de conexión entre las unidades, o pesos, obtenidos por un proceso de aprendizaje a partir de un conjunto de patrones de entrenamiento. El objetivo de las RNA es conseguir que la red aprenda automáticamente las propiedades deseadas a partir de un conjunto de datos de entrada (suficientemente significativo). Se componen de unidades simples llamadas neuronas. Cada neurona tiene asociada una función matemática (función de transferencia) que genera la salida de la neurona a partir de las señales de entrada. La entrada de la función es la suma de todas las señales de entrada por el peso asociado a la conexión de entrada de la señal. De este modo, la función de transferencia es la relación entre la señal de salida y de entrada.

Las ventajas de las RNA es que permiten modelar problemas complejos en los que puede haber interacciones no lineales entre variables. El principal inconveniente es que tiene el efecto de "caja negra". Los datos entran en la "caja negra" y se obtienen las predicciones, pero no se revela normalmente la naturaleza de las relaciones entre las variables independientes y dependientes. En una RNA el conjunto de los pesos determina el conocimiento de esa red y permiten resolver el problema para el que ha sido entrenada. Sin embargo, el conocimiento de la red expresado de este modo (en forma de pesos), impide la inteligibilidad de las asignaciones de clase que se realizan. Estos pesos son ocultos y no pueden ser modificados por el operador. Es decir, de las variables descriptoras no se extrae nuevo conocimiento para el usuario, sino que esa extracción de conocimiento es interna de la red y no revierte en el usuario salvo en la asignación de clases realizada. Otro de los inconvenientes de las RNA es que el modelo aprendido es difícilmente comprensible, y requieren gran cantidad de datos para su entrenamiento.

2.3 Métodos de Evaluación

Para entrenar y probar un modelo se parten los datos en dos conjuntos: el conjunto de entrenamiento (*training set*) y el conjunto de prueba o de test (*test set*). Esta separación es necesaria para garantizar que la validación de la precisión del modelo es una medida independiente. Si no se usan conjuntos diferentes de entrenamiento y prueba, la precisión del modelo será sobreestimada, es decir, tendremos estimaciones muy optimistas.

En los modelos predictivos, el uso de esta separación entre entrenamiento y prueba es fácil de interpretar. Por ejemplo, para una tarea de clasificación, después de generar el modelo con el conjunto de entrenamiento, éste se puede usar para predecir la clase de los datos de prueba (*test*). Entonces, la razón de precisión (o simplemente precisión), se obtiene dividiendo el número de clasificaciones correctas por el número total de instancias. La precisión es una buena estimación de cómo se comportará el modelo para datos futuros similares a los de test. Esta forma de proceder no garantiza que el modelo sea correcto, sino que simplemente indica que si usamos la misma técnica con una base de datos con datos similares a los de prueba, la precisión media será bastante parecida a la obtenida con éstos.

El método de evaluación más básico, la **validación simple**, reserva un porcentaje de la base de datos como conjunto de prueba, y no lo usa para construir el modelo. Este porcentaje suele variar entre el cinco por ciento y el 50 por ciento. La división de los datos en estos dos grupos debe ser aleatoria para que la estimación sea correcta.

Si tenemos una cantidad no muy elevada de datos para construir el modelo, puede que no podamos permitirnos el lujo de reservar parte de los mismos para la etapa de evaluación. En estos casos se usa un método conocido como validación cruzada (*cross validation*). Los datos se dividen aleatoriamente en dos conjuntos equitativos con los que se estima la precisión predictiva del modelo. Para ello, primero se construye un modelo con el primer conjunto y se usa para predecir los resultados en el segundo conjunto y calcular así un ratio de error (o de precisión). A continuación, se construye un modelo con el segundo conjunto y se usa para predecir los resultados del primer conjunto, obteniéndose un segundo ratio de error. Finalmente, se construye un modelo con todos los datos, se calcula un promedio de los ratios de error y se usa para estimar mejor su precisión.

El método que se usa normalmente es **la validación cruzada con n pliegues** (*n-fold cross validation*). En este método los datos se dividen aleatoriamente en n grupos. Un grupo se reserva para el conjunto de prueba y con los otros n-1 restantes (juntando todos sus datos) se construye un modelo y se usa para predecir el resultado de los datos del grupo reservado. Este proceso se repite n veces, dejando cada vez un grupo diferente para la prueba. Esto significa que se calculan

n ratios de error independientes. Finalmente, se construye un modelo con todos los datos y se obtienen sus ratios de error y precisión promediando los n ratios de error disponibles.

2.4 Técnicas de Selección de Atributos

Weka (plataforma de software para el aprendizaje automático y la minería de datos, descrito en el epígrafe 2.7) incorpora una gran variedad de técnicas de selección de atributos que tratan de explorar qué subconjuntos de atributos son los que mejor pueden clasificar la clase de la instancia. Esta selección de atributos tiene dos componentes:

1. Un método de evaluación que determina la calidad del conjunto de atributos para discriminar la clase. Se pueden distinguir dos categorías de métodos de evaluación, en la primera se utiliza directamente un clasificador específico para medir la calidad del subconjunto de atributos a través de la tasa de error del clasificador. Estos métodos necesitan un proceso completo de entrenamiento y evaluación en cada caso de búsqueda, por eso resultan de un elevado coste computacional. La alternativa es la utilización de métodos que no utilizan un clasificador específico, por ejemplo el método ***CfsSubsetEval*** que se encuentra implementado en ***Weka*** y que se basa en calcular la correlación de la clase con cada atributo, y eliminar atributos que tienen una correlación muy alta como atributos redundantes.

2. Un método de búsqueda que determina la forma de realizar la búsqueda de conjuntos. La evaluación exhaustiva de todos los posibles subconjuntos se convierte en un problema combinatorio inabordable cuando el número de atributos es elevado. Por tanto, se necesitan estrategias de búsqueda más eficientes. Una de las estrategias más efectiva, por su rapidez, es el **ForwardSelection**, que se basa en elegir primero el mejor atributo, y realizar un proceso iterativo de ir añadiendo atributos que aporten más información hasta llegar a la situación en la que añadir un nuevo atributo empeora la situación.

2.5 Problemas presentes en los datos

El trabajo (Moreo, Rodríguez, Sicilia, Riquelme, & Ruiz, 2009), afirma que con frecuencia los datos presentan diferentes problemas que dificultan la labor de los clasificadores y disminuyen la calidad de la clasificación realizada. Los problemas más destacables son, el ruido, es un problema derivado de la naturaleza de los datos consistente en el gran parecido que presentan entre sí datos pertenecientes a clases distintas o datos erróneos; el solapamiento entre clases ocurre cuando datos de clases distintas ocupan un espacio común debido a que algunos de los atributos de dichas clases comparten un mismo rango de valores; y el desbalanceo de clases ocurre cuando el número de instancias de cada clase es muy diferente.

En estas circunstancias los clasificadores presentan una tendencia de clasificación hacia la clase mayoritaria, minimizando de ésta manera el error de clasificación y clasificando correctamente instancias de clase mayoritaria en detrimento de instancias de clase minoritaria. Entre las medidas que podemos tomar para el tratamiento del ruido se encuentran usar algoritmos tolerantes al ruido; usar algoritmos que reduzcan el ruido filtrando las instancias ruidosas y usar algoritmos que corrijan las instancias que generan el ruido. Para el tratamiento del solapamiento podemos utilizar algoritmos de “limpieza” que reduzcan las áreas de solapamiento.

Finalmente hay más opciones para el tratamiento del tratamiento del desbalanceo, entre las que se encuentran:

- *Sampling*: Consiste en balancear la distribución de las clases añadiendo ejemplos de las clase minoritaria. A ésta técnica se le denomina *oversampling*. Algunos de los algoritmos más representativos son SMOTE, y *Resampling.* También es posible realizar lo contrario: eliminar ejemplos de la clase mayoritaria. Ésta técnica se conoce como *undersampling*. Un algoritmo bastante representativo es RUS [3]. Ambas técnicas tienen ventajas e inconvenientes. Entre los inconvenientes del *undersampling* está la pérdida de información que se produce al eliminar instancias de la muestra. Sin embargo tiene la ventana de que reduce el tiempo de procesado del conjunto de datos. *Oversampling* tiene la ventaja de no perder información pero puede repetir muestras con ruido además de aumentar el tiempo necesario para procesar el conjunto de datos.
    - *Oversampling*:
        - ✓ SMOTE: Genera nuevas instancias de la clase minoritaria interpolando los valores de las instancias minoritarias más cercanas a una dada.
        - ✓ *Resampling*: Duplica al azar instancias de la clase minoritaria
        - ✓ ROSE: Crea nuevos registros aleatoriamente.
    - *Undersampling*:
        - ✓ *Random undersampling*: Elimina al azar instancias de la clase mayoritaria
        - ✓ *Tomek Links*: Elimina sólo instancias de la clase mayoritaria quesean redundantes o que se encuentren muy cerca de instancias de la clase minoritaria.
        - ✓ *Wilson Editing*: También conocido como ENN (Editing Nearest Neighbor) elimina aquellas instancias donde la mayoría de sus vecinos pertenecen a otra clase.
- *Boosting*: consiste en asociar pesos a cada instancia que se van modificando en cada iteración del clasificador. Inicialmente todas las instancias tienen el mismo peso y después de cada iteración, en función del error cometido en la clasificación se reajustan los pesos con objeto de reducir dicho error:
    - ✓ *AdaBoost*: Implementa el algoritmo de *Boosting* descrito. En cada iteración *AdaBoost* genera nuevas instancias utilizando *Resampling*
    - ✓ SMOTEBoost: Es similar a AdaBoost pero usa SMOTE en lugar del Resampling para generar nuevas instancias.
    - ✓ RUSBoost: Aplica AdaBoost pero en cada iteración utiliza RUS (*Random Undersampling*) que reducen el tamaño de la muestra de datos y simplifican y aumentan el rendimiento del clasificador.

Posteriormente utilizaremos una de las técnicas más utilizadas por su potencia y efectividad: SMOTE; y ROSE para balancear los datos; ya que como se podrá ver más adelante, el data set presenta problemas de desbalanceo de datos.

2.6 Herramientas de Minería de Datos

Dos de las herramientas más potentes para el tratamiento de datos son “R” y “Weka”. A continuación se describe brevemente estas dos potentes herramientas.

En The Comprehensive R Archive Network (R Development Core Team, 2000) se define a **R** como un conjunto integrado de programas para manipulación de datos, cálculo y gráficos. Entre otras características dispone de: almacenamiento y manipulación efectiva de datos, operadores para cálculo sobre variables indexadas (Arrays), en particular matrices, una amplia, coherente e integrada colección de herramientas para análisis de datos, posibilidades gráficas para análisis de datos, que funcionan directamente sobre pantalla o impresora, y un lenguaje de programación

bien desarrollado, simple y efectivo, que incluye condicionales, ciclos, funciones recursivas y posibilidad de entradas y salidas. (Debe destacarse que muchas de las funciones suministradas con el sistema están escritas en el lenguaje R)

El término "entorno" lo caracteriza como un sistema completamente diseñado y coherente, antes que como una agregación incremental de herramientas muy específicas e inflexibles, como ocurre frecuentemente con otros programas de análisis de datos.

R es en gran parte un vehículo para el desarrollo de nuevos métodos de análisis interactivo de datos. Como tal es muy dinámico y las diferentes versiones no siempre son totalmente compatibles con las anteriores. Algunos usuarios prefieren los cambios debido a los nuevos métodos y tecnología que los acompañan, a otros sin embargo les molesta ya que algún código anterior deja de funcionar. Aunque R puede entenderse como un lenguaje de programación, los programas escritos en R deben considerarse esencialmente efímeros.

En Wikipedia (Wikipedia, 2015), se define a **Weka** (Waikato Environment for Knowledge Analysis, en español «entorno para análisis del conocimiento de la Universidad de Waikato») como una plataforma de software para el aprendizaje automático y la minería de datos escrito en Java y desarrollado en la Universidad de Waikato. Weka es software libre distribuido bajo la licencia GNU-GPL.

El paquete Weka contiene una colección de herramientas de visualización y algoritmos para análisis de datos y modelado predictivo, unidos a una interfaz gráfica de usuario para acceder fácilmente a sus funcionalidades. La versión original de Weka fue un front-end (interfaz frontal de un software) en TCL/TK (lenguaje de herramientas de comando) para modelar algoritmos implementados en otros lenguajes de programación, más unas utilidades para preprocesamiento de datos desarrolladas en C (lenguaje de programación) para hacer experimentos de aprendizaje automático. Esta versión original se diseñó inicialmente como herramienta para analizar datos procedentes del dominio de la agricultura, pero la versión más reciente basada en Java (WEKA 3), que empezó a desarrollarse en 1997, se utiliza en muchas y muy diferentes áreas, en particular con finalidades docentes y de investigación.

Para la manipulación de los datos, utilizaremos en su gran mayoría, "R", y solamente para la selección de los atributos, "Weka", ya que posee varias técnicas implementadas.

CAPÍTULO 3.
RECOPILACIÓN Y LIMPIEZA DE DATOS

CAPÍTULO 3.

RECOPILACIÓN Y LIMPIEZA DE DATOS

3.1 Integración y Recopilación de Datos

En esta primera etapa, determinaremos la fuente de información de donde obtendremos los datos relevantes para realizar nuestro estudio.

En primera instancia, trabajaremos con un dataset de accidentes de tráfico en Reino Unido. Para ello, descargamos los archivos de la siguiente dirección web: **http://data.gov.uk/dataset/road-accidents-safety-data**, que contienen los datos publicados por el Departamento de Transporte del Gobierno del Reino Unido. Dentro del sitio web se puede apreciar los datos distribuidos por años. El archivo más antiguo hace referencia a accidentes que ocurrieron desde el año 1979 hasta el 2004, no obstante, encontramos que posee demasiados registros de diversos campos sin datos, por tal motivo, no es considerado en el estudio posterior.

Luego encontramos datos de accidentes ocurridos desde el 2005 hasta el 2013, en donde se pudo encontrar información relevante con detalles de los accidentes y de los automóviles involucrados en los mismos. Al explorar cada carpeta de cada año, no se encuentra información importante referido a los detalles de los automóviles del año 2005 al 2008, por tal motivo, se trabajará con información perteneciente a los años 2009 al 2013, ya que no registran ausencia de datos para el presente estudio.

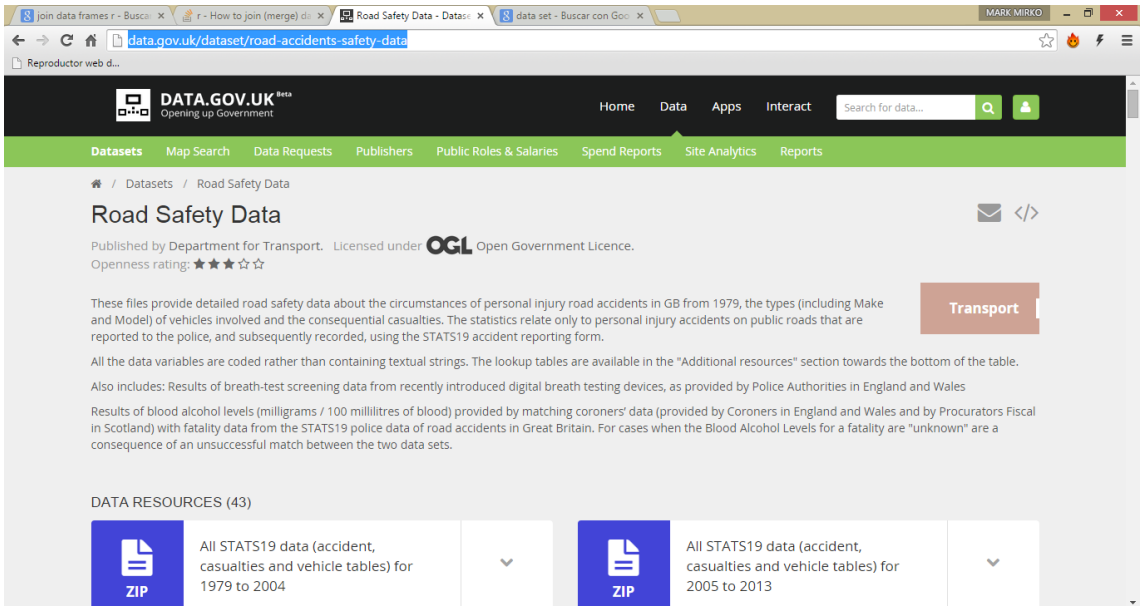


Ilustración 1. Portal web de datos libres del Reino Unido

Finalmente trabajaremos con los siguientes datasets:

Dataset	**Descripción**
DfTRoadSafety_Accidents_2009.csv, DfTRoadSafety_Accidents_2010.csv, DfTRoadSafety_Accidents_2011.csv, DfTRoadSafety_Accidents_2012.csv, DfTRoadSafety_Accidents_2013.csv	Proporcionan datos detallados de seguridad vial sobre las circunstancias de lesiones personales en accidentes de tráfico ocurridos en Reino Unido desde el año 2009 al 2013.
STATS19 Vehicle Accident Data 2009 with Make & Model.csv, STATS19 Vehicle Accident Data 2010 with Make & Model.csv, STATS19 Vehicle Accident Data 2011 with Make & Model.csv, Road Safety - Vehicles by Make and Model 2012 v2.csv, Road Safety - Vehicles by Make and Model 2013.csv	Proporciona datos detallados sobre los vehículos involucrados en accidentes de tráfico ocurridos en Reino Unido desde el año 2009 al 2013.

Tabla 2. Descripción de los data sets que se utilizarán para el estudio

Para integrar y realizar la limpieza de los datos, utilizaremos **"R"** (herramienta descrita en el epígrafe 2.6).

Después de descargar los archivos, lo que haremos seguidamente es unir los campos pertenecientes al dataset de los vehículos con los del dataset de los accidentes, de esta manera tendremos la información en un solo dataset.

Filtramos al dataset por aquellos registros que pertenecen al primer vehículo involucrado en el accidente.

```
datos_vehiculos_2009=subset(datos_vehiculos_2009,datos_vehiculos_2009$Vehicle_Reference
=="1")
```

Quitamos el campo "Vehicle_Reference", cuyo valor es 1 para todos, ya que hace referencia al primer vehículo involucrado en el accidente. Cabe recalcar que posteriormente se eliminarán campos irrelevantes después del análisis respectivo.

```
datos_vehiculos_2009=datos_vehiculos_2009[,-3]
```

Extraemos los nombres de las columnas

```
nombre_columnas=colnames(datos_vehiculos_2009[,-1])
```

Agregamos al dataset accidentes_2009, las columnas extraídas anteriormente

```
accidentes_2009[nombre_columnas] <- NA
```

Ahora, ya con las columnas agregadas, crearemos una función para introducir los registros de acuerdo al "Acc_Index", que es el campo que identifica a los accidentes.

```
agregar_datos_vehiculo_2009=function(){

    print("La hora de inicio del proceso es:")
    print(Sys.time())
    hora_inicio<-Sys.time()
```

```
  n=nrow(accidentes_2009)

  for (i in 1:n){
    print("Actualizado la fila:")
    print(i)

    filtro=accidentes_2009$Accident_Index[i]
    dataset=subset(datos_vehiculos_2009, datos_vehiculos_2009$Acc_Index==filtro)

    accidentes_2009$Vehicle_Type[i]=dataset$Vehicle_Type[1]
    accidentes_2009$Towing_and_Articulation[i]=dataset$Towing_and_Articulation[1]
    accidentes_2009$Vehicle_Manoeuvre[i]=dataset$Vehicle_Manoeuvre[1]
   accidentes_2009$Vehicle_Location.Restricted_Lane[i]=
   dataset$Vehicle_Location.Restricted_Lane[1]
   accidentes_2009$Skidding_and_Overturning[i]=dataset$Skidding_and_Overturning[1]
   accidentes_2009$X1st_Point_of_Impact[i]=dataset$Point_of_Impact[1]
   accidentes_2009$Journey_Purpose_of_Driver[i]=dataset$Journey_Purpose_of_Driver[1]
   accidentes_2009$Sex_of_Driver[i]=dataset$Sex_of_Driver[1]
   accidentes_2009$Age_Band_of_Driver[i]=dataset$Age_Band_of_Driver[1]
   accidentes_2009$Engine_Capacity_.CC.[i]=dataset$Engine_Capacity_.CC.[1]
   accidentes_2009$Propulsion_Code[i]=dataset$Propulsion_Code[1]
   accidentes_2009$Age_of_Vehicle[i]=dataset$Age_of_Vehicle[1]
   accidentes_2009$make[i]=dataset$make[1]
   accidentes_2009$model[i]=dataset$model[1]

 }

 print("La hora de finalización del proceso es:");
 print(Sys.time());
 hora_fin<-Sys.time();
 print("Tiempo de duración del proceso:");
 duracion=hora_fin - hora_inicio;
 print(duracion);

 assign('accidentes_2009',accidentes_2009,envir=.GlobalEnv)
}
```

Realizamos los mismos pasos para los datasets correspondientes a los años 2010, 2011, 2012 y 2013

Luego, unimos los 5 datasets para construir la **vista minable**

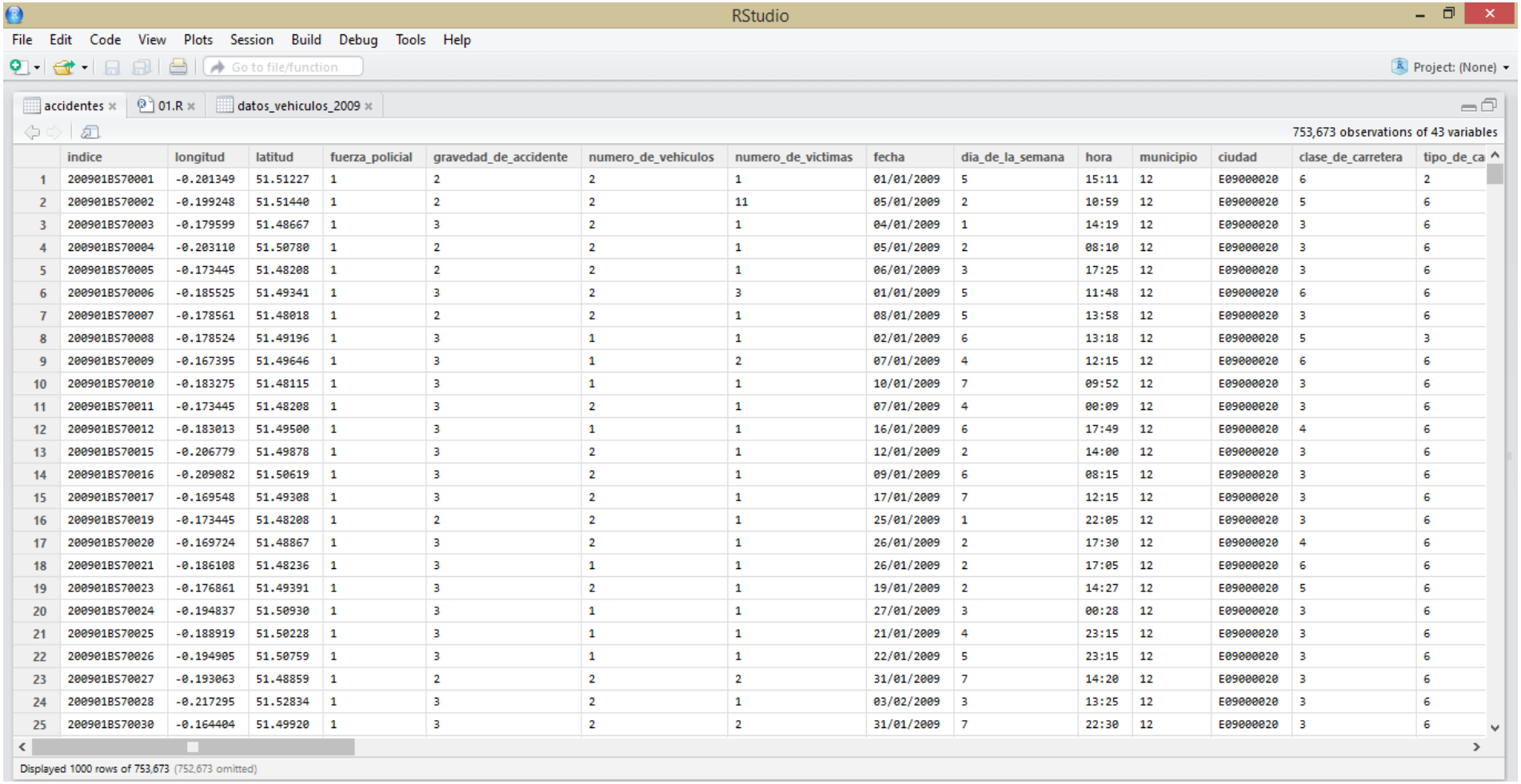
```
accidentes <- rbind(accidentes_2009, accidentes_2010, accidentes_2011, accidentes_2012,
accidentes_2013)
```

Ahora, tenemos un solo dataset con 40 variables, que son las siguientes:

N°	Atributo	Descripción
1	indice	Identificador del accidente
2	longitud	Longitud
3	latitud	Latitud
4	fuerza_policial	Condado al que pertenece la policía (Cumbria, Lancashire, etc.)
5	gravedad_de_accidente	Gravedad del accidente (fatal, serio, leve)
6	numero_de_vehiculos	Número de vehículos involucrados en el accidente
7	numero_de_victimas	Número de víctimas
8	fecha	Fecha del accidente
9	dia_de_la_semana	Día de la semana en el que ocurrió el accidente (Lunes, Martes, etc.)
10	hora	Hora del accidente
11	municipio	Municipio donde ocurrió el accidente (Westminster, Camden, etc.)
12	ciudad	Ciudad donde ocurrió el accidente
13	clase_de_carretera	Clase de carretera (autopista, A(M), A, B, C)
14	tipo_de_carretera	Tipo de carretera (rotonda, calle de un solo sentido, autovía, etc.)
15	limite_de_velocidad	Límite de velocidad
16	detalle_conexion_carretera	Detalle de la unión de la carretera (rotonda, mini rotonda, etc.)
17	control_conexion_carretera	Control de la unión de la carretera (persona autorizada, señal de stop, etc.)
18	control_humano_paso_peatones	Control humano para el pase de peatones (patrulla escolar, etc.)
19	patinaje_volcadura	Instalaciones físicas para el pase de peatones (cebra, pelícano, pasarela, etc.)
20	condicion_de_iluminacion	Condición de iluminación (luz, oscuro con luces encendidas, etc.)
21	condicion_climatica	Condición climática (lloviendo, nevando, etc.)
22	condicion_superficie_carretera	Condición de la superficie de la carretera (seco, mojado, nieve, etc.)
23	condicion_especial_carretera	Condición especial de la carretera (obras de carretera, superficie defectuosa, etc.)
24	riesgo_via	Riesgo en la vía (carga de vehículo, otro objeto, accidente anterior, etc.)
25	rural_o_urbano	Zona rural o urbana
26	presencia_policial	Presencia policial (si/no)
27	ubicación_bloque_accidente	Ubicación específica del accidente (Kensington and Chelsea 006A, etc.)
28	tipo_de_vehiculo	Tipo de vehículo (taxi, motocicleta, etc.)
29	remolque_y_articulacion	Remolque, articulación
30	maniobra_vehicular	Maniobra vehicular ocurrida durante el accidente (estacionado, etc.)
31	ubicacion_del_vehiculo	Ubicación del vehículo (carril del autobús, carril de las bicicletas, etc.)
32	patinaje_volcadura	Patinaje o volcadura
33	punto_de_impacto	Punto de impacto (frente, atrás, lateral, etc.)
34	proposito_de_viaje	Propósito del viaje
35	sexo_de_conductor	Sexo del conductor
36	edad_de_conductor	Edad del conductor
37	edad_del_vehiculo	Tiempo del vehículo
38	marca	Marca del vehículo (Honda, Ford, Suzuki, etc.)
39	modelo	Modelo del vehículo (C50LAC, CONNECT L230 D, RV 125 K7, ETC.)
40	anio_de_accidente	Año del accidente

Tabla 3. Descripción de los atributos del dataset principal después de la unificación

El dataset tiene el siguiente aspecto:



	indice	longitud	latitud	fuerza_policial	gravedad_de_accidente	numero_de_vehiculos	numero_de_victimas	fecha	dia_de_la_semana	hora	municipio	ciudad	clase_de_carretera	tipo_de_ca
1	200901BS70001	-0.201349	51.51227	1	2	2	1	01/01/2009	5	15:11	12	E09000020	6	2
2	200901BS70002	-0.199248	51.51440	1	2	2	11	05/01/2009	2	10:59	12	E09000020	5	6
3	200901BS70003	-0.179599	51.48667	1	3	2	1	04/01/2009	1	14:19	12	E09000020	3	6
4	200901BS70004	-0.203110	51.50780	1	2	2	1	05/01/2009	2	08:10	12	E09000020	3	6
5	200901BS70005	-0.173445	51.48208	1	2	2	1	06/01/2009	3	17:25	12	E09000020	3	6
6	200901BS70006	-0.185525	51.49341	1	3	2	3	01/01/2009	5	11:48	12	E09000020	6	6
7	200901BS70007	-0.178561	51.48018	1	2	2	1	08/01/2009	5	13:58	12	E09000020	3	6
8	200901BS70008	-0.178524	51.49196	1	3	1	1	02/01/2009	6	13:18	12	E09000020	5	3
9	200901BS70009	-0.167395	51.49646	1	3	1	2	07/01/2009	4	12:15	12	E09000020	6	6
10	200901BS70010	-0.183275	51.48115	1	3	1	1	10/01/2009	7	09:52	12	E09000020	3	6
11	200901BS70011	-0.173445	51.48208	1	3	2	1	07/01/2009	4	00:09	12	E09000020	3	6
12	200901BS70012	-0.183013	51.49500	1	3	1	1	16/01/2009	6	17:49	12	E09000020	4	6
13	200901BS70015	-0.206779	51.49878	1	3	2	1	12/01/2009	2	14:00	12	E09000020	3	6
14	200901BS70016	-0.209082	51.50619	1	3	2	1	09/01/2009	6	08:15	12	E09000020	3	6
15	200901BS70017	-0.169548	51.49308	1	3	2	1	17/01/2009	7	12:15	12	E09000020	3	6
16	200901BS70019	-0.173445	51.48208	1	2	2	1	25/01/2009	1	22:05	12	E09000020	3	6
17	200901BS70020	-0.169724	51.48867	1	3	2	1	26/01/2009	2	17:30	12	E09000020	4	6
18	200901BS70021	-0.186108	51.48236	1	3	1	1	26/01/2009	2	17:05	12	E09000020	6	6
19	200901BS70023	-0.176861	51.49391	1	3	2	1	19/01/2009	2	14:27	12	E09000020	5	6
20	200901BS70024	-0.194837	51.50930	1	3	1	1	27/01/2009	3	00:28	12	E09000020	3	6
21	200901BS70025	-0.188919	51.50228	1	3	1	1	21/01/2009	4	23:15	12	E09000020	3	6
22	200901BS70026	-0.194905	51.50759	1	3	1	1	22/01/2009	5	23:15	12	E09000020	3	6
23	200901BS70027	-0.193063	51.48859	1	2	2	2	31/01/2009	7	14:20	12	E09000020	3	6
24	200901BS70028	-0.217295	51.52834	1	3	2	1	03/02/2009	3	13:25	12	E09000020	3	6
25	200901BS70030	-0.164404	51.49920	1	3	2	2	31/01/2009	7	22:30	12	E09000020	3	6

Displayed 1000 rows of 753,673 (752,673 omitted)

Ilustración 2. Dataset después de la unificación

3.2 Selección, Limpieza y Transformación

En esta fase se eliminan o corrigen los datos incorrectos y se decide la estrategia a seguir con los datos incompletos. Además, se proyectan los datos para considerar únicamente aquellas variables o atributos que van a ser relevantes, con el objetivo de hacer más fácil la tarea propia de minería y para que los resultados de la misma sean más útiles.

En primer lugar, eliminamos atributos que son innecesarios para el estudio que realizaremos: índice, latitud y longitud.

accidentes=accidentes[,-1]
accidentes=accidentes[,-1]
accidentes=accidentes[,-1]

Exploramos, y averiguamos el tipo de datos del dataset:

str(accidentes)

```
$ fuerza_policial                : Factor w/ 43 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...
$ gravedad_de_accidente          : num  2 2 3 2 2 3 3 3 3 2 ...
$ numero_de_vehiculos            : int  2 2 2 2 2 1 1 2 2 2 ...
$ numero_de_victimas             : int  1 11 1 1 1 2 1 1 1 1 ...
$ fecha                          : chr  "01/01/2009" "05/01/2009" "04/01/2009" "05/01/2009" ...
$ dia_de_la_semana               : int  5 2 1 2 5 4 7 4 7 1 ...
$ hora                           : chr  "15-16" "10-11" "14-15" "08-09" ...
$ municipio                      : int  12 12 12 12 12 12 12 12 12 12 ...
$ ciudad                         : chr  "E09000020" "E09000020" "E09000020" "E09000020" ...
$ clase_de_carretera             : chr  "6" "5" "3" "3" ...
$ tipo_de_carretera              : chr  "Un sólo sentido" "Calzada unica" "Calzada unica" "Calzada unica" ...
$ limite_de_velocidad            : int  30 30 30 30 30 30 30 30 30 30 ...
$ detalle_conexion_carretera     : int  3 6 3 3 3 3 6 6 3 6 ...
$ control_conexion_carretera     : int  4 4 4 2 4 4 2 2 4 2 ...
$ control_humano_paso_peatones   : int  0 0 0 0 0 0 0 0 0 0 ...
$ instalacion_fisica_paso_peatones: int  0 1 0 5 0 0 0 5 0 5 ...
$ condicion_de_iluminacion       : int  1 1 1 1 1 1 1 4 1 4 ...
$ condicion_climatica            : int  1 1 1 8 1 1 8 1 1 1 ...
$ condicion_superficie_carretera : int  1 2 1 4 1 1 2 1 1 1 ...
$ condicion_especial_carretera   : int  0 0 0 0 0 0 4 0 0 0 ...
$ riesgo_via                     : int  0 0 0 0 0 0 0 0 0 0 ...
$ rural_o_urbano                 : int  1 1 1 1 1 1 1 1 1 1 ...
$ presencia_policial             : int  1 1 1 1 1 1 1 1 1 1 ...
$ ubicacion_bloque_accidente     : chr  "E01002882" "E01002886" "E01002912" "E01002871" ...
$ tipo_de_vehiculo               : num  2 13 3 5 7 7 2 7 3 7 ...
$ remolque_y_articulacion        : int  0 0 0 0 0 0 0 0 0 0 ...
$ maniobra_vehicular             : int  18 18 18 18 9 1 5 9 17 9 ...
$ ubicacion_del_vehiculo         : int  0 3 0 0 0 0 0 0 0 0 ...
$ patinaje_volcadura             : int  0 0 0 0 0 0 1 0 0 0 ...
$ punto_de_impacto               : int  1 1 0 3 1 4 0 4 1 1 ...
$ proposito_de_viaje             : int  15 1 15 1 1 15 1 15 15 15 ...
$ sexo_de_conductor              : int  1 1 1 1 2 1 2 1 1 1 ...
$ edad_de_conductor              : int  11 7 7 6 8 10 5 6 7 6 ...
$ capacidad_del_motor            : int  49 1753 124 535 5998 2967 49 1995 124 1598 ...
```

Ilustración 3. Tipo de datos del dataset

Convertimos a factor los atributos que lo requieran

accidentes$fuerza_policial=as.factor(accidentes$fuerza_policial)
accidentes$gravedad_de_accidente=as.factor(accidentes$gravedad_de_accidente)
accidentes$dia_de_la_semana=as.factor(accidentes$dia_de_la_semana)
accidentes$hora=as.factor(accidentes$hora)
accidentes$municipio=as.factor(accidentes$municipio)
accidentes$ciudad=as.factor(accidentes$ciudad)
accidentes$clase_de_carretera=as.factor(accidentes$clase_de_carretera)
accidentes$tipo_de_carretera=as.factor(accidentes$tipo_de_carretera)
accidentes$detalle_conexion_carretera=as.factor(accidentes$detalle_conexion_carretera)
accidentes$control_conexion_carretera=as.factor(accidentes$control_conexion_carretera)
accidentes$control_humano_paso_peatones=as.factor(accidentes$control_humano_paso_peat ones)

accidentes$instalacion_fisica_paso_peatones=as.factor(accidentes$instalacion_fisica_paso_peatones)
accidentes$condicion_de_iluminacion=as.factor(accidentes$condicion_de_iluminacion)
accidentes$condicion_climatica=as.factor(accidentes$condicion_climatica)
accidentes$condicion_superficie_carretera=as.factor(accidentes$condicion_superficie_carretera)
accidentes$condicion_especial_carretera=as.factor(accidentes$condicion_especial_carretera)
accidentes$riesgo_via=as.factor(accidentes$riesgo_via)
accidentes$presencia_policial=as.factor(accidentes$presencia_policial)
accidentes$ubicacion_bloque_accidente=as.factor(accidentes$ubicacion_bloque_accidente)
accidentes$tipo_de_vehiculo=as.factor(accidentes$tipo_de_vehiculo)
accidentes$remolque_y_articulacion=as.factor(accidentes$remolque_y_articulacion)
accidentes$maniobra_vehicular=as.factor(accidentes$maniobra_vehicular)
accidentes$patinaje_volcadura=as.factor(accidentes$patinaje_volcadura)
accidentes$marca=as.factor(accidentes$marca)
accidentes$modelo=as.factor(accidentes$modelo)
accidentes$anio_de_accidente=as.factor(accidentes$anio_de_accidente)

Mostramos un resumen del dataset:
summary(accidentes)

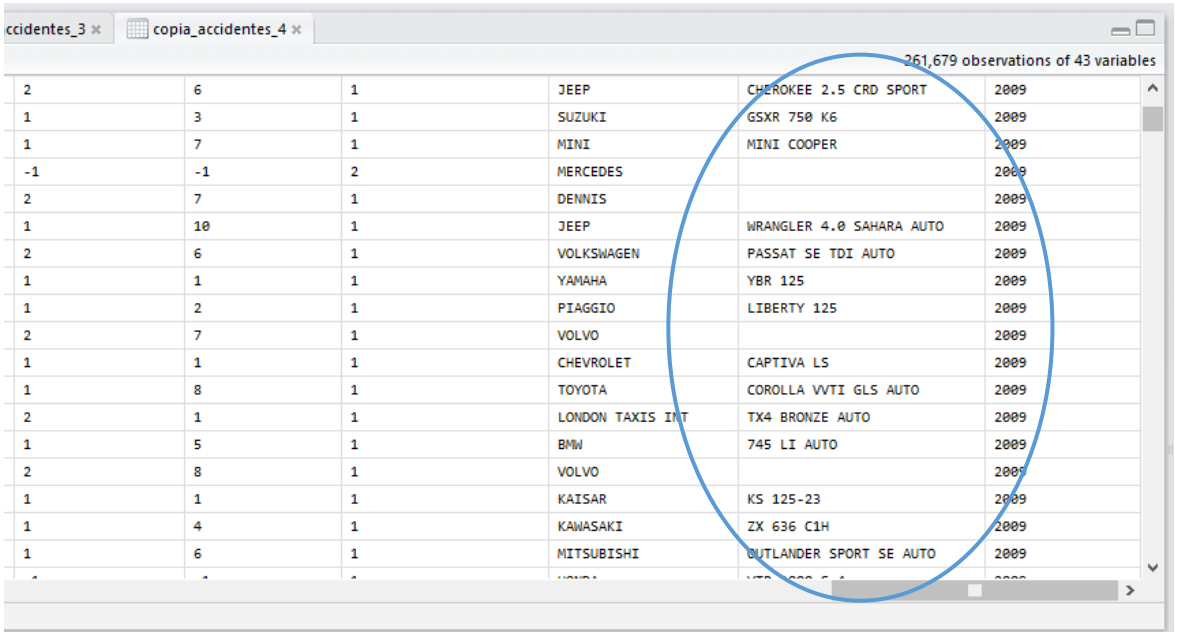
```
fuerza_policial  gravedad_de_accidente numero_de_vehiculos numero_de_victimas    fecha          dia_de_la_semana      hora
1      :117412   1:  8830             Min.   : 1.000      Min.   : 1.000     Length:753673     1: 81639         17:00  :  7491
20     : 28827   2:103948             1st Qu.: 1.000      1st Qu.: 1.000     Class :character  2:107557         17:30  :  7043
43     : 27813   3:640895             Median : 2.000      Median : 1.000     Mode  :character  3:113224         16:00  :  6707
13     : 27699                        Mean   : 1.824      Mean   : 1.346                       4:113201         18:00  :  6668
6      : 25320                        3rd Qu.: 2.000      3rd Qu.: 1.000                       5:114096         15:30  :  6561
46     : 24512                        Max.   :67.000      Max.   :87.000                       6:123644         16:30  :  6342
(Other):502090                                                                                  7:100312         (Other):712861
  municipio           ciudad       clase_de_carretera tipo_de_carretera limite_de_velocidad detalle_conexion_carretera
300    : 13334   E10000016: 21390   1: 27812           1: 51552          Min.   :10.00       0      :297873
204    :  9795   E10000030: 19680   2:  2162           2: 14964          1st Qu.:30.00       3      :238776
1      :  7323   E10000017: 16427   3:345901           3:108792          Median :30.00       6      : 73882
200    :  6780   E10000012: 15887   4: 96826           6:567084          Mean   :38.71       1      : 67379
102    :  6665   E10000014: 15534   5: 68600           7:  7880          3rd Qu.:50.00       8      : 27827
596    :  6574   E08000025: 13334   6:212372           9:  3401          Max.   :70.00       9      : 18952
(Other):703202   (Other)  :651421                                                           (Other): 28984
control_conexion_carretera control_humano_paso_peatones instalacion_fisica_paso_peatones condicion_de_iluminacion condicion_climatica
-1:273040                  0:749168                     0:620888                         1:554961                 1      :603729
0 : 25378                  1:  1841                     1: 21576                         4:145849                 2      : 86658
1 :  1168                  2:  2664                     4: 41307                         5:  3420                 8      : 17163
2 : 79611                                               5: 52692                         6: 40223                 9      : 13891
3 :  4052                                               7:  2346                         7:  9220                 5      :  9986
4 :370424                                               8: 14864                                                  4      :  8979
                                                                                                                  (Other): 13267
condicion_superficie_carretera condicion_especial_carretera riesgo_via  rural_o_urbano  presencia_policial ubicacion_bloque_accidente
-1:  1061                      0      :735623                -1:     5   Min.   :1.000   -1:     7                  : 51450
1 :519604                      4      :  8479                0 :740435   1st Qu.:1.000   1 :611364          E01000004:  1182
2 :202850                      6      :  2513                1 :   815   Median :1.000   2 :141762          E01004736:   737
3 :  8042                      7      :  2247                2 :  5331   Mean   :1.354   3 :   540          E01011365:   671
4 : 21062                      5      :  2000                3 :  1128   3rd Qu.:2.000                      E01008440:   489
5 :  1054                      1      :  1352                6 :  1872   Max.   :2.000                      E01004764:   398
                               (Other):  1459                7 :  4087                                      (Other)  :698746
tipo_de_vehiculo remolque_y_articulacion maniobra_vehicular ubicacion_del_vehiculo patinaje_volcadura punto_de_impacto proposito_de_viaje
9      :208431   0      :270963          18     :121236     0      :269559         0      :228676     -1  :     9     6      :108603
19     : 15078   1      :  2381          9      : 35636     9      :  1899         1      : 31792     0   : 16425     15     : 78390
11     : 11782   4      :  1020          4      : 20357     2      :  1579         2      :  8257     1   :152030     1      : 55496
5      :  9481   3      :   187          5      : 14965     6      :   514         5      :  5764     2   : 27903     2      : 25180
3      :  7504   5      :   178          17     : 13626     8      :   448         3      :   174     3   : 38707     5      :  3533
(Other): 22502   (Other):    49          (Other): 68958     (Other):   779         (Other):   115     4   : 39704     (Other):  3576
NA's   :478895   NA's   :478895          NA's   :478895     NA's   :478895         NA's   :478895     NA's:478895     NA's   :478895
sexo_de_conductor edad_de_conductor capacidad_del_motor propulsion_vehiculo edad_del_vehiculo tipo_area_casa_conductor
1   :188767       Min.   :-1.0      Min.   :   -1       1      :157510      Min.   :-1.0      Min.   :-1.0
2   : 81564       1st Qu.: 5.0      1st Qu.: 1229       2      : 98714      1st Qu.: 3.0      1st Qu.: 1.0
3   :  4447       Median : 7.0      Median : 1596       -1     : 17084      Median : 6.0      Median : 1.0
NA's:478895       Mean   : 6.5      Mean   : 1876       8      :  1023      Mean   : 6.4      Mean   : 1.1
                  3rd Qu.: 8.0      3rd Qu.: 1995       7      :   267      3rd Qu.:10.0      3rd Qu.: 1.0
                  Max.   :11.0      Max.   :91000       (Other):   180      Max.   :83.0      Max.   : 3.0
                  NA's   :478895    NA's   :478895      NA's   :478895      NA's   :478895    NA's   :478895
                  marca                          modelo       anio_de_accidente
FORD                 : 32440                      : 26550   2009:163554
VAUXHALL             : 32115   KA                 :  1143   2010:154414
PEUGEOT              : 18798   206 LX             :   879   2011:151474
VOLKSWAGEN           : 16661   CLIO DYNAMIQUE 16V :   842   2012:145571
RENAULT              : 15392   CORSA CLUB 12V     :   672   2013:138660
(Other)              :159372   (Other)            :244692
NA's                 :478895   NA's               :478895
```

Ilustración 4. Resumen de los atributos del data set

Como se puede observar hay valores NA's (datos vacíos) y -1's (datos perdidos, según la descripción de la base de datos) que tienen que ser eliminados del dataset.

accidentes<-na.omit(accidentes)
accidentes=subset(accidentes, accidentes$condicion_climatica!=-1)
accidentes=subset(accidentes, accidentes$condicion_superficie_carretera!=-1)
accidentes=subset(accidentes, accidentes$condicion_especial_carretera!=-1)
accidentes=subset(accidentes, accidentes$riesgo_via!=-1)
accidentes=subset(accidentes, accidentes$presencia_policial!=-1)
accidentes=subset(accidentes, accidentes$remolque_y_articulacion!=-1)
accidentes=subset(accidentes, accidentes$maniobra_vehicular!=-1)
accidentes=subset(accidentes, accidentes$ubicacion_del_vehiculo!=-1)
accidentes=subset(accidentes, accidentes$patinaje_volcadura!=-1)
accidentes=subset(accidentes, accidentes$punto_de_impacto!=-1)
accidentes=subset(accidentes, accidentes$proposito_de_viaje!=-1)
accidentes=subset(accidentes, accidentes$edad_de_conductor!=-1)
accidentes=subset(accidentes, accidentes$capacidad_del_motor!=-1)
accidentes=subset(accidentes, accidentes$propulsion_vehiculo!=-1)
accidentes=subset(accidentes, accidentes$edad_del_vehiculo!=-1)
accidentes=subset(accidentes, accidentes$control_conexion_carretera!=-1)

Observando el dataset, se puede ver que el campo "modelo", que hace referencia al modelo del vehículo, tiene muchos campos vacíos, así como otros campos, por tal motivo, convertimos los datos vacíos en NA's


ccidentes_3 × copia_accidentes_4 ×

261,679 observations of 43 variables

2	6	1	JEEP	CHEROKEE 2.5 CRD SPORT	2009
1	3	1	SUZUKI	GSXR 750 K6	2009
1	7	1	MINI	MINI COOPER	2009
-1	-1	2	MERCEDES		2009
2	7	1	DENNIS		2009
1	10	1	JEEP	WRANGLER 4.0 SAHARA AUTO	2009
2	6	1	VOLKSWAGEN	PASSAT SE TDI AUTO	2009
1	1	1	YAMAHA	YBR 125	2009
1	2	1	PIAGGIO	LIBERTY 125	2009
2	7	1	VOLVO		2009
1	1	1	CHEVROLET	CAPTIVA LS	2009
1	8	1	TOYOTA	COROLLA VVTI GLS AUTO	2009
2	1	1	LONDON TAXIS INT	TX4 BRONZE AUTO	2009
1	5	1	BMW	745 LI AUTO	2009
2	8	1	VOLVO		2009
1	1	1	KAISAR	KS 125-23	2009
1	4	1	KAWASAKI	ZX 636 C1H	2009
1	6	1	MITSUBISHI	OUTLANDER SPORT SE AUTO	2009

Ilustración 5. Data set mostrando los campos vacíos en el atributo "modelo del vehículo"

accidentes[accidentes==""]<-NA

Eliminamos los campos cuyo valor sea NA

accidentes<-na.omit(accidentes)

Ahora buscaremos datos anómalos, para ello tenemos en cuenta la Ilustración 4 y observamos que en algunos campos, los valores máximos son muy superiores a la media. Utilizaremos diagramas de caja (box plots) para estudiar las frecuencias de las variables y detectar datos anómalos.

Atributo: numero_de_vehiculos

boxplot(accidentes$numero_de_vehiculos)

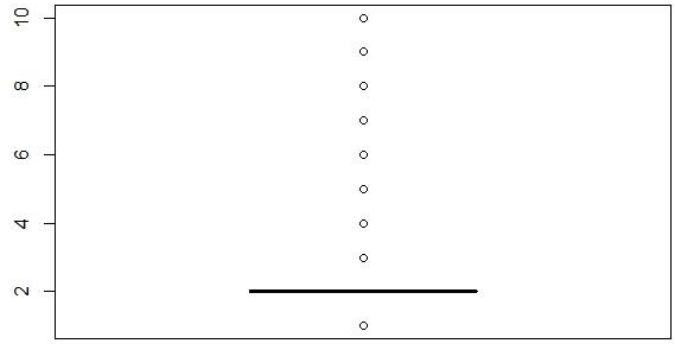


Ilustración 6. Diagrama de caja correspondiente al atributo "numero_de_vehiculos"

Después de eliminar registros con datos faltantes (=-1), se eliminaron también datos anómalos correspondientes al atributo "numero_de_vehiculos"

Atributo: numero_de_victimas

boxplot(accidentes$numero_de_victimas)

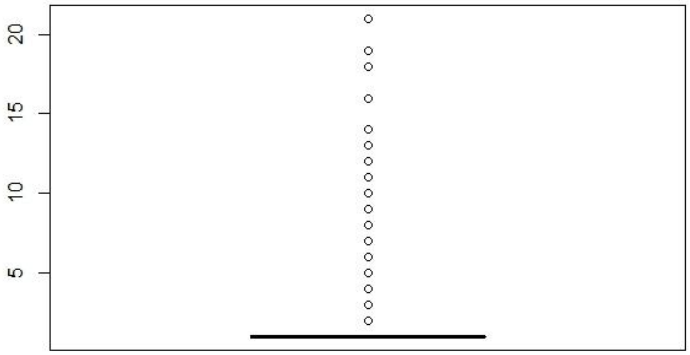


Ilustración 7. Diagrama de caja correspondiente al atributo "numero_de_victimas"

De la misma manera, se borraron datos anómalos correspondientes al atributo "numero_de_victimas"

Atributo: capacidad_de_motor

boxplot(accidentes$capacidad_del_motor)

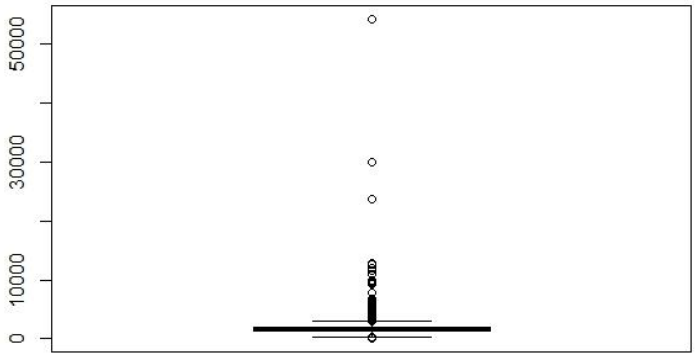


Ilustración 8. Diagrama de caja correspondiente al atributo “capacidad_de_motor”

Eliminamos datos anómalos

accidentes=subset(accidentes, accidentes$capacidad_del_motor<20000)

Atributo: edad_del_vehiculo

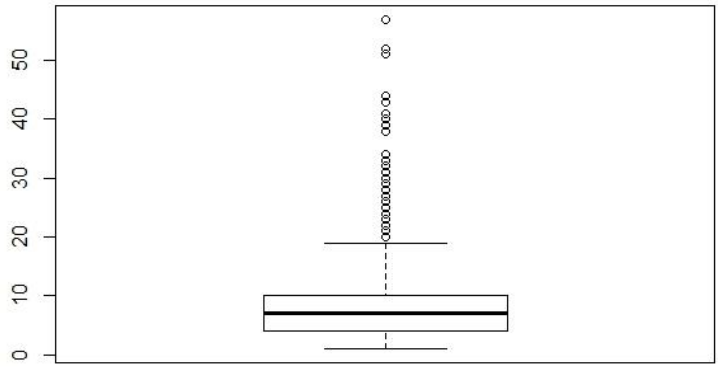


Ilustración 9. Diagrama de caja correspondiente al atributo “edad_del_vehiculo”

Eliminamos posibles datos anómalos

accidentes=subset(accidentes, accidentes$edad_del_vehiculo<45)

Eliminamos registros en donde el atributo “tipo_de_carretera” es igual a “Desconocido”

accidentes=subset(accidentes, accidentes$tipo_de_carretera!="Desconocido")

Se decidió borrar los datos faltantes (NA’s) y datos anómalos ya que contamos con muchos registros (261679) para el estudio.

Transformamos el atributo fecha a Date y, extraemos el mes

accidentes$fecha = as.POSIXct(accidentes$fecha)
meses=format(accidentes$fecha, "%m")
accidentes["mes"]=meses

accidentes=accidentes[c(1,2,3,4,5,40,39,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38)]
accidentes$mes=as.factor(accidentes$mes)

Después de todo el proceso de limpieza y transformación de los datos, obtenemos una base de datos sin datos erróneos. Finalmente contamos con 129627 registros y 39 atributos, que será nuestra vista minable.

CAPÍTULO 4.
ESTUDIO EXPERIMENTAL

CAPÍTULO 4.

ESTUDIO EXPERIMENTAL

En este capítulo, después de unificar, limpiar y transformar los datos, se realiza, en primer lugar, un análisis estadístico de los accidentes ocurridos, que nos darán información relevante sobre las tendencias que éstos poseen y bajo qué condiciones ocurren con más frecuencia. Seguidamente, teniendo en cuenta el objetivo principal de este estudio, aplicaremos diversas técnicas de minería de datos con el fin de encontrar patrones en los accidentes de tráfico cuya severidad es Fatal.

Para realizar lo anteriormente mencionado, utilizaremos en su gran mayoría R y Weka (sólo para el tema de la selección de atributos)

4.1 Análisis Estadístico

En esta sección se mostrarán gráficos estadísticos que describen la relación entre la frecuencia de los accidentes y los atributos más relevantes. Cabe recalcar que los gráficos están basados en la última base de datos que quedó como resultado después de la limpieza y transformación de los datos, es decir, todos los campos de los atributos son datos válidos, no anómalos.

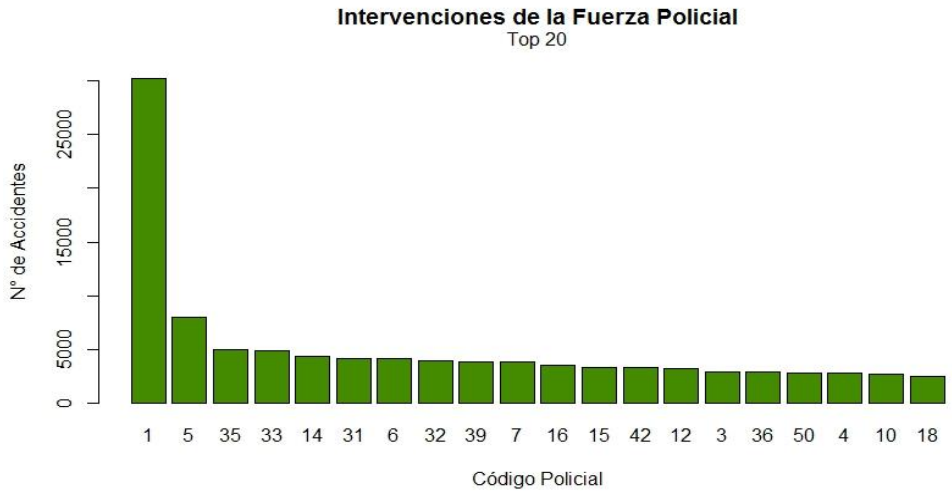


Ilustración 10. Gráfica de frecuencia de Intervenciones de la Fuerza Policial

Código	Descripción	Código	Descripción	Código	Descripción
1	Metropolitan Police	12	Humberside	33	Sussex
3	Lancashire	14	West Midlands	35	Devon and Cornwall
4	Merseyside	15	Staffordshire	36	Avon and Somerset
5	Greater Manchester	16	West Mercia	39	Dorset
6	Cheshire	18	Derbyshire	42	South Wales
7	Northumbria	31	Surrey	50	Strathclyde
10	West Yorkshire	32	Lincolnshire		

Tabla 4. Código y descripción de las Fuerzas Policiales

En este primer gráfico (Ilustración 10) se observa que la mayoría de accidentes se producen en un área metropolitana, mientras que en las otras áreas no hay grandes diferencias en cuanto al número de accidentes (por debajo de 5000 al año).

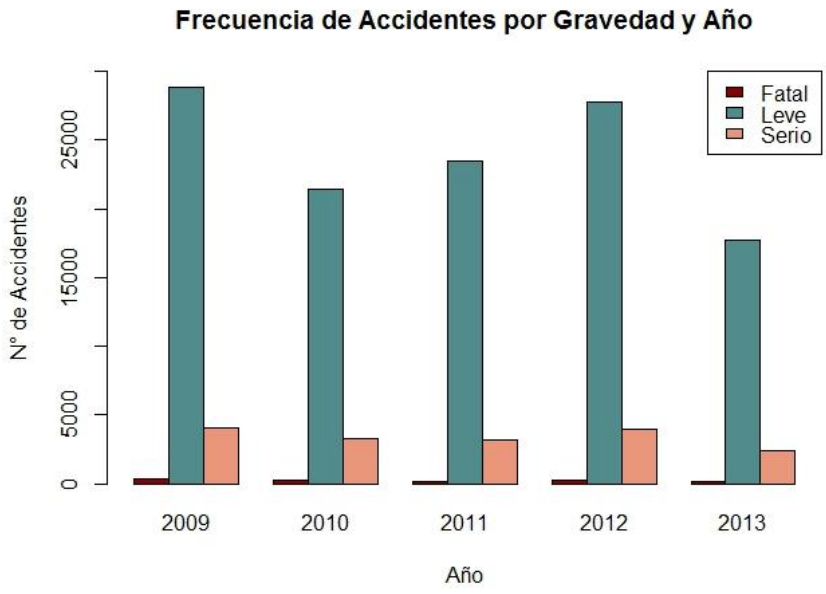


Ilustración 11. Gráfica de frecuencia de accidentes por Gravedad y año

En esta gráfica (Ilustración 11) podemos observar que la gravedad que predomina en todos los años es Leve, seguidamente Serio y Fatal; también notamos que la cantidad de accidentes ha disminuido en todos los casos.

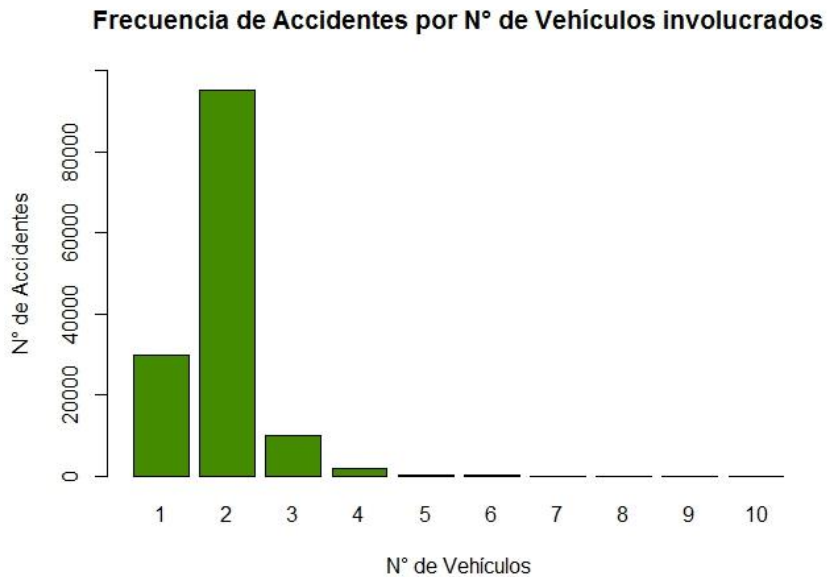


Ilustración 12. Gráfica de frecuencia de accidentes por n° de vehículos involucrados

En el gráfico superior (Ilustración 12), se observa que la mayoría de los accidentes ocurren entre dos vehículos (más de 80000).

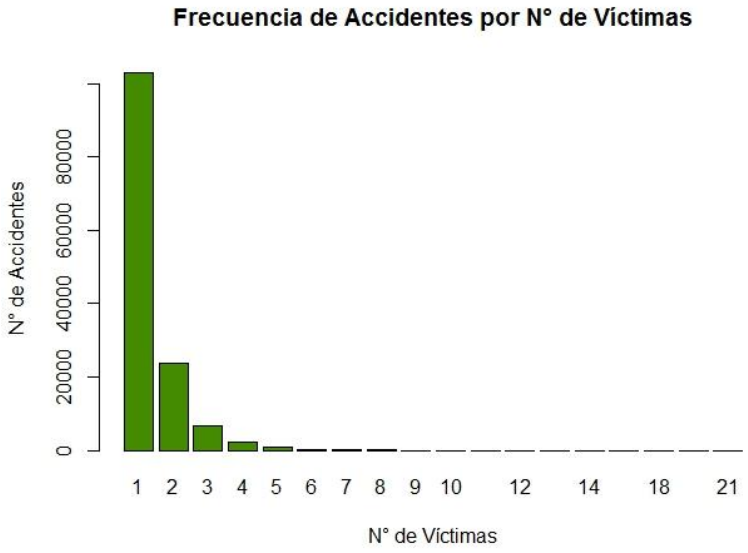


Ilustración 13. Gráfica de frecuencia de accidentes por n° de víctimas

La Ilustración 13 nos muestra que en la mayoría de los accidentes, el número de víctimas sólo es 1. Hasta 5 es la cantidad de víctimas con mayor frecuencia. Hay pocos accidentes con 6 o más víctimas.

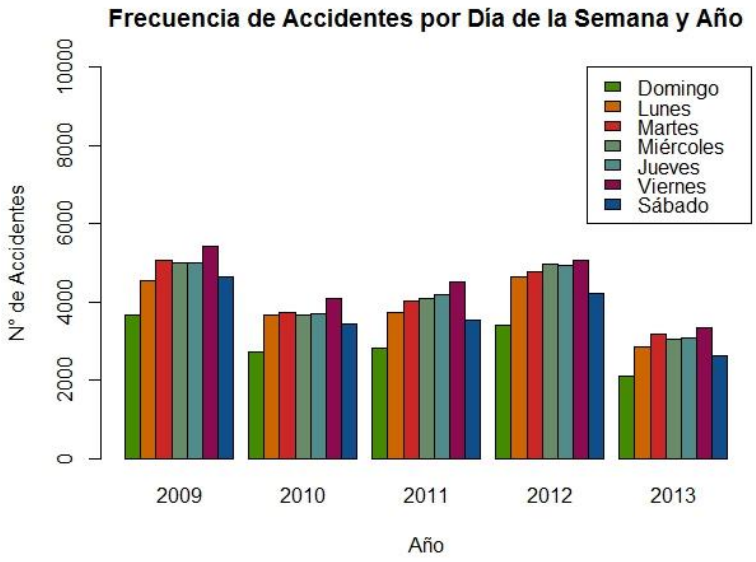


Ilustración 14. Gráfica de frecuencia de accidentes por día de la semana

Como podemos observar en la gráfica (Ilustración 14), el día con mayor frecuencia de accidentes es viernes y el de menor es Domingo.

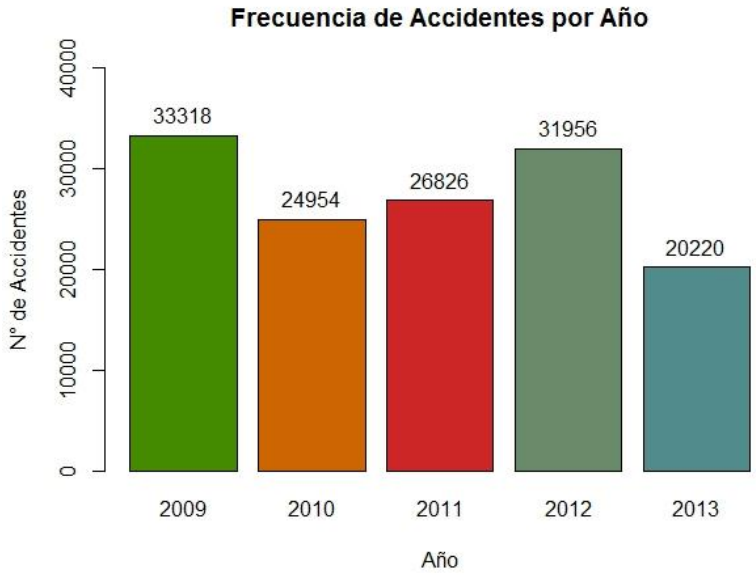


Ilustración 15. Gráfica de frecuencia de accidentes por año

El año 2010 presentó una disminución notoria (con respecto al año anterior) en la frecuencia de los accidentes, no obstante, en el 2011 y en el 2012 el número de accidentes se fue incrementando hasta que en el 2013 volvió a reducirse la cantidad de los incidentes (Ver Ilustración 15).

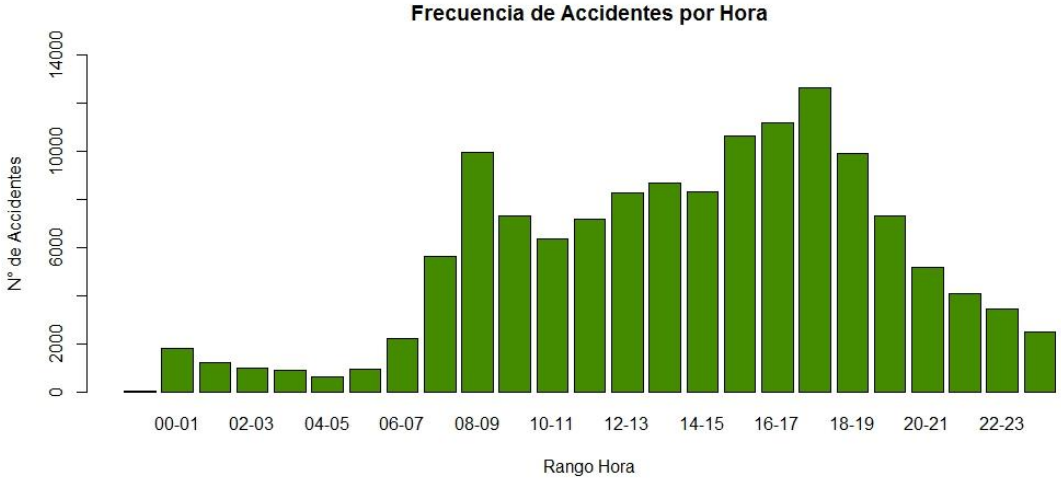


Ilustración 16. Gráfica de frecuencia de accidentes por hora

En la gráfica (Ilustración 16), podemos apreciar que en las tardes, la ocurrencia de los accidentes es mayor, predominando entre las 17:00 y 18:00 horas. También se observa una importante cantidad de accidentes ocurridos entre las 08:00 y 09:00 de la mañana.

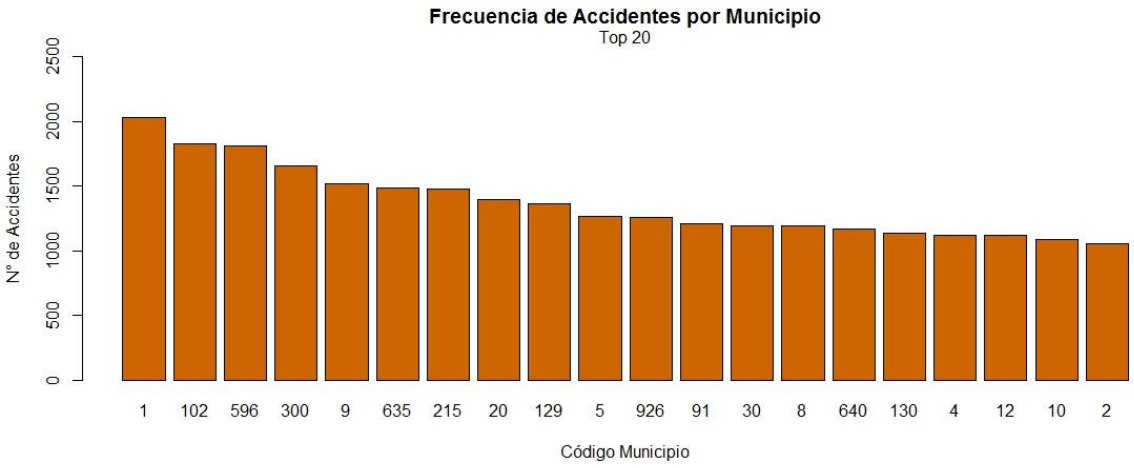


Ilustración 17. Gráfica de frecuencia de accidentes por municipio

Código	Descripción	Código	Descripción	Código	Descripción
1	Westminster	20	Croydon	640	Bournemouth
102	Manchester	129	Cheshire East	130	Cheshire West and Chester
596	Cornwall	5	Tower Hamlets	4	Hackney
300	Birmingham	926	Glasgow City	12	Kensington and Chelsea
9	Lambeth	91	Liverpool	10	Wandsworth
635	Wiltshire	30	Barnet	2	Camden
215	Sheffield	8	Southwark		

Tabla 5. Código y descripción de los Municipios

En la imagen superior (Ilustración 17), se observa que en el Municipio de Westminster, ocurren más accidentes, seguido de Manchester y Cornwall.

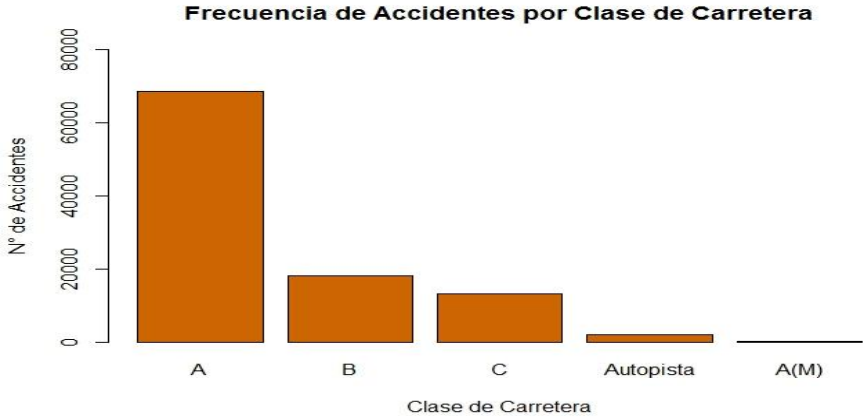


Ilustración 18. Gráfica de frecuencia de accidentes por clase de carretera

En las carreteras cuya clase es A, la ocurrencia de los accidentes sucede con mayor frecuencia según la Ilustración 18.

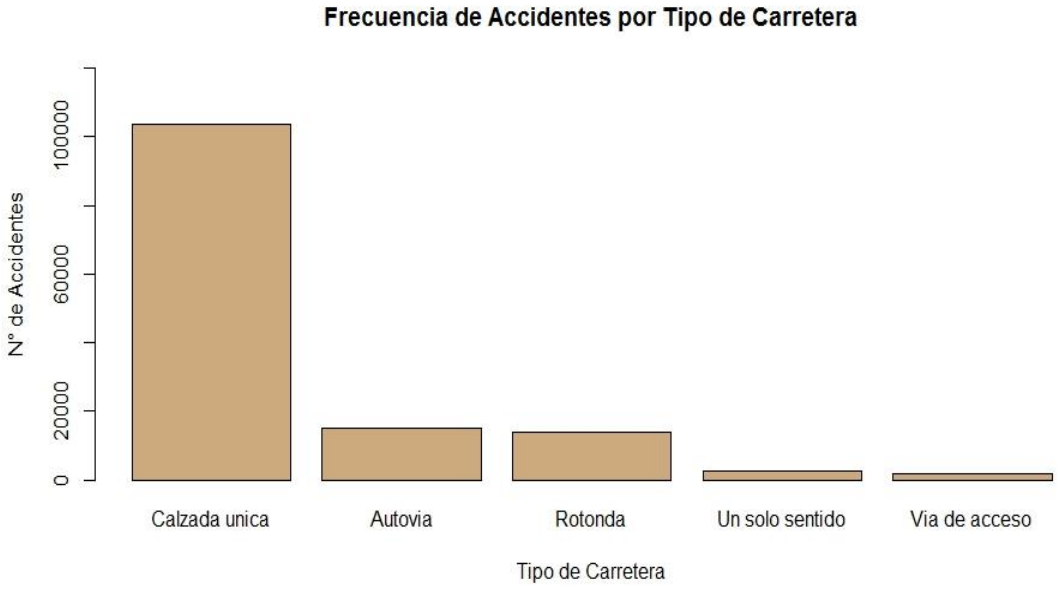


Ilustración 19. Gráfica de frecuencia de accidentes por tipo de carretera

Las carreteras del tipo Calzada única (que tienen una sola calzada para ambos sentidos de circulación), son las que presentan mayor ocurrencia de accidentes según la Ilustración 19.

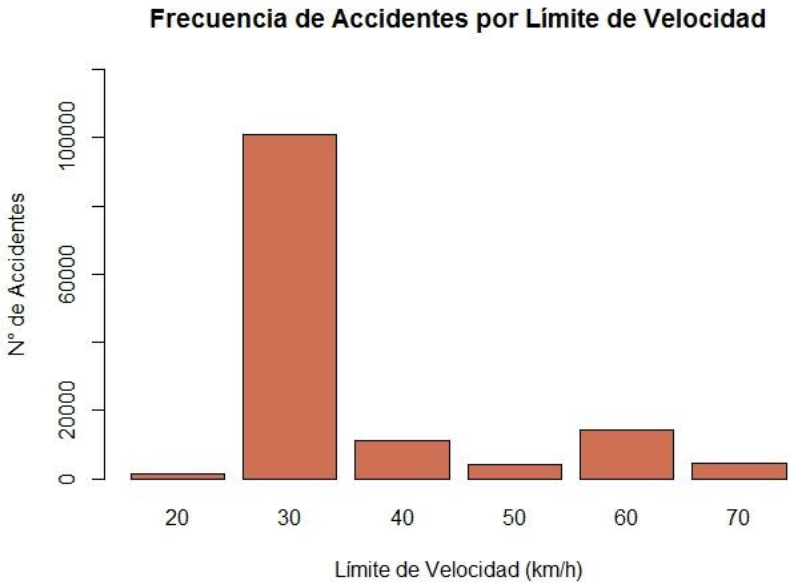


Ilustración 20. Gráfica de frecuencia de accidentes por límite de velocidad

Según el gráfico (Ilustración 20), los accidentes ocurren con mayor frecuencia en las carreteras cuyo límite de velocidad es de 30 km/h.

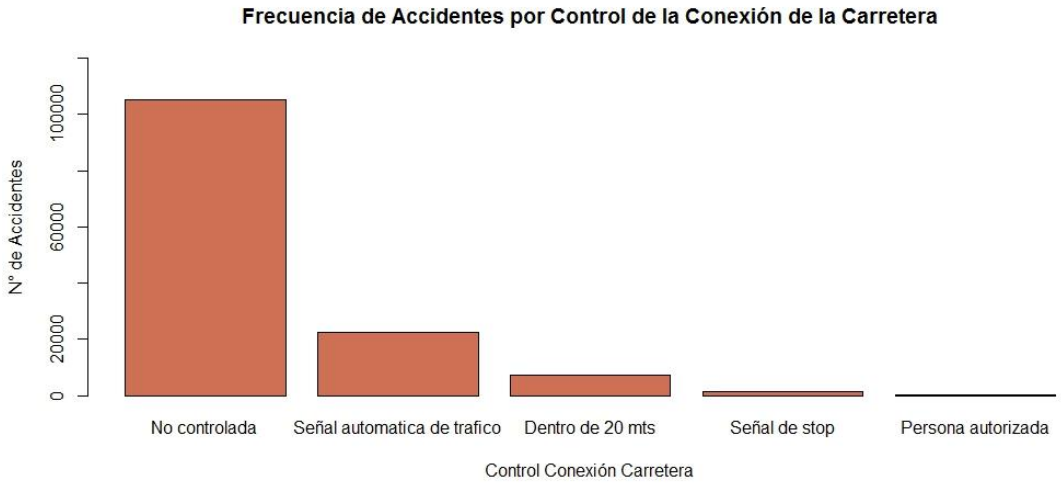


Ilustración 22. Gráfica de frecuencia de accidentes por control de la conexión de la carretera

En la imagen (Ilustración 21), podemos observar que cuando no hay control alguno en la conexión de la carretera (tramos de carretera cortos entre carreteras divididas, que permiten hacer cambios de sentido), la ocurrencia de los accidentes es mayor que cuando existe algún mecanismo de control.

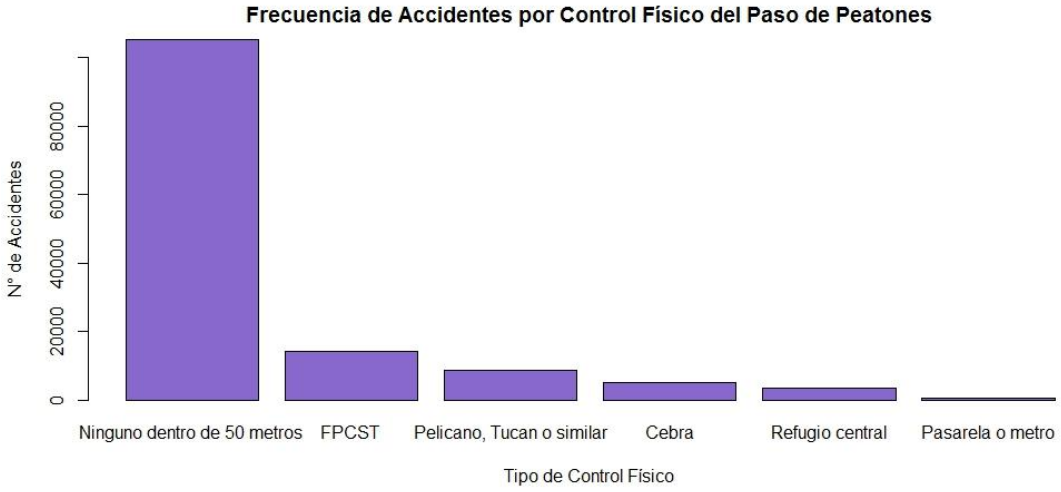


Ilustración 24. Gráfica de frecuencia de accidentes por control físico del paso de peatones

La Ilustración 22, nos muestra que cuando no existe un control físico (cebra, pelícano, etc.), la ocurrencia de los accidentes es mayor que cuando existe alguno.

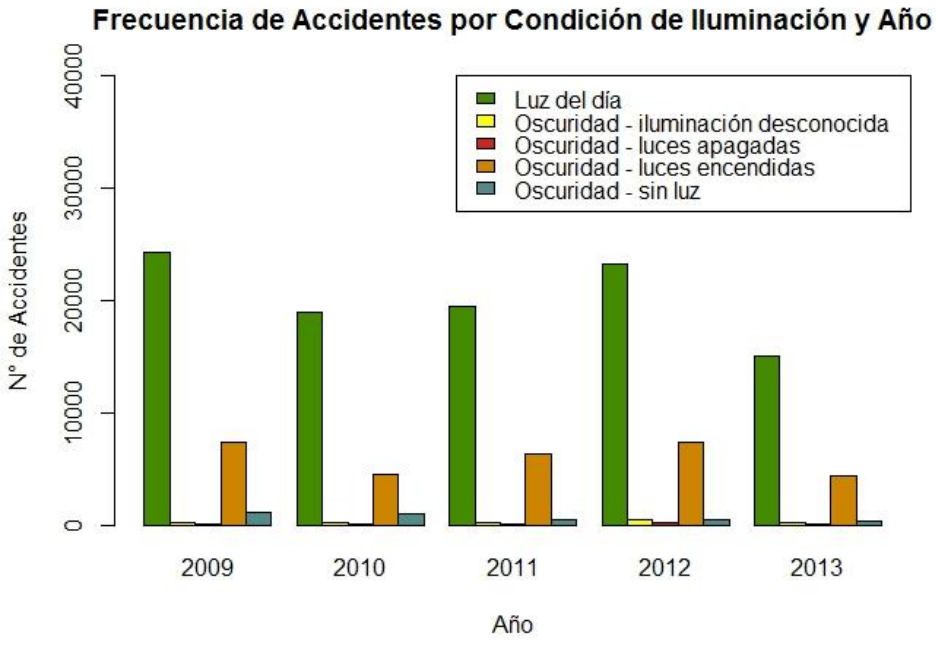


Ilustración 25. Gráfica de frecuencia de accidentes por condición de iluminación

El gráfico (Ilustración 23) nos muestra que la mayoría de los accidentes ocurren en el día.

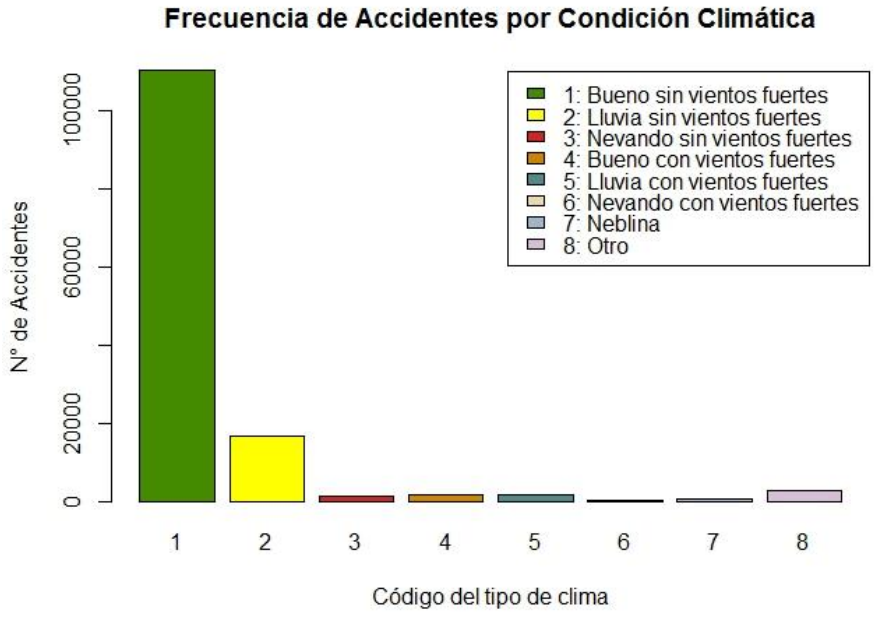


Ilustración 26. Gráfica de frecuencia de accidentes por condición climática

Teniendo en cuenta la condición climática, lo accidentes ocurren más cuando el clima es bueno y cuando hay lluvias sin vientos fuertes, según la Ilustración 24.

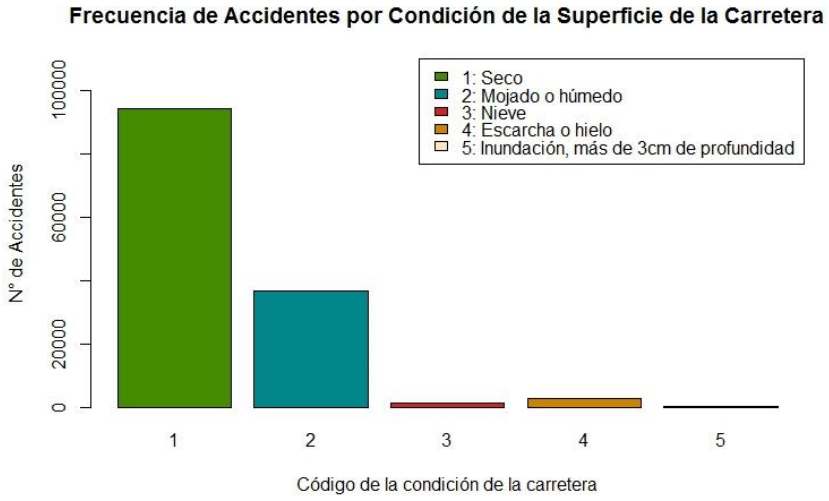


Ilustración 27. Gráfica de frecuencia de accidentes por condición de la superficie de la carretera

La Ilustración 25 nos muestra que cuando la superficie de la carretera es seca, mojada o húmeda, los accidentes ocurren con mayor frecuencia.

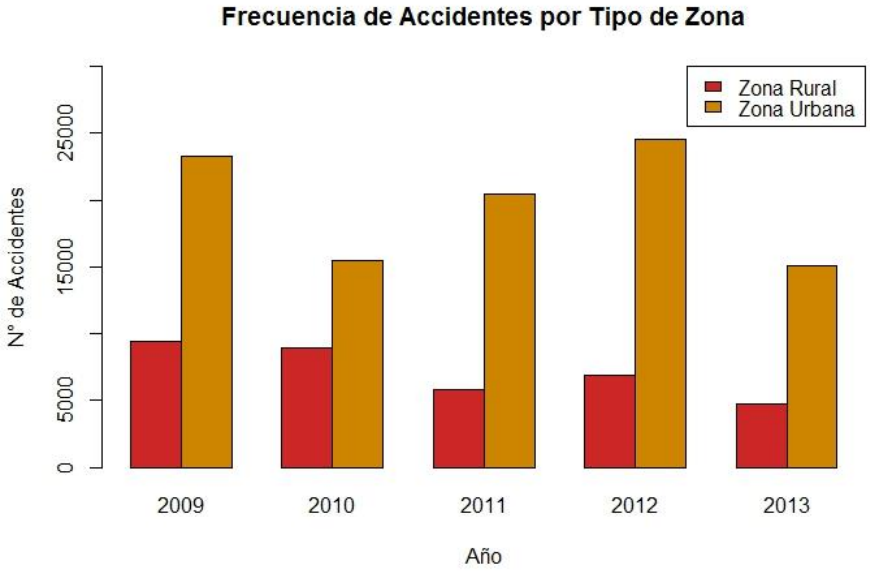


Ilustración 28. Gráfica de frecuencia de accidentes por tipo de zona

Se puede observar en la Ilustración 26, que mayormente los accidentes ocurren en zonas urbanas. En todos los años la tendencia se mantiene. Además identificamos este atributo como redundante ya que en la Ilustración 10, que representa la frecuencia de intervenciones policiales, se observa claramente que la Policía metropolitana tiene intervenciones superiores a los demás.

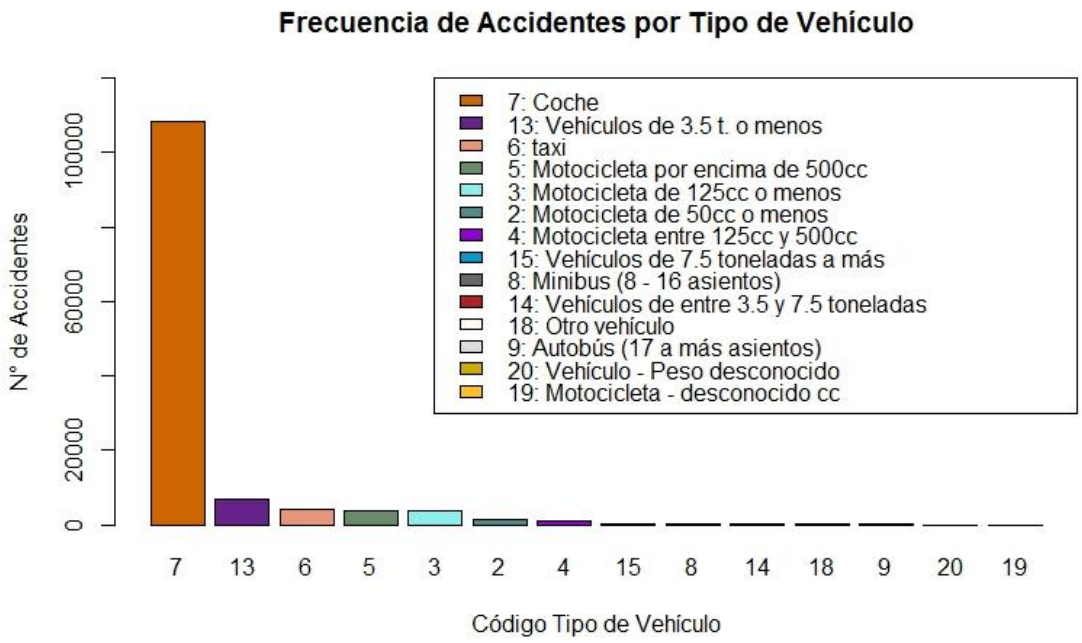


Ilustración 30. Gráfica de frecuencia de accidentes por tipo de vehículo

Los vehículos que están involucrados en la mayoría de los accidentes son los coches, vehículos de 3.5 t. o menos, taxis y motocicletas, como se puede observar en la Ilustración 27.

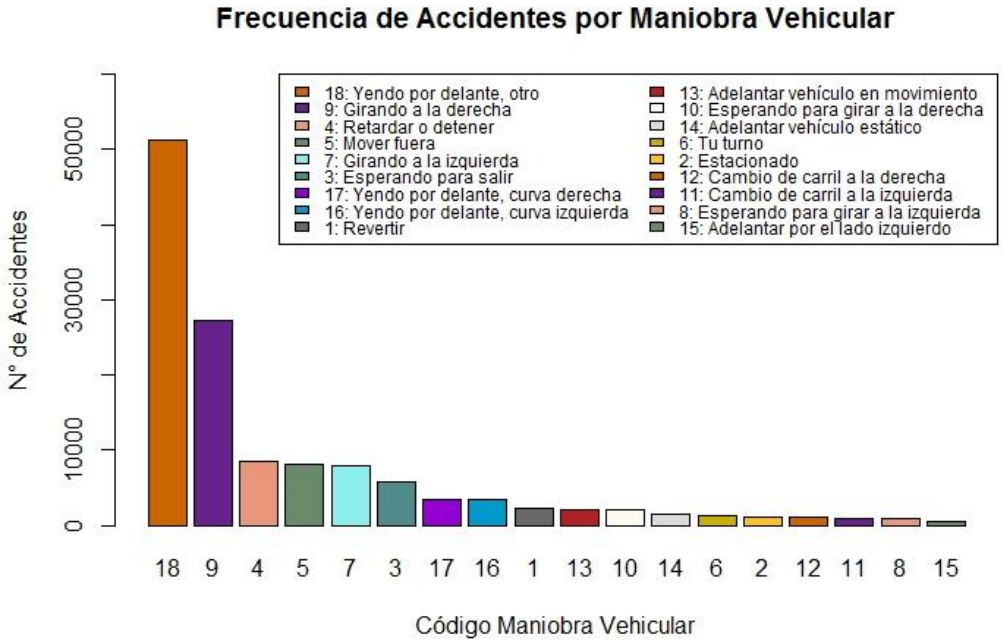


Ilustración 31. Gráfica de frecuencia de accidentes por maniobra vehicular.

Muchos de los accidentes son causados por la maniobra vehicular que el conductor realiza (como se podrá ver más adelante). En la Ilustración 28, descubrimos que cuando el coche se encuentra por delante de otro o cuando gira a la derecha, ocurren accidentes con mayor frecuencia.

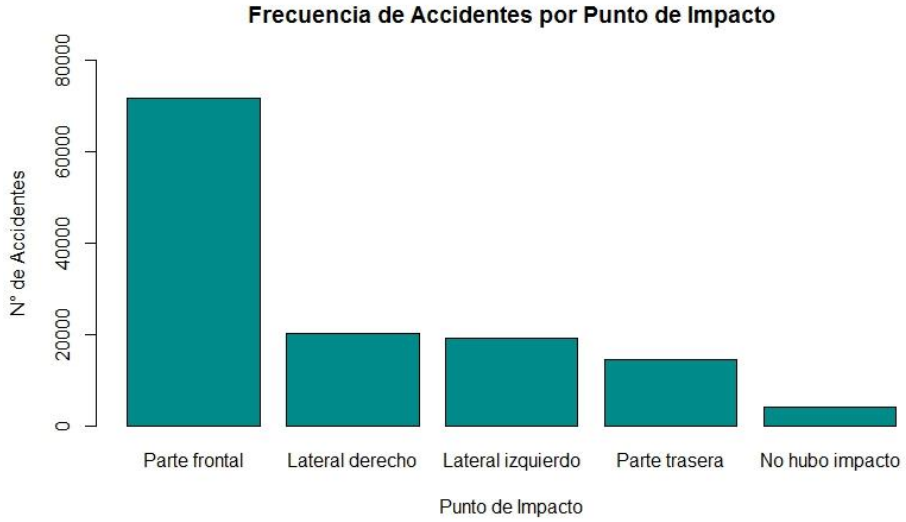


Ilustración 32. Gráfica de frecuencia de accidentes por punto de impacto.

La parte del coche con mayor frecuencia de impacto en un accidente es la frontal, según la imagen (Ilustración 29). Ambas partes laterales derecha e izquierda muestran cantidades de impacto similares.

4.2 Minería de Datos, Evaluación e Interpretación

En esta etapa, utilizaremos diversas técnicas de minería de datos, descritos anteriormente, con el propósito de cumplir nuestro objetivo general, que consiste básicamente, en identificar patrones de accidentes que ocurren dentro de las carreteras, cuya gravedad del accidente es “Fatal.

El atributo “gravedad_de_accidente” va a ser el atributo a predecir en nuestro problema de clasificación. Este atributo tiene 3 valores: fatal, serio y leve. Para nuestro estudio nos centraremos en los accidentes fatales, ya que queremos en última instancia obtener reglas de decisión sobre cuáles son las causas que más influyen en este tipo de accidentes, a partir de las cuales se podrían efectuar recomendaciones para evitar futuros accidentes mortales.

En primer lugar, convertiremos los accidentes serios y leves en “no fatales”, de esta manera sólo tendremos 2 valores de clase: “fatal” y “no fatal”.

accidentes$gravedad_de_accidente[accidentes$gravedad_de_accidente==’serio’]=’no fatal’
accidentes$gravedad_de_accidente[accidentes$gravedad_de_accidente==’leve’]=’no fatal’

Dado que la gravedad de los accidentes en zona rural es superior a la de los accidentes en zona urbana, trabajaremos sólo con datos de accidentes ocurridos en zona rural.

accidentes=subset(accidentes, accidentes$rural_o_urbano=="2")

La gravedad del accidente depende de factores de diversa naturaleza (tales como el vehículo, la carretera o el conductor). Sin embargo, los factores que afectan a los accidentes con un solo vehículo involucrado pueden ser diferentes de los que afectan a accidentes con más vehículos involucrados. En los accidentes múltiples la gravedad está altamente relacionada con factores tales como el tipo de colisión, el tamaño y peso de los vehículos involucrados en el accidente, los puntos de contacto, etc. (Krull et al., 2000).

La gravedad de los accidentes de un sólo vehículo ha sido previamente estudiada por al menos tres razones (Chang and Yeh, 2006). En primer lugar, por lo general, estos accidentes presentan mayor gravedad que los accidentes múltiples, por lo que son un objetivo prioritario de cara al desarrollo de estrategias de mejora de la seguridad vial. En segundo lugar, en estos accidentes el comportamiento del conductor o los factores humanos que contribuyen a la gravedad del accidente pueden ser explorados con mayor eficacia. Y en tercer lugar, estudiar la gravedad de los accidentes con un solo vehículo puede simplificar el diseño de la investigación mediante la exclusión de los efectos relativos a otros vehículos.

Por lo tanto, en este estudio sólo se estudiará accidentes con un vehículo involucrado.

accidentes= subset(accidentes,accidentes$numero_de_vehiculos=="1")

Después del proceso de filtrado contamos con una muestra de 6749 accidentes válidos y 39 variables.

Seguidamente construiremos nuestros modelos aplicando las distintas técnicas de aprendizaje (descritas en el epígrafe 2.2.1) y los evaluaremos utilizando **validación cruzada de 10 folds** (grupos). Los paquetes utilizados implementados en **R** fueron los siguientes: "rpart" para Árboles de decisión), "kknn" para k vecinos más cercanos (es un algoritmo de knn con pesos), "e1071" para Naive Bayes, "class" para Máquinas de soporte vectorial, "randomForest" para Bosques aleatorios, "ada" para Potenciación y "nnet" para Redes neuronales. Para todos ellos se usaron los valores por defecto de todos sus parámetros.

La siguiente tabla, en la que los métodos se denotan con las abreviaturas definidas en la sección 2.2.1, muestra los resultados:

Técnica de aprendizaje	**Acierto: clase "Fatal"**	**Acierto: clase "No fatal"**	**Accuracy**
ADD	0.0000	1.0000	0.9747
K-NN	0.0090	0.9980	0.9731
NB	0.0169	0.9962	0.9714
SVM	0.1185	0.9096	0.8901
RF	0.0000	0.9998	0.9746
AdaBoost	0.0098	0.9959	0.9710
RNA	0.0000	1.0000	0.9747

Tabla 6. Resultados de precisión de los modelos, evaluados con validación cruzada de 10 grupos, utilizando el dataset de Reino Unido.

Como se puede observar en la Tabla 6, aunque la precisión (columna "Accuracy") global es muy buena en la mayoría de modelos, la precisión alcanzada al predecir los accidentes de tipo "Fatal", es muy mala. De hecho, algunos modelos como el árbol de decisión o la red neuronal clasifican todas las instancias de clase "No fatal". En estos experimentos hemos trabajado con todas las variables (39) del dataset "accidentes". Ya hemos visto en el apartado anterior que hay algunos atributos que pueden ser redundantes. También puede ser que no todos los atributos sean igual de relevantes a la hora de predecir la gravedad del accidente. Por lo tanto lo que haremos en el siguiente apartado es aplicar estrategias de selección automática de atributos con mayor poder discriminante de la severidad del accidente. Para tal objetivo, utilizaremos **Weka**, ya que esta herramienta posee técnicas implementadas para la selección de atributos.

4.2.1 Filtrado de variables

En la siguiente tabla se pueden observar los subconjuntos de atributos obtenidos por ***Weka*** utilizando el método de evaluación ***CfsSubsetEval*** y diferentes métodos de búsqueda de atributos relevantes.

Método de búsqueda	N° de atributos	Atributos
BestFirst	8	dia_de_la_semana tipo_de_carretera condicion_de_iluminacion condicion_especial_carretera presencia_policial maniobra_vehicular patinaje_volcadura sexo_de_conductor
GeneticSearch	7	numero_de_victimas dia_de_la_semana condicion_especial_carretera presencia_policial maniobra_vehicular patinaje_volcadura sexo_de_conductor
GreedyStepwise	8	dia_de_la_semana tipo_de_carretera condicion_de_iluminacion condicion_especial_carretera presencia_policial maniobra_vehicular patinaje_volcadura sexo_de_conductor
LinearForwardSelection	8	dia_de_la_semana tipo_de_carretera condicion_de_iluminacion condicion_especial_carretera presencia_policial maniobra_vehicular patinaje_volcadura sexo_de_conductor
RankSearch	8	dia_de_la_semana tipo_de_carretera condicion_de_iluminacion condicion_especial_carretera presencia_policial maniobra_vehicular patinaje_volcadura sexo_de_conductor
ScatterSearchV1	8	dia_de_la_semana tipo_de_carretera condicion_de_iluminacion condicion_especial_carretera presencia_policial maniobra_vehicular patinaje_volcadura sexo_de_conductor

SubsetSizeForwardSelection	8	dia_de_la_semana tipo_de_carretera condicion_de_iluminacion condicion_especial_carretera presencia_policial maniobra_vehicular patinaje_volcadura sexo_de_conductor

Tabla 7. Subconjuntos de atributos seleccionados con el evaluador de atributos CfsSubsetEval, utilizando el dataset de Reino Unido.

Como puede apreciarse, los subconjuntos obtenidos en la mayoría de los métodos, excepto uno, son iguales, por lo tanto, seleccionaremos dicho subconjunto para el estudio posterior.

accidentes <- subset(accidentes, select = c("gravedad_de_accidente","riesgo_via","condicion_especial_carretera","ubicacion_del_vehiculo","condicion_climatica","tipo_de_carretera","condicion_de_iluminacion","remolque_y_articulacion","maniobra_vehicular","tipo_de_vehiculo"))

Paso siguiente, volvemos a construir nuestros modelos con las mismas técnicas de aprendizaje ya mencionadas y los evaluaremos utilizando validación cruzada de 10 grupos.

La siguiente tabla muestra los resultados:

Técnica de aprendizaje	**Acierto: clase "Fatal"**	**Acierto: clase "No fatal"**	**Accuracy**
ADD	0.0000	1.0000	0.9748
K-NN	0.0000	0.9995	0.9743
NB	0.0000	1.0000	0.9748
SVM	0.0000	0.9998	0.9746
RF	0.0000	1.0000	0.9748
AdaBoost	0.0000	1.0000	0.9748
RNA	0.0000	1.0000	0.9748

Tabla 8. Resultados de precisión de los modelos, evaluados con validación cruzada de 10 grupos, después de aplicar la técnica de selección de atributos, utilizando el dataset de Reino Unido

De manera similar que en la tabla anterior (Tabla 7), se puede apreciar en la Tabla 8, que la precisión alcanzada al predecir los accidentes de tipo “Fatal”, es muy mala (incluso peor que la obtenida usando todos los atributos del dataset). Para ver cuál puede ser la causa de estos resultados analizamos cuántos accidentes Fatales y No Fatales existen en el dataset.

table(accidentes$gravedad_de_accidente)

Fatal	No Fatal
170	6579

El resultado nos muestra la existencia de un **problema de desbalanceo de las clases**, ya que sólo el 2.51% de los accidentes son Fatales y el 97.49% No Fatales. Esto hace que los modelos tiendan a ajustarse a la clase mayoritaria, obviando la clase minoritaria la cual es, en nuestro caso, el principal objetivo del estudio. Por tal motivo aplicaremos técnicas de balanceado de clases con la intención de mejorar los resultados.

4.2.2 Balanceado de clases

Dado que las categorías de la variable “gravedad_de_accidente” no se encuentran balanceadas, y que este hecho afecta tanto a la precisión total del modelo como a la probabilidad predicha de cada clase, utilizaremos algunos algoritmos para el balanceado de los datos.

Utilizaremos los paquetes “DMwR” y “ROSE” para aplicar las técnicas de SMOTE y ROSE respectivamente (que se describieron en la sección 2.5). También dividiremos aleatoriamente el dataset principal en 10 grupos utilizando validación cruzada. Cabe recalcar que el balanceado de datos se aplicará únicamente a la partición de entrenamiento y no a la de test. El proceso consiste básicamente en dividir la tabla aleatoriamente en 10 grupos iguales, reservo 1 grupo para el testeo y las 9 restantes para el entrenamiento, a ésta tabla (entrenamiento) aplicamos balanceo de clases y entrenamos el modelo. Después de entrenar el modelo, testeamos con la tabla reservada para el testeo. Obtenemos los resultados (acierto clase fatal, acierto clase no fatal, Accuracy y AUC) y lo guardamos. Este proceso se repite 10 veces, pero en cada iteración se seleccionará 1 grupo distinto para el testeo (2do, 3er, 4to, etc. grupo), y los 9 restantes para el entrenamiento.

En la siguiente ilustración, extraída de Wikipedia (Wikipedia, 2015) , muestra una base de datos dividida en 4 grupos.

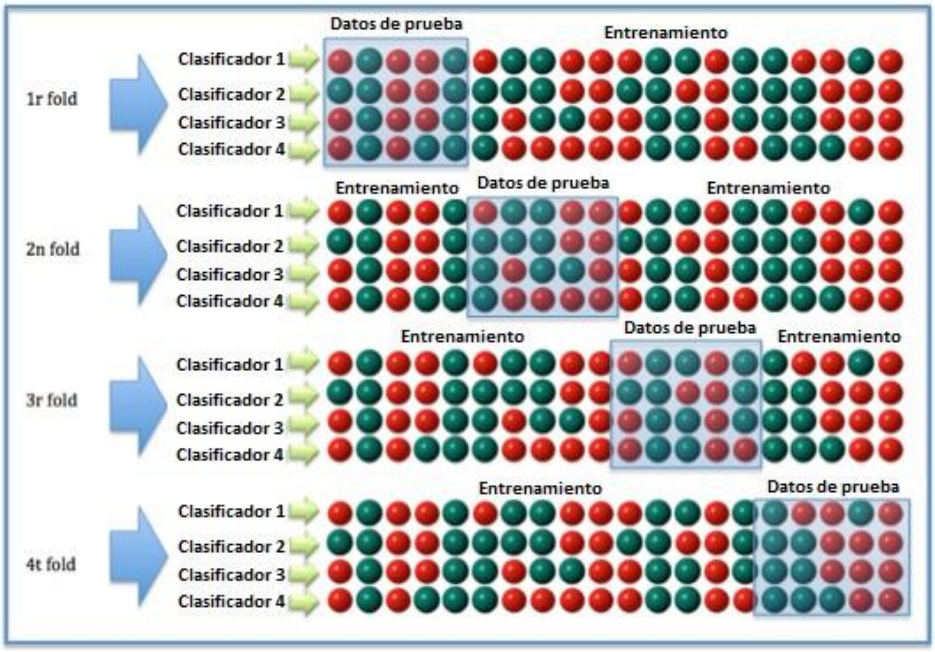


Ilustración 34. Validación cruzada con 4 grupos y 4 clasificadores

Finalmente promediamos los 10 resultados obtenidos, convirtiéndose en el resultado final.

En la siguiente tabla 9 se muestran los resultados obtenidos después de balancear los datos con algunos algoritmos, y aplicar las diversas técnicas de aprendizaje. Hemos añadido una nueva columna (AUC) que mide el Area Under the ROC Curve (Flach, Hernandez-Orallo, & Ferri, 2011), una medida de evaluación de clasificadores más adecuada para evaluar problemas desbalanceados (López, Fernández, García, Palade, & Herrera, 2013).

Técnica de aprendizaje	**Técnica de balanceo de datos**	**Acierto clase "Fatal"**	**Acierto clase "No fatal"**	**Accuracy**	**AUC**
ADD	SMOTE	0.4814	0.7713	0.7637	0.5145
	ROSE	**0.5383**	**0.7264**	0.7215	**0.5162**
K-NN	SMOTE	0.5890	0.5827	0.5834	0.5094
	ROSE	0.5066	0.6303	0.6273	0.5074
NBs	SMOTE	0.5190	0.7762	0.7702	**0.5211**
	ROSE	**0.6699**	**0.6137**	0.6135	0.5116
SVM	SMOTE	0.5013	0.7448	0.7389	**0.5185**
	ROSE	**0.6135**	**0.6861**	0.6842	0.5163
RF	SMOTE	0.5016	0.6774	0.6737	0.5134
	ROSE	0.6065	0.6388	0.6375	0.5141
AdaBoost	SMOTE	**0.7746**	0.3500	0.3616	0.5092
	ROSE	0.3306	0.4896	0.4860	0.5095
RNs	SMOTE	0.4594	0.7702	0.7631	0.5139
	ROSE	0.5599	0.6430	0.6410	0.5117

Tabla 9. Resultados de precisión de los modelos, evaluados con validación cruzada de 10 grupos, después de aplicar la técnica de selección de atributos y balanceo de datos, utilizando el dataset de Reino Unido.

Después de balancear las clases (Fatal y No Fatal), se puede ver claramente un cambio radical en los resultados que se manifiestan en la Tabla 9. Aunque el mejor resultado para predecir la clase Fatal es el obtenido por AdaBoost (0.7746%) este algoritmo comete muchas equivocaciones en la clase No fatal (0.3500 %) por lo que el accuracy global y el AUC no son buenos. Los árboles de decisión y las máquinas de vectores soporte obtienen unos buenos valores tanto en las medidas globales como en el acierto de cada clase por separado, al igual que las redes bayesianas.

Con objeto de mejorar el modelo vamos a comparar los resultados con los obtenidos haciendo una evaluación usando 5 pliegues para la validación cruzada. A continuación se muestran los resultados obtenidos:

Técnica de aprendizaje	Técnica de balanceo de datos	Acierto clase "Fatal"	Acierto clase "No fatal"	Accuracy	AUC
ADD	SMOTE	0.6157	0.6035	0.6038	0.5127
	ROSE	**0.6652**	**0.6506**	**0.6509**	0.5168
K-NN	SMOTE	0.5511	0.6519	0.6495	0.5113
	ROSE	0.3112	0.7784	0.7667	0.5062
NBs	SMOTE	0.6183	0.6532	0.6523	0.5145
	ROSE	0.7321	0.6268	0.6294	0.5187
SVM	SMOTE	0.6252	0.6665	0.6654	0.5159
	ROSE	**0.7310**	**0.6539**	**0.6559**	**0.5209**
RF	SMOTE	0.6826	0.5058	0.5105	0.5119
	ROSE	0.5058	0.7432	0.7372	0.5160
AdaBoost	SMOTE	0.7222	0.4347	0.4424	0.5131
	ROSE	0.2248	0.5051	0.4982	0.5137
RNA	SMOTE	0.5506	0.6622	**0.6598**	0.5157
	ROSE	0.5767	0.6549	0.6531	0.5148

Tabla 10. Resultados de precisión de los modelos, evaluados con validación cruzada de 5 grupos, después de aplicar la técnica de selección de atributos y balanceo de datos, utilizando el dataset de Reino Unido.

Finalmente como se puede observar en la Tabla 10, logramos mejorar los resultados. Según los valores de la Razón de Verdaderos Positivos (Acierto clase "Fatal"), la Razón de Falsos Positivos (Acierto clase "No Fatal") y el Accuracy, los mejores resultados obtenemos a partir del uso de Máquinas de Soporte Vectorial y redes Neuronales, no obstante, utilizando Árboles de decisión también obtenemos resultados aceptables tal y como vimos en la tabla 9.

Cabe destacar que los valores de Accuracy se encuentran dentro del rango de los valores obtenidos en otros estudios en los que se han aplicado métodos de clasificación con objetivos similares. López, G. (2013) obtuvo una precisión del **55.57%** con Árboles de decisión. Abdelwahab y Abdel-Aty (2001) obtuvieron una precisión del **60.4%** al aplicar redes Neuronales. De Oña, J., Oqab, R. y Calvo, F. (2011) obtuvieron precisiones del **58%, 59% y 61%** aplicando redes Bayesianas con diferentes algoritmos (AIC, MDL Y BDeu respectivamente). En el último estudio realizado por De Oña, J., López, G., Mujalli, R. y Calvo, F. (2013), aplicando Análisis Clúster y redes Bayesianas, obtuvieron precisiones que variaban de un **55,1%** a un **64%**.

4.2.3 Extracción de Reglas de Decisión (RDs)

A partir del mejor modelo creado basado en Árboles de Decisión en el experimento 4 (Tabla 10), extraemos las reglas de decisión. A continuación mostramos el árbol generado por el modelo:

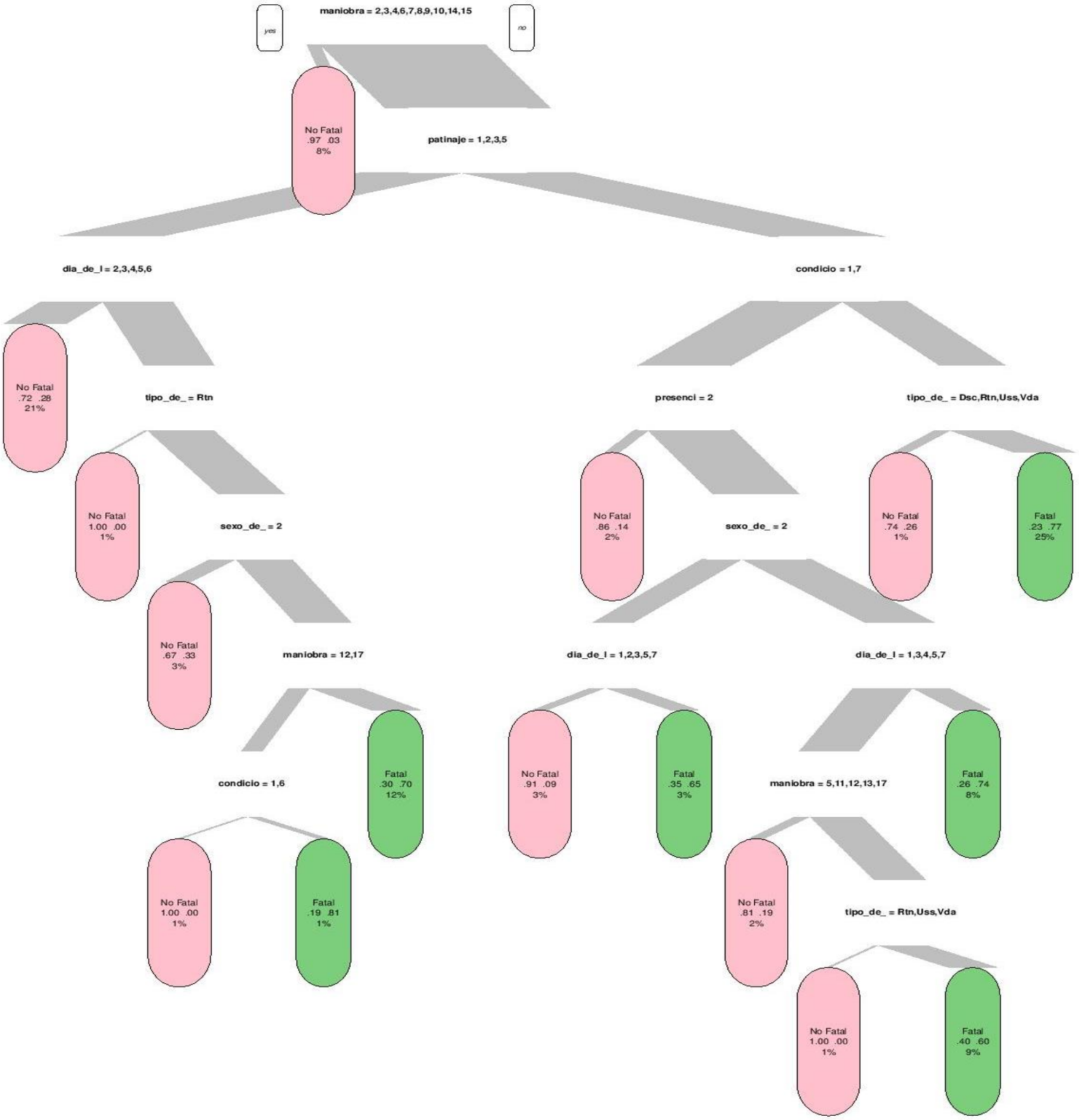


Ilustración 35. Árbol de decisión generado por el modelo (dataset: Reino Unido).

Las RDs extraídas de un Árbol de decisión dependen de la propia estructura del árbol. De modo que la extracción de conocimiento (en forma de RDs) sólo se realiza en el sentido dictado desde el nodo raíz hasta cada uno de los nodos terminales del árbol.

Todos los nodos del árbol pueden ser transformados en reglas de decisión de la forma: "Si (X) entonces (Y)". En la Tabla 11 se muestra las 6 posibles RDs (coincidentes con los nodos terminales del árbol), donde los accidentes ocurridos presentan gravedad Fatal.

N°	Regla	Entonces
1	Si [(maniobra_vehicular= "Cambio de carril a la derecha" ó maniobra_vehicular= "Yendo por delante, curva derecha") y (patinaje_volcadura= "Patinaje" ó patinaje_volcadura= "Patinaje y Volcadura" ó patinaje_volcadura= "plegadura del remolque" ó patinaje_volcadura= "Volcadura") y (dia_de_la_semana= "Sábado" ó dia_de_la_semana= "Domingo") y (tipo_de_carretera != "Rotonda") y (sexo_conductor = "Hombre") y (condicion_de_iluminacion= "Oscuridad, luces encendidas" ó condicion_de_iluminacion= "Oscuridad, luces apagadas" ó condicion_de_iluminacion= "Oscuridad, iluminación desconocida")]	gravedad_de_accidente = Fatal
2	Si [(maniobra_vehicular= "Revertir" ó maniobra_vehicular= "Mover hacia afuera" ó maniobra_vehicular= "Cambio de carril a la izquierda" ó maniobra_vehicular= "Adelantar vehículo en movimiento" ó maniobra_vehicular= "Yendo por delante, curva izquierda" ó maniobra_vehicular= "Yendo por delante, otro") y (patinaje_volcadura= "Patinaje" ó patinaje_volcadura= "Patinaje y Volcadura" ó patinaje_volcadura= "plegadura del remolque" ó patinaje_volcadura= "Volcadura") y (dia_de_la_semana= "Sábado" ó dia_de_la_semana= "Domingo") y (tipo_de_carretera != "Rotonda") y (sexo_conductor = "Hombre")]	gravedad_de_accidente = Fatal
3	Si [(maniobra_vehicular= "Revertir" ó maniobra_vehicular= "Mover hacia afuera" ó maniobra_vehicular= "Cambio de carril a la izquierda" ó maniobra_vehicular= "Cambio de carril a la derecha" ó maniobra_vehicular= "Adelantar vehículo en movimiento" ó maniobra_vehicular= "Yendo por delante, curva izquierda" ó maniobra_vehicular= "Yendo por delante, curva derecha" ó maniobra_vehicular= "Yendo por delante, otro") y	gravedad_de_accidente = Fatal

	(patinaje_volcadura= “plegadura del remolque y volcadura”) y (condicion_de_iluminacion= “Luz del día” ó condicion_de_iluminacion= “Oscuridad, iluminación desconocida”) y (presencia_policial= “Si”) y (sexo_de_conductor= “Mujer”) y (dia_de_la_semana= “Miércoles” ó dia_de_la_semana= “Viernes”)]	
4	Si [(maniobra_vehicular= “Yendo por delante, curva izquierda” ó maniobra_vehicular= “Yendo por delante, otro”) y (patinaje_volcadura= “plegadura del remolque y volcadura”) y (condicion_de_iluminacion= “Luz del día” ó condicion_de_iluminacion= “Oscuridad, iluminación desconocida”) y (presencia_policial= “Si”) y (sexo_de_conductor= “Hombre”) y (dia_de_la_semana= “Domingo” ó dia_de_la_semana= “Martes” ó dia_de_la_semana= “Miércoles” ó dia_de_la_semana= “Jueves” ó dia_de_la_semana= “Sábado”) y (tipo_de_carretera= “Autovía” ó tipo_de_carretera= “Calzada única”)]	gravedad_de_accidente = Fatal
5	Si [(maniobra_vehicular= “Revertir” ó maniobra_vehicular= “Mover hacia afuera” ó maniobra_vehicular= “Cambio de carril a la izquierda” ó maniobra_vehicular= “Cambio de carril a la derecha” ó maniobra_vehicular= “Adelantar vehículo en movimiento” ó maniobra_vehicular= “Yendo por delante, curva izquierda” ó maniobra_vehicular= “Yendo por delante, curva derecha” ó maniobra_vehicular= “Yendo por delante, otro”) y (patinaje_volcadura= “plegadura del remolque y volcadura”) y (condicion_de_iluminacion= “Luz del día” ó condicion_de_iluminacion= “Oscuridad, iluminación desconocida”) y (presencia_policial= “Si”) y (sexo_de_conductor= “Hombre”) y (dia_de_la_semana= “Lunes” ó dia_de_la_semana= “Viernes”)]	gravedad_de_accidente = Fatal
6	Si [(maniobra_vehicular= “Revertir” ó maniobra_vehicular= “Mover hacia afuera” ó maniobra_vehicular= “Cambio de carril a la izquierda” ó maniobra_vehicular= “Adelantar vehículo en movimiento” ó	gravedad_de_accidente = Fatal

	maniobra_vehicular= “Yendo por delante, curva izquierda” ó maniobra_vehicular= “Yendo por delante, otro”) y (patinaje_volcadura= “plegadura del remolque y volcadura”) y (condicion_de_iluminacion= “Oscuridad, luces encendidas” ó condicion_de_iluminacion= “Oscuridad, luces apagadas” ó condicion_de_iluminacion= “Oscuridad, sin luz”) y (tipo_de_carretera= “Autovía” ó tipo_de_carretera= “Calzada única”)]	

Tabla 11. Reglas de decisión extraídas del árbol de decisión (dataset: Reino Unido).

En la mayoría de los casos, los accidentes cuya severidad es Fatal, son ocasionados por personas de sexo masculino.

Para conductores masculinos, en cualquier tipo de carretera (un sólo sentido, autovía, calzada única o vía de acceso), excepto Rotonda, cuando el vehículo patina, sufre volcadura o plegadura del remolque los fines de semana, es decir, sábados y domingos, en función a la maniobra vehicular y a la condición de iluminación de la carretera, se distinguen 2 patrones:

- **Regla 1:** cuando la maniobra que realiza el conductor consiste en cambiarse de carril a la derecha (es decir en adelantamientos) o cuando el vehículo se encuentra en una curva derecha y cuando hay ausencia de la luz del día, la probabilidad de accidente Fatal es del 81%.
- **Regla 2:** cuando la maniobra que realiza el conductor es uno de los siguientes: dar la vuelta o girar, mover hacia afuera (mover el vehículo hacia el borde de la carretera), cambio de carril a la izquierda, adelantar vehículo en movimiento o yendo por una curva izquierda, la probabilidad de accidente Fatal es del 70%.

Para conductores masculinos, cuando el vehículo sufre de plegadura del remolque y volcadura en el día o noche con iluminación desconocida, en función al día de la semana, maniobra vehicular y tipo de carretera, se distinguen 2 patrones:

- **Regla 3:** cuando el día es Domingo, Martes, Miércoles, Jueves o Sábado y el vehículo se encuentra por delante de otro y el tipo de carretera es Autovía o Calzada única, la probabilidad de accidente Fatal es del 60%.
- **Regla 4:** cuando la maniobra que realiza el conductor es uno de los siguientes: revertir, mover hacia afuera, cambio de carril, adelantar vehículo en movimiento o yendo por delante de otro vehículo y el día es Lunes o Viernes, la probabilidad de accidente Fatal es del 74%.

Para conductores femeninos los días Miércoles y Viernes, se distingue 1 patrón:

- **Regla 5:** cuando la maniobra que realiza la conductora es uno de los siguientes: revertir, mover hacia afuera, cambio de carril, adelantar vehículo en movimiento o el vehículo se encuentra delante de otro y el vehículo sufre de plegadura del remolque y volcadura en el día o noche con iluminación desconocida, la probabilidad de accidente Fatal es del 65%.

Y finalmente:

- **Regla 6:** cuando la maniobra que realiza el conductor es uno de los siguientes: revertir, mover hacia afuera, cambio de carril, adelantar vehículo en movimiento o el vehículo se encuentra delante de otro y el vehículo sufre de plegadura del remolque y volcadura en

la noche con luces encendidas, apagadas o sin luz, la probabilidad de accidente Fatal es del 77%.

Teniendo en cuenta los patrones encontrados, podemos concluir, que los accidentes dependen mucho de la maniobra que realiza el conductor y del suceso que ocurra después con el vehículo (patinaje, volcadura o plegadura del remolque). La probabilidad de ocurrencia de un accidente fatal aumenta cuando el conductor realiza maniobras como girar, moverse hacia el borde la pista, cambiarse de carril, adelantar un vehículo o cuando el vehículo se encuentra delante de otro, además después de esta maniobra, el vehículo sufre de plegadura del remolque y volcadura. Si estos sucesos ocurren en la noche, hay más probabilidades de accidente fatal que en otros casos o circunstancias.

4.3 Análisis de otra base de datos

Ahora trabajaremos con otra base de datos, que se obtuvo del portal web de datos abiertos del estado de Queensland (Australia).Utilizaremos el mismo procedimiento para obtener los patrones pertenecientes a los accidentes cuya severidad sea Fatal. El objetivo es doble: por un lado ver si los resultados son similares para otras bases de datos, y por otro lado, comprobar hasta qué punto el estudio realizado puede generalizarse a otros países o si por el contrario, cada base de datos requiere de un procesamiento y técnicas específicas.

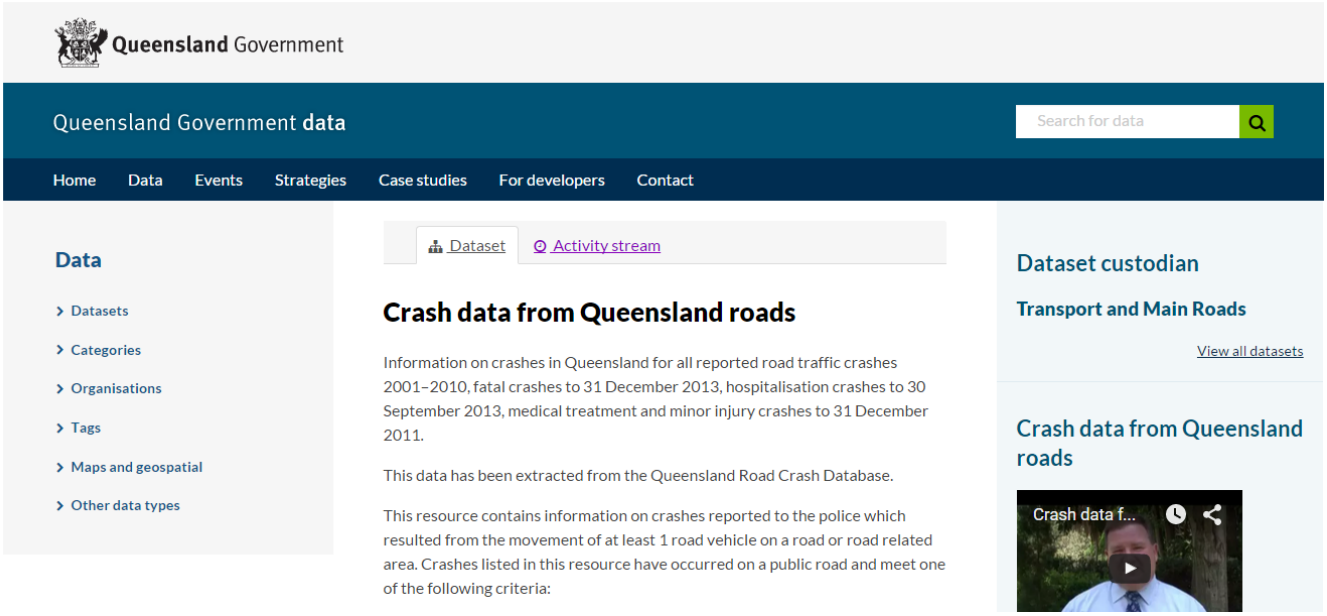


Ilustración 36. Portal web de datos libres del estado de Queensland (Australia)

Después de realizar el proceso de limpieza y transformación de los datos (similar a la realizada en la sección 3.2), contamos con una muestra de 13488 registros y 21 variables.

Seguidamente construiremos nuestros modelos aplicando las distintas técnicas de aprendizaje y los evaluaremos utilizando **validación cruzada de 10 Folds** (grupos).

La siguiente tabla muestra los resultados:

Técnica de aprendizaje	Acierto: clase "Fatal"	Acierto: clase "No fatal"	Accuracy
ADD	0.0000	1.0000	0.9042
K-NN	0.0569	0.9761	0.8964
NBs	0.3937	0.8269	0.7854
SVM	0.0243	1	0.9132
RF	0.0162	1	0.9125
AdaBoost	0.0207	0.9940	0.9008
RNA	0.0000	1.0000	0.9042

Tabla 12. Resultados de precisión de los modelos, evaluados con validación cruzada de 10 grupos, utilizando el dataset de Queensland (Australia).

Observando la Tabla 12, la mejor predicción lo obtenemos utilizando Naive Bayes, no obstante, nuestro objetivo radica en obtener una predicción aceptable con Árboles de decisión (para poder extraer reglas comprensibles). Se está trabajando con 21 variables y no todas son de vital importancia al momento de predecir la gravedad del accidente. Utilizaremos Weka para seleccionar los atributos relevantes.

En la siguiente tabla se muestra los resultados obtenidos:

Método de búsqueda	N° de atributos	Atributos
BestFirst	7	Crash_Nature Loc_ABS_Remoteness Crash_Roadway_Feature Crash_Speed_Limit Crash_Road_Surface_Condition Crash_Atmospheric_Condition Crash_Lighting_Condition
GeneticSearch	7	Crash_Nature Loc_ABS_Remoteness Crash_Traffic_Control Crash_Speed_Limit Crash_Road_Surface_Condition Crash_Atmospheric_Condition Crash_Lighting_Condition
GeneticSearch	6	Crash_Nature Loc_ABS_Remoteness Crash_Speed_Limit Crash_Road_Surface_Condition Crash_Atmospheric_Condition Crash_Lighting_Condition
GreedyStepwise	7	Crash_Nature Loc_ABS_Remoteness Crash_Roadway_Feature Crash_Speed_Limit Crash_Road_Surface_Condition Crash_Atmospheric_Condition

		Crash_Lighting_Condition
LinearForwardSelection	7	Crash_Nature Loc_ABS_Remoteness Crash_Roadway_Feature Crash_Speed_Limit Crash_Road_Surface_Condition Crash_Atmospheric_Condition Crash_Lighting_Condition
RandomSearch	9	Crash_Nature Loc_ABS_Remoteness Crash_Roadway_Feature Crash_Traffic_Control Crash_Speed_Limit Crash_Road_Surface_Condition Crash_Atmospheric_Condition Crash_Lighting_Condition Crash_Road_Horiz_Align
RankSearch	7	Crash_Nature Loc_ABS_Remoteness Crash_Traffic_Control Crash_Speed_Limit Crash_Road_Surface_Condition Crash_Atmospheric_Condition Crash_Lighting_Condition
ScatterSearchV1	7	Crash_Nature Loc_ABS_Remoteness Crash_Roadway_Feature Crash_Speed_Limit Crash_Road_Surface_Condition Crash_Atmospheric_Condition Crash_Lighting_Condition
SubsetSizeForwardSelection	7	Crash_Nature Loc_ABS_Remoteness Crash_Roadway_Feature Crash_Speed_Limit Crash_Road_Surface_Condition Crash_Atmospheric_Condition Crash_Lighting_Condition

Tabla 13. Subconjuntos de atributos seleccionados con el evaluador de atributos CfsSubsetEval, utilizando el dataset de Queensland (Australia).

Utilizaremos las variables obtenidas por la mayoría de los métodos, que son las siguientes: Crash_Nature, Loc_ABS_Remoteness, Crash_Roadway_Feature, Crash_Speed_Limit, Crash_Road_Surface_Condition, Crash_Atmospheric_Condition y Crash_Lighting_Condition. Por tal motivo, seleccionamos dicho subconjunto y construimos los modelos con las mismas técnicas de aprendizaje utilizadas anteriormente.

La siguiente tabla muestra los resultados después de la respectiva evaluación con validación cruzada de 10 pliegues:

Técnica de aprendizaje	**Acierto: clase "Fatal"**	**Acierto: clase "No fatal"**	**Accuracy**
ADD	0.0000	1.0000	0.9043
K-NN	0.3541	0.7976	0.7547
NBs	0.1014	0.9658	0.8831
SVM	0.0338	1.0000	0.9123
RF	0.0000	0.9991	0.9085
AdaBoost	0.0008	0.9998	0.9042
RNA	0.0000	1.0000	0.9043

Tabla 14. Resultados de precisión de los modelos, evaluados con validación cruzada de 10 grupos, después de aplicar la técnica de selección de atributos, utilizando el dataset de Queensland (Australia).

Como observamos en la Tabla 14, no mejoramos el modelo. Al igual que con la base de datos de Reino Unido, es muy probable que el dataset presente problemas de balanceo de datos.

table(accidentes_australia$gravedad_de_accidente)

Fatal	No Fatal
1291	12197

El cuadro anterior manifiesta un problema de balanceo de datos. Utilizaremos las técnicas de balanceo de datos SMOTE y ROSE y también se evaluará mediante validación cruzada de 10 grupos.

En la siguiente tabla se muestra los resultados obtenidos después de balancear los datos, y aplicar las diversas técnicas de aprendizaje.

Técnica de aprendizaje	Técnica de balanceo de datos	Acierto clase "Fatal"	Acierto clase "No fatal"	Accuracy	AUC
ADD	SMOTE	0.4662	0.6970	0.6749	0.5324
	ROSE	**0.5945**	**0.6831**	0.6747	**0.5535**
K.NN	SMOTE	0.5026	0.5951	0.5861	0.5183
	ROSE	0.6152	0.5798	0.5832	0.5344
NBs	SMOTE	0.4458	0.7040	0.6796	0.5305
	ROSE	**0.6636**	**0.6520**	0.6531	**0.5583**
SVM	SMOTE	0.4413	0.7448	0.7159	0.5401
	ROSE	0.7291	0.5531	0.5699	0.5501
RF	SMOTE	0.2369	0.8216	0.7655	0.5169
	ROSE	0.6989	0.5863	0.5973	0.5516
AdaBoost	SMOTE	0.7990	0.2871	0.3360	0.5387
	ROSE	0.2198	0.4802	0.4551	0.5524
RNA	SMOTE	0.2324	0.8350	0.7775	0.5187
	ROSE	**0.6896**	**0.6180**	0.6247	**0.5524**

Tabla 15. Resultados de precisión de los modelos, evaluados con validación cruzada de 10 grupos, después de aplicar la técnica de selección de atributos y balanceo de datos, utilizando el dataset de Queensland (Australia).

Como podemos ver en la Tabla 15, hay un cambio radical en los resultados, alcanzando una precisión de 59.45% para la clase "Fatal" y un 68.31% para la clase "No Fatal", utilizando Árboles de decisión.

Como último procedimiento evaluaremos el modelo con validación cruzada de 5 grupos y veremos si los resultados mejoran. En la siguiente tabla se muestran los resultados:

Técnica de aprendizaje	Técnica de balanceo de datos	Acierto clase "Fatal"	Acierto clase "No fatal"	Accuracy	AUC
ADD	SMOTE	0.4470	0.7065	0.6816	0.5312
	ROSE	**0.6057**	**0.6756**	0.6690	**0.5537**
K-NN	SMOTE	0.3756	0.6664	0.6394	0.5077
	ROSE	0.6888	0.5143	0.5311	0.5352
NBs	SMOTE	0.4021	0.7324	0.7007	0.5285
	ROSE	**0.6580**	**0.6548**	0.6550	**0.5577**
SVM	SMOTE	0.4154	0.7479	0.7159	0.5358
	ROSE	0.7269	0.5413	0.5593	0.5475
RF	SMOTE	0.3034	0.7854	0.7398	0.5214
	ROSE	0.7609	0.5036	0.5279	0.5484
AdaBoost	SMOTE	0.6201	0.4615	0.4782	0.5299
	ROSE	0.1072	0.6672	0.6138	0.5471
RNA	SMOTE	0.3175	0.7775	0.7342	0.5236
	ROSE	**0.6166**	**0.6478**	0.6451	0.5471

Tabla 16. Resultados de precisión de los modelos, evaluados con validación cruzada de 5 grupos, después de aplicar la técnica de selección de atributos y balanceo de datos, utilizando el dataset de Queensland (Australia).

Observando la Tabla 16, con validación cruzada de 5 grupos, se logra mejorar mínimamente el modelo con respecto al acierto de los accidentes con severidad “Fatal”, de un 59.45% a un 60.57%.

Hemos aplicado la misma metodología que con el dataset del Reino Unido (limpieza de valores nulos y anómalos, selección de atributos, aplicación de ROSE y SMOTE para el balanceado de las clases) y con ello hemos obtenido resultados similares: las redes bayesianas y las redes neuronales obtienen los mejores resultados (por clase y AUC), pero los árboles de decisión son también competitivos.

La recomendación tras este estudio es efectuar los siguientes pasos cuando se realice un estudio similar con otra base de datos si se quieren obtener reglas de decisión:

- Limpieza de valores nulos y anómalos (eliminando los registros)
- Selección de atributos (con cualquier método de los proporcionados por Weka)
- Aplicación del algoritmo ROSE para el balanceado de clases
- Usar la técnica de aprendizaje de árboles de decisión

Finalmente, a partir del modelo creado basado en Árboles de Decisión, extraeremos las reglas de decisión.

A continuación mostramos el árbol generado por el modelo:

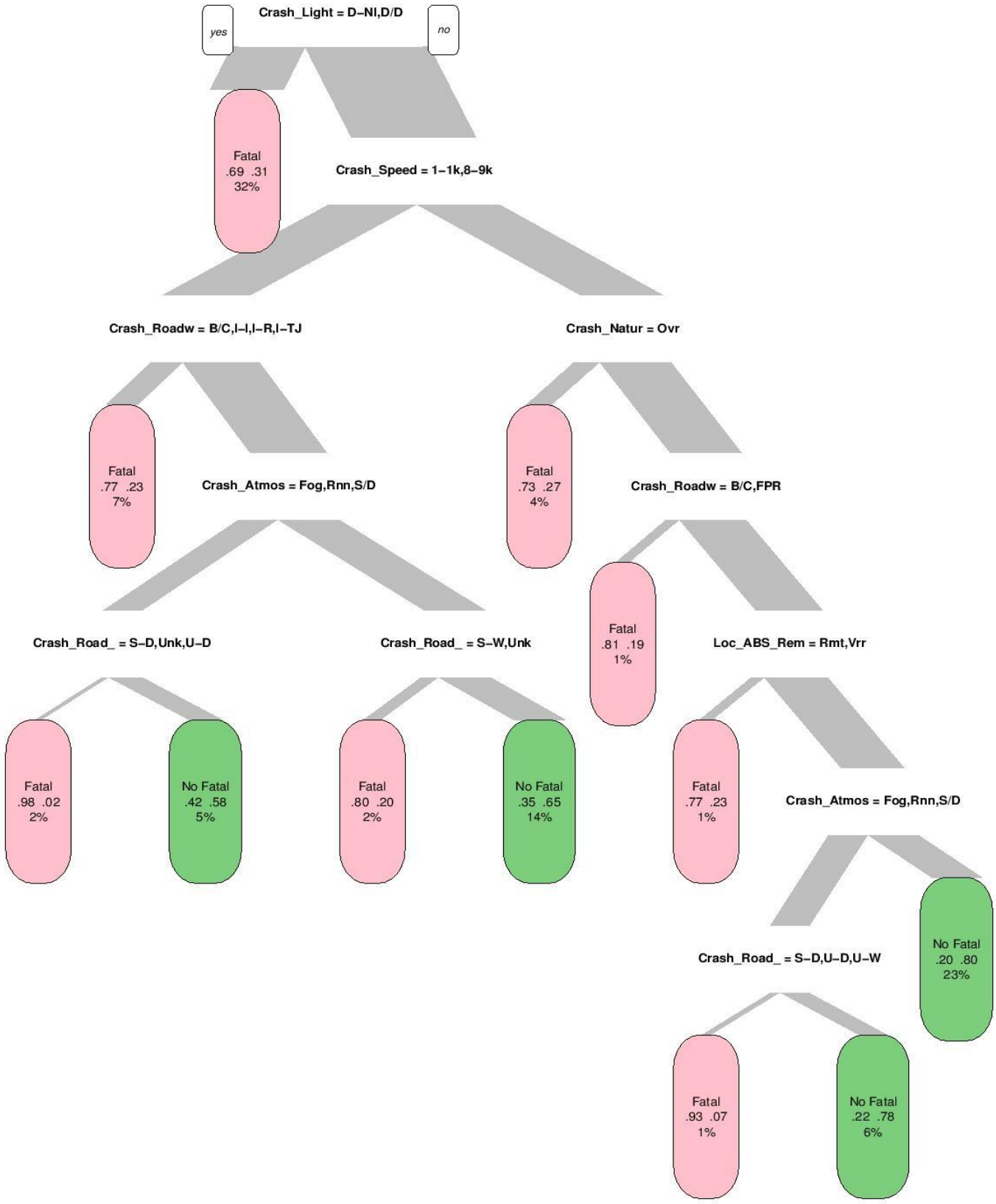


Ilustración 37. Árbol de decisión generado por el modelo (dataset: Queensland).

En el árbol se puede identificar sólo dos patrones para accidentes cuya severidad es Fatal. En la siguiente tabla se definen.

N°	Regla	Entonces
1	Si [(Crash_Lighting_Condition= “Darkness - Not Lighted” ó Crash_Lighting_Condition= “Dawn/Dusk”)]	gravedad_de_accidente = Fatal
2	Si [(Crash_Lighting_Condition= "Darkness - Lighted" ó Crash_Lighting_Condition= "Daylight") y (Crash_Speed_Limit= "100-110 km/h" ó Crash_Speed_Limit= "89-90 km/h") y (Crash_Roadway_Feature= "Bridge/Causeway" ó Crash_Roadway_Feature= "Intersection - Interchange" ó Crash_Roadway_Feature= "Intersection - Roundabout" ó Crash_Roadway_Feature= "Intersection - T Junction")]	gravedad_de_accidente = Fatal
3	Si [(Crash_Lighting_Condition= "Darkness - Lighted" ó Crash_Lighting_Condition= "Daylight") y (Crash_Speed_Limit= "100-110 km/h" ó Crash_Speed_Limit= "89-90 km/h") y (Crash_Roadway_Feature!= "Bridge/Causeway" y Crash_Roadway_Feature!= "Intersection - Interchange" y Crash_Roadway_Feature!= "Intersection - Roundabout" y Crash_Roadway_Feature!= "Intersection - T Junction") y (Crash_Atmospheric_Condition= "Fog" ó Crash_Atmospheric_Condition= "Raining" ó Crash_Atmospheric_Condition= "Smoke/Dust") y (Crash_Road_Surface_Condition= "Sealed - Dry" ó Crash_Road_Surface_Condition= "Unsealed - Dry")]	gravedad_de_accidente = Fatal
4	Si [(Crash_Lighting_Condition= "Darkness - Lighted" ó Crash_Lighting_Condition= "Daylight")	gravedad_de_accidente = Fatal

	y (Crash_Speed_Limit= "100-110 km/h" ó Crash_Speed_Limit= "89-90 km/h") y (Crash_Roadway_Feature!= "Bridge/Causeway" ó Crash_Roadway_Feature!= "Intersection - Interchange" ó Crash_Roadway_Feature!= "Intersection - Roundabout" ó Crash_Roadway_Feature!= "Intersection - T Junction") y (Crash_Atmospheric_Condition= "Clear") y (Crash_Road_Surface_Condition= "Sealed - Wet")]	
5	Si [(Crash_Lighting_Condition= "Darkness - Lighted" ó Crash_Lighting_Condition= "Daylight") y (Crash_Speed_Limit= "0-50 km/h" ó Crash_Speed_Limit= "60 km/h" ó Crash_Speed_Limit= "70 km/h") y (Crash_Nature= "Overturned")]	gravedad_de_accidente = Fatal
6	Si [(Crash_Lighting_Condition= "Darkness - Lighted" ó Crash_Lighting_Condition= "Daylight") y (Crash_Speed_Limit= "0-50 km/h" ó Crash_Speed_Limit= "60 km/h" ó Crash_Speed_Limit= "70 km/h") y (Crash_Nature!= "Overturned") y (Crash_Roadway_Feature= "Bridge/Causeway" ó Crash_Roadway_Feature= "Forestry/National Park Road")]	gravedad_de_accidente = Fatal
7	Si [(Crash_Lighting_Condition= "Darkness - Lighted" ó Crash_Lighting_Condition= "Daylight") y	gravedad_de_accidente = Fatal

	(Crash_Speed_Limit= "0-50 km/h" ó Crash_Speed_Limit= "60 km/h" ó Crash_Speed_Limit= "70 km/h") y (Crash_Nature!= "Overturned") y (Crash_Roadway_Feature!= "Bridge/Causeway" y Crash_Roadway_Feature!= "Forestry/National Park Road") y (Loc_ABS_Remoteness= "Remote" ó Loc_ABS_Remoteness= "Very remote")]	
8	Si [(Crash_Lighting_Condition= "Darkness - Lighted" ó Crash_Lighting_Condition= "Daylight") y (Crash_Speed_Limit= "0-50 km/h" ó Crash_Speed_Limit= "60 km/h" ó Crash_Speed_Limit= "70 km/h") y (Crash_Nature!= "Overturned") y (Crash_Roadway_Feature!= "Bridge/Causeway" y Crash_Roadway_Feature!= "Forestry/National Park Road") y (Loc_ABS_Remoteness!= "Remote" y Loc_ABS_Remoteness!= "Very remote") y (Crash_Atmospheric_Condition= "Fog" ó Crash_Atmospheric_Condition= "Raining" ó Crash_Atmospheric_Condition= "Smoke/Dust") y (Crash_Road_Surface_Condition= "Sealed - Dry" ó Crash_Road_Surface_Condition= "Unsealed - Dry" ó Crash_Road_Surface_Condition= "Unsealed - Wet")]	gravedad_de_accidente = Fatal

Tabla 17. Reglas de decisión extraídas del árbol de decisión (dataset: Australia)

Los patrones, que pertenecen a los accidentes cuya severidad es Fatal, que se pueden extraer son los siguientes:

Regla 1: Si el accidente ocurre al amanecer, anochecer o en la noche con ausencia de luz, la probabilidad de accidente Fatal es del 69%.

Cuando el accidente ocurre en el día o en la noche con iluminación y cuyo límite de velocidad de la carretera oscila entre 80 y 110 km/h, obtenemos los siguientes patrones:

- **Regla 2:** Si el accidente ocurre en un puente o en una intersección (intercambio, rotonda o unión T), la probabilidad de accidente Fatal es del 77%.
- **Regla 3:** Si el accidente ocurre en lugares diferentes a un puente o una intersección (intercambio, rotonda o unión T) y niebla o hay humo o polvo y la superficie de la carretera está seca, la probabilidad de accidente Fatal es del 69%.
- **Regla 4:** Si el accidente ocurre en lugares diferentes a un puente o una intersección (intercambio, rotonda o unión T) y el ambiente está claro y la superficie mojada, la probabilidad de accidente Fatal es del 80%.

Cuando el accidente ocurre en el día o en la noche con iluminación y cuyo límite de velocidad de la carretera oscila entre 0 y 70 km/h, obtenemos los siguientes patrones:

- **Regla 5:** Si el vehículo se vuelca, la probabilidad de accidente Fatal es del 73%.
- **Regla 6:** Si el vehículo se colisiona con algún objeto ó sufre alguna caída en un puente o carretera de un parque nacional, la probabilidad de accidente Fatal es del 81%.
- **Regla 7:** Si el vehículo se colisiona con algún objeto ó sufre alguna caída en una Intersección,ciclovía o cruce de ferrocarril y se encuentra en una zona remota o muy remota, la probabilidad de accidente Fatal es del 77%.
- **Regla 8:** Si el vehículo se colisiona con algún objeto ó sufre alguna caída en una Intersección, ciclovía o cruce de ferrocarril y no se encuentra en una zona remota, además hay lluvia, niebla, humo o polvo, y la carretera está seca (pavimentada o sin pavimentar) ó húmeda sin pavimentar, la probabilidad de accidente Fatal es del 93%.

Después de analizar los patrones y el gráfico mostrado (Ilustración 36), se concluye que los accidentes en este territorio (Queensland, Australia) dependen principalmente de las condiciones de iluminación de la carretera y del límite de velocidad que éstas poseen. La probalbilidad de ocurrencia de un accidente fatal es muy alta en carreteras secas o húmedas sin pavimentar con límite de velocidad de 70 km/h, cuando el vehículo colisiona con algún objeto o sufre alguna caída en una intersección, ciclovía o cruce de ferrocarril en zonas urbanas, con la presencia de lluvia, niebla, humo o polvo.

CAPÍTULO 5.
CONCLUSIONES Y TRABAJO FUTURO

CAPÍTULO 5.

CONCLUSIONES Y TRABAJO FUTURO

Los Árboles de decisión son una herramienta que permiten analizar los accidentes de tráfico de una manera sencilla y fácilmente comprensible para los analistas de la seguridad vial. Por tal motivo, los Árboles de decisión se presentan como un método alternativo a los modelos paramétricos debido a su capacidad para identificar patrones en los datos.

Los modelos de clasificación se pueden utilizar para determinar las interacciones entre las variables que serían imposibles de establecer directamente, utilizando las técnicas de modelización tradicionales.

Los Árboles de decisión permiten identificar determinadas reglas, potencialmente útiles, que pueden ser utilizadas por los analistas y gestores de seguridad vial para establecer determinadas contramedidas y/o acciones de carácter preventivo. En una primera instancia, los gestores pueden centrarse en los accidentes cuya severidad es "Fatal", y posteriormente intervenir en los accidentes "No Fatales". En este estudio se hace énfasis a los accidentes con severidad Fatal, identificando y describiendo los patrones obtenidos después de aplicar las técnicas de minería de datos.

Desde una perspectiva de la gestión de la seguridad vial, se destacan las conclusiones generales sobre los resultados particulares obtenidos:

Para aquellos accidentes ocurridos en Reino Unido:

- Los accidentes cuya severidad es Fatal son producidos fundamentalmente por conductores de sexo masculino.
- En el árbol generado, se puede observar que la variable raíz es la "maniobra vehicular", seguidamente la variable que genera las dos grandes ramas, es la variable "patinaje volcadura". Por tales hechos, se puede concluir que la severidad de los accidentes depende principalmente de la maniobra realizada por el conductor y, si el vehículo sufre de volcadura, patinaje o plegadura del remolque .
- La probabilidad de un accidente Fatal es alta, en cualquier tipo de carretera (excepto en una rotonda) los fines de semana cuando el conductor es hombre y trata de cambiarse al carril derecho o cuando se encuentra en una curva derecha, con ausencia de luz del día y el vehículo patina, sufre de volcadura o plegadura del remolque.
- Cuando una conductora está involucrada en un accidente, el método predice un accidente Fatal los miércoles y viernes en el día o noche con iluminación desconocida, después de realizar una de las siguientes maniobras: revertir, mover hacia afuera, cambio de carril, adelantar vehículo en movimiento o el vehículo se encuentra delante de otro y el vehículo sufre de plegadura del remolque y volcadura.

Para accidentes ocurridos en Queensland (Australia):

- Las condiciones de iluminación son un factor vital para la severidad de los accidentes. En el árbol formado, representan la raíz y claramente se puede observar que si el accidente ocurre en circunstancias de ausencia de luz, la gravedad se vuelve Fatal.
- La probabilidad de un accidente Fatal aumenta, cuando el vehículo colisiona con algún objeto o sufre alguna caída en una intersección, ciclovía o cruce de ferrocarril en zonas no alejadas de la ciudad, además de la existencia de lluvia, niebla, humo o polvo en carreteras secas o húmedas sin pavimentar.

Comparando los resultados obtenidos de ambas bases de datos, podemos identificar que las condiciones de iluminación coinciden en ambos estudios, por lo tanto, son de carácter relevante en la ocurrencia de los accidentes fatales. Se debe tener en cuenta que ambas bases de datos no cuentan con los mismos atributos, y que por tal motivo, las causas de los accidentes varían en los resultados que se obtienen.

Se debe destacar que el método con mayor precisión a la hora de predecir la severidad del accidente fue las Máquinas de soporte vectorial, como se indica en la Tabla 10, no obstante, no se puede extraer las reglas de decisión, que es el principal objetivo de este estudio.

Como trabajo futuro, se propone lo siguiente:

- Tratar de aumentar el Accuracy (precisión del modelo), agregando más información a la base de datos mediante atributos que estén relacionados con la gravedad del accidente.
- En la presente investigación, han sido analizados sólo los accidentes ocurridos con un vehículo, ya que por lo general la gravedad es mayor. Se pretende ampliar el estudio para el caso de accidentes múltiples.
- El estudio se ha enfocado a obtener patrones de accidentes con severidad fatal, no obstante, se pretende también estudiar los patrones de accidentes no fatales, con el objetivo de disminuir la frecuencia de los mismos.

Finalmente, se puede destacar que en esta tesis de máster se han analizado las problemáticas concretas de las carreteras de las zonas rurales del Reino Unido y carreteras de Queensland (Australia), sin embargo, la metodología utilizada puede ser aplicada para el estudio de las problemáticas existentes en otras vías de tránsito.

CAPÍTULO 6.
REFERENCIAS BIBLIOGRÁFICAS

CAPÍTULO 6.

REFERENCIAS BIBLIOGRÁFICAS

Abdelwahab, H., & Abdel-Aty, M. (2001). Development of Artificial Neural Network Models to Predict Driver Injury Severity in Traffic Accidents at Signalized Intersections. *Transportation Research Record: Journal of the Transportation Research Board, 1746*.

de Oña, J., López, G., Mujalli, R., & Calvo, F. (2013). Analysis of traffic accidents on rural highways using Latent Class Clustering and Bayesian Networks. *Accident Analysis & Prevention, 51*.

de Oña, J., Oqab Mujalli, R., & Calvo, F. (2011). Analysis of traffic accident injury severity on Spanish rural highways using Bayesian networks. *Accident Analysis & Prevention, 43*.

Depaire, B., Wets, G., & Vanhoof, K. (2008). Traffic accident segmentation by means of latent class clustering. *Accident Analysis & Prevention, 40*.

Flach, P. A., Hernandez-Orallo, J., & Ferri, C. (2011). A coherent interpretation of AUC as a measure of aggregated classification performance. *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, (págs. 657-664).

Flores López, R., & Fernández Fernández, J. M. (2008). Las Redes Neuronales Artificiales: Fundamentos teóricos y aplicaciones prácticas. La Coruña: Netbiblo, S.L.

Heras Martínez, A., Tolmos, P., & Hernández-March, J. (2010). *UN ANÁLISIS COMPARATIVO DE UNA SVM Y UN MODELO.* Madrid.

López Maldonado, G. (2013). *Análisis de la Severidad de Accidentes de Tráfico, utilizando Técnicas de Minería de Datos.* Granada: Editorial de la Universidad de Granada.

López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences, 250*, 113-141.

Moreo, J., Rodríguez, D., Sicilia, M., Riquelme, J., & Ruiz, R. (2009). SMOTE-I: mejora del algoritmo SMOTE para balanceo de clases minoritarias. *Actas de los Talleres de las Jornadas de Ingeniería del Software y Bases de Datos, 3*.

Orozco Guillén, E. E., Iruretagoyena Garcia, G., Vazquez y Montiel, S., Delgado-Atencio, J. A., Castro Ramos, J., & Gutierrez Delgado, F. (2010). Métodos de clasificación para identificar lesiones. *Revista Ingeniería Biomédica, 4*.

R Development Core Team. (16 de Mayo de 2000). *Notas sobre R: Un entorno de programación para Análisis de Datos y Gráficos.* Obtenido de The Comprehensive R Archive Network: https://cran.r-project.org/

The R Foundation. (8 de Setiembre de 2015). *The R Project for Statistical Computing*. Obtenido de https://www.r-project.org/

Weiss, S. M., & Indurkhya, N. (1998). *Predictive Data Mining: A Practical Guide.*

Wikipedia. (30 de Abril de 2015). *Validación cruzada*. Obtenido de Wikipedia: https://es.wikipedia.org/wiki/Validación_cruzada

Wikipedia. (18 de Agosto de 2015). *Weka (aprendizaje automático)*. Obtenido de Wikipedia: https://es.wikipedia.org/

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques.* San Francisco: Elsevier Inc.