

ANÁLISIS DE ACCIDENTES DE TRÁFICO

Proyecto de Minería de datos y el paradigma Big Data



CARLA PAOLA PEÑARRIETA URIBE
NOELIA CALDERÓN MATEO
ÁLVARO DELGADO GUTIÉRREZ

Índice

1. Introducción

2. Objetivos

3. Metodología

3.1. Abstracción del escenario

3.2. Selección de los datos

3.3. Limpieza y preprocesamiento

3.4. Transformación de los datos

3.5. Selección de la apropiada tarea de minería de datos

3.6. Elección del algoritmo de minería de datos

3.7. Aplicación del algoritmo

3.8. Evaluación

3.9. Aplicación

4. Desarrollo

4.1. Selección de los datos

4.2. Limpieza y preprocesamiento

4.3. Transformación de los datos

4.4. Exploración

4.5. Minería de datos

5. Resultados y evaluación

6. Conclusiones

7. Bibliografía

Introducción

Con este trabajo pretendemos conocer las causas que generan el mayor número de accidentes de tráfico en Madrid. Aprender a manejar las técnicas de minería de datos y comprobar cuáles son las más indicadas para obtener patrones confiables de los accidentes.

El objetivo es construir un modelo de predicción para tomar medidas necesarias o que ayuden a evitar que se produzcan dichos accidentes o a reducir las consecuencias. Nos basaremos en variables como puede ser el clima en el momento del accidente o, por ejemplo, si ese día fue festivo.

Se procesarán los datos del año 2017 y se generarán gráficas que ayuden a sacar conclusiones.

Para realizar este trabajo utilizaremos como herramienta “R”, que es un lenguaje de programación especializado en el análisis estadístico de datos.

Objetivos

El principal objetivo consiste en el reconocimiento de patrones para intentar disminuir la cantidad de accidentes que se producen en Madrid.

Esto puede conseguirse con la mejora de carreteras, de los vehículos y, sobre todo, creando conciencia y modificando los hábitos de comportamiento de los conductores y demás usuarios de la vía pública.

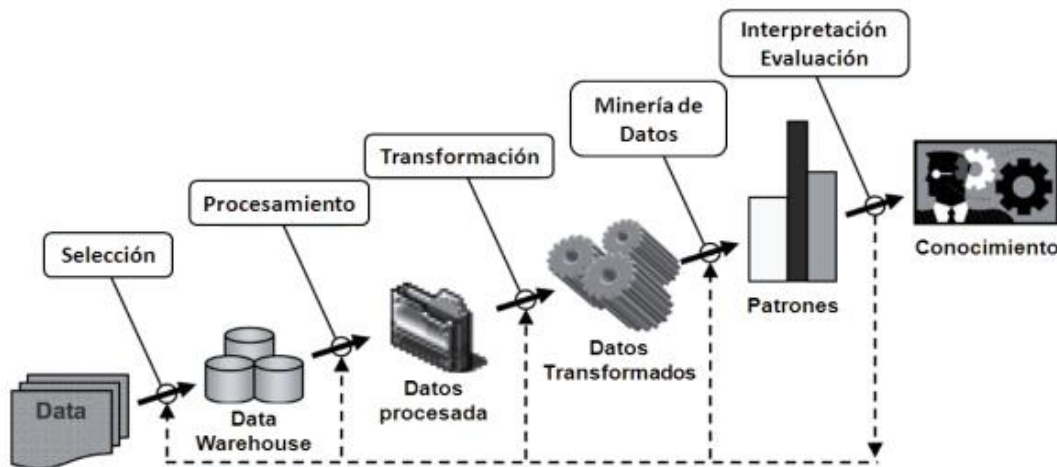
Trabajando y analizando las distintas variables de los data sets, trataremos de averiguar cuáles son las mayores causas de que se produzcan los accidentes para poder tomar las medidas necesarias.

Para conseguir este objetivo procederemos a realizar varios puntos específicos: recogida de data sets para obtener la información necesaria, limpieza de estos datos para su procesamiento, identificar variables útiles, realizar un análisis estadístico de los accidentes para obtener información relevante sobre bajo qué condiciones suelen ocurrir con mayor frecuencia y, usaremos también, técnicas no supervisadas como las reglas de asociación.

Metodología

La metodología que trataremos de seguir en este proyecto es *Knowledge Discovery in Databases*, KDD. Ayuda a procesar grandes cantidades de datos para encontrar conocimiento útil en ellos.

El proceso se compone de cinco etapas (selección, procesamiento, transformación, minería de datos e interpretación y evaluación) descritas a continuación.



Abstracción del escenario

Para conseguir nuestro objetivo debemos tener claro dónde queremos enfocar el análisis dentro de las limitaciones que tendremos con los datos obtenidos. Para ello es necesario entender el problema al que nos enfrentamos, conocer las restricciones y establecer un objetivo.

Estudiaremos claves importantes como el clima, la fecha y la vía en el momento del accidente.

También tendremos en cuenta otros aspectos como la edad o el género a la hora de analizar los distintos problemas que causan estos accidentes.

Nos centraremos en un conjunto de data sets de accidentes en Madrid en el año 2017 recogidos previamente de la página de datos abiertos de Madrid. Estos datos contendrán distintos atributos con toda la información necesaria.

Selección de los datos

Se definen las fuentes de datos que serán útiles para la realización del proyecto. Lo normal es que los datos provengan de distintos sitios y con distintas fuentes.

Cada fuente se compone de distintas variables. Dependiendo del problema nos centraremos en una en concreto. Por ejemplo, si consideramos que el clima es un

factor fundamental para que se produzca un accidente, entonces nos centraremos en variables como el clima, la fecha o el estado de la calzada.

Una vez recopilados los datos, tenemos que integrarlos y prepararlos para su procesamiento.

Limpieza y preprocesamiento

Hay que revisar el contenido de los datos ya que tienen diferentes formatos y fuentes. Es importante la limpieza de los datos para no conducir el trabajo a resultados inválidos y poco fiables.

Dado que los datos contienen diversas variables, es necesario eliminar aquellas que no sean necesarias o sean erróneas.

Esta es la fase en la que se corrigen los data sets, se eliminan duplicados y los datos aislados y se decide qué hacer con cada dato.

Además, se determinan aquellas variables que serán relevantes, para centralizarse en ello y simplificar el problema.

Una vez realizada esta fase, se realiza la transformación de los datos.

Transformación de los datos

En esta etapa se podrá seleccionar los datos concretos a utilizar, trasladar a la aplicación y la reducción de la dimensionalidad en el caso de que existan variables que no influyan en el análisis.

Selección de la apropiada tarea de Minería de Datos

Esta fase se refiere a elegir la tarea que más nos conviene para realizar nuestro trabajo.

Existen diferentes técnicas de minería de datos o algoritmos que podemos separar en supervisados o predictivos (se caracterizan por partir de ejemplos. Existe un atributo especial o clase en los casos, que especifica si un caso pertenece o no a un concepto.) y no supervisados o de descubrimiento (el aprendizaje en este caso se realiza mediante observación. Se construyen hipótesis a partir de un conjunto de datos).

Algunas de las técnicas de cada categoría son:

- **Supervisados:** Árboles de decisión, Bosques aleatorios, Redes neuronales, etc.
- **No Supervisados:** Segmentación, Reglas de asociación, Agrupamiento (clustering), etc.

Elección del algoritmo de Minería de Datos

El objetivo de este trabajo es usar algunas de estas técnicas para aplicarlas al tema de los accidentes de tráfico. Trataremos de realizar predicciones y sacar conclusiones que puedan ayudar a prevenir esos accidentes.

Se escogerán distintas técnicas dependiendo de los atributos con las que queramos aplicarlas.

- **Agrupación**

El clustering o agrupamiento es el proceso de particionar un conjunto de datos (u objetos) en un conjunto de subclases significativas llamadas grupos o clústeres. Un grupo es una colección de objetos de datos que son similares a otros y así pueden ser tratados colectivamente como un grupo.

Es una forma de clasificación no supervisada en la que, a diferencia de la supervisada, no se conocen las etiquetas de las clases.

Los algoritmos más populares de esta técnica son PAM, CLARA y CLARANS, que usan k-means y k-medoids.

- **Reglas de asociación**

Las reglas de asociación en la minería de datos se usan para encontrar hechos que ocurren en común dentro de un conjunto de datos. Para ello se combinan condiciones para que haya una consecuencia.

$$(\text{Antecedente}) \ X \Rightarrow Y \ (\text{Consecuente})$$

El algoritmo más frecuente es **A PRIORI**, diseñado para operar en bases de datos transicionales. Se basa en el conocimiento previo de los conjuntos frecuentes. Esto sirve para reducir el espacio de búsqueda y aumentar la eficiencia de los resultados.

Se utilizan 3 parámetros para medir el interés de una regla:

- ★ **Support:** mide la cantidad de veces que aparecen los ítems de dicha regla en la base de datos.
- ★ **Confidence:** es el porcentaje de veces que, apareciendo en una instancia los ítems del antecedente, aparecen también los del consecuente.
- ★ **Lift:** indica qué probabilidad existe de encontrar el consecuente limitando la búsqueda a aquellos conjuntos de ítems donde el antecedente está presente.

Aplicación del algoritmo

Este es el momento de aplicar los algoritmos para ver que resultados obtenemos. La aplicación de estos la realizamos en el apartado *Desarrollo*.

Evaluación

Esta parte de la metodología se realiza después de aplicar los algoritmos. Se evalúan los patrones que se han generados y el rendimiento. Ver apartado *Resultados y Evaluación*.

Aplicación

Si todos los pasos han resultado satisfactorios en esta última fase podemos aplicar el conocimiento obtenido y sacar conclusiones. Ver apartado *Conclusión*.

Desarrollo

Selección de los datos

Los datos recogidos son dos data sets con los accidentes ocurridos en Madrid en el año 2017, uno de ellos en el centro y otro en las carreteras de las afueras con distintas variables que proporcionan información sobre dónde, cuándo y cómo se produjo cada accidente, y otro data set del calendario laboral de ese mismo año que nos servirá para analizar si una de las causas es que el día sea festivo o no.

Data set calendario laboral madrileño

	A	B	C
1	Fecha	laboral	
2	01/01/2017	domingo	
3	02/01/2017	laborable	
4	03/01/2017	laborable	
5	04/01/2017	laborable	
6	05/01/2017	laborable	
7	06/01/2017	festivo	
8	07/01/2017	sabado	
9	08/01/2017	domingo	
10	09/01/2017	laborable	
11	10/01/2017	laborable	
12	11/01/2017	laborable	
13	12/01/2017	laborable	
14	13/01/2017	laborable	
15	14/01/2017	sabado	
16	15/01/2017	domingo	
17	16/01/2017	laborable	
18	17/01/2017	laborable	
19	18/01/2017	laborable	
20	19/01/2017	laborable	

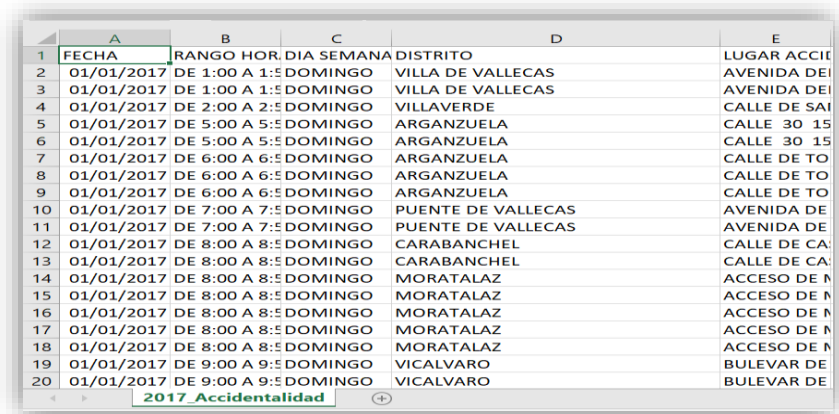
calendario_solo_2017

Data set de accidentes en la M-30

	A	B	C	D
1	Fecha	Hora	Dia	Carretera
2	01/01/2017	5:48	Domingo	M30
3	01/01/2017	7:05	Domingo	M30
4	01/01/2017	23:26	Domingo	M30
5	02/01/2017	23:55	Lunes	M30
6	03/01/2017	14:45	Martes	M30
7	03/01/2017	17:03	Martes	M30
8	03/01/2017	18:25	Martes	M30
9	03/01/2017	18:32	Martes	M30
10	04/01/2017	12:35	Miércoles	M30
11	04/01/2017	19:35	Miércoles	M30
12	06/01/2017	13:45	Viernes	M30
13	06/01/2017	14:50	Viernes	M30
14	06/01/2017	20:25	Viernes	M30
15	08/01/2017	10:09	Domingo	M30
16	08/01/2017	15:07	Domingo	M30
17	08/01/2017	19:08	Domingo	M30
18	09/01/2017	6:52	Lunes	A5
19	09/01/2017	7:24	Lunes	M30
20	09/01/2017	9:07	Lunes	M607

Listado_accidentes_2017_filtrad

Data set de los accidentes en los barrios de Madrid capital



	A	B	C	D	E
	FECHA	RANGO HOR	DÍA SEMANA	DISTRITO	LUGAR ACCI
1	01/01/2017	DE 1:00 A 1:5	DOMINGO	VILLA DE VALLECAS	AVENIDA DE
2	01/01/2017	DE 1:00 A 1:5	DOMINGO	VILLA DE VALLECAS	AVENIDA DE
3	01/01/2017	DE 2:00 A 2:5	DOMINGO	VILLAVERDE	CALLE DE SAI
4	01/01/2017	DE 5:00 A 5:5	DOMINGO	ARGANZUELA	CALLE 30 15
5	01/01/2017	DE 5:00 A 5:5	DOMINGO	ARGANZUELA	CALLE 30 15
6	01/01/2017	DE 6:00 A 6:5	DOMINGO	ARGANZUELA	CALLE DE TO
7	01/01/2017	DE 6:00 A 6:5	DOMINGO	ARGANZUELA	CALLE DE TO
8	01/01/2017	DE 6:00 A 6:5	DOMINGO	ARGANZUELA	CALLE DE TO
9	01/01/2017	DE 6:00 A 6:5	DOMINGO	ARGANZUELA	CALLE DE TO
10	01/01/2017	DE 7:00 A 7:5	DOMINGO	PUENTE DE VALLECAS	AVENIDA DE
11	01/01/2017	DE 7:00 A 7:5	DOMINGO	PUENTE DE VALLECAS	AVENIDA DE
12	01/01/2017	DE 8:00 A 8:5	DOMINGO	CARABANCHEL	CALLE DE CA
13	01/01/2017	DE 8:00 A 8:5	DOMINGO	CARABANCHEL	CALLE DE CA
14	01/01/2017	DE 8:00 A 8:5	DOMINGO	MORATALAZ	ACCESO DE N
15	01/01/2017	DE 8:00 A 8:5	DOMINGO	MORATALAZ	ACCESO DE N
16	01/01/2017	DE 8:00 A 8:5	DOMINGO	MORATALAZ	ACCESO DE N
17	01/01/2017	DE 8:00 A 8:5	DOMINGO	MORATALAZ	ACCESO DE N
18	01/01/2017	DE 8:00 A 8:5	DOMINGO	MORATALAZ	ACCESO DE N
19	01/01/2017	DE 9:00 A 9:5	DOMINGO	VICALVARO	BULEVAR DE
20	01/01/2017	DE 9:00 A 9:5	DOMINGO	VICALVARO	BULEVAR DE

Limpieza y preprocesamiento

Al revisar los datos, nos damos cuenta de que hay columnas que requieren de una limpieza previa para poder usarlas en los algoritmos utilizados en el proyecto.

Las variables requieren una limpieza ya que pueden tener caracteres erróneos, espacios, datos en blanco o no asignados e incluso, palabras únicamente distinguidas con mayúsculas y minúsculas que hay que corregir.

A continuación, explicamos la limpieza que hemos realizado en cada uno de los data sets:

- *Limpieza de listado_accidentes_2017.csv*

En las columnas de tipo numérico en las que había NA's, hemos rellenado con 0 para obtener resultados correctos.

```
datos_accidentes$Total_Implicados[is.na(datos_accidentes$Total_Implicados)] = 0
datos_accidentes$Turismos[is.na(datos_accidentes$Turismos)] = 0
datos_accidentes$Motocicletas[is.na(datos_accidentes$Motocicletas)] = 0
datos_accidentes$Pesados[is.na(datos_accidentes$Pesados)] = 0
```

Para poder utilizar la columna de los días de la semana "Dia" en las técnicas de minería, hemos cambiado los datos a tipo numérico según el día de la semana.

Como vemos aquí: lunes es 1, martes 2, etc.

```
datos_accidentes$Dia[datos_accidentes$Dia == "Lunes" | datos_accidentes$Dia == "lunes"] = 1
datos_accidentes$Dia[datos_accidentes$Dia == "Martes" | datos_accidentes$Dia == "martes"] = 2
datos_accidentes$Dia[datos_accidentes$Dia == "Miércoles" | datos_accidentes$Dia == "miércoles"] = 3
datos_accidentes$Dia[datos_accidentes$Dia == "Jueves" | datos_accidentes$Dia == "jueves"] = 4
datos_accidentes$Dia[datos_accidentes$Dia == "Viernes" | datos_accidentes$Dia == "viernes"] = 5
datos_accidentes$Dia[datos_accidentes$Dia == "Sábado" | datos_accidentes$Dia == "sábado"] = 6
datos_accidentes$Dia[datos_accidentes$Dia == "Domingo" | datos_accidentes$Dia == "domingo"] = 7
datos_accidentes$Dia<-as.numeric(datos_accidentes$Dia)
```

Algunas variables estaban distinguidas únicamente por alguna mayúscula/minúscula. Para que no haga esa distinción entre palabras que son iguales, tenemos que ponerlas en el mismo formato. Por tanto, hemos cambiado todo a mayúsculas.


```
datos_accidentes$Vertido[datos_accidentes$Vertido == "No"] = "NO"
datos_accidentes$Vertido[datos_accidentes$Vertido == "Si"] = "SI"
datos_accidentes$Vertido[is.na(datos_accidentes$Vertido) | datos_accidentes$Vertido == ""] = "NN"
```

También hemos cambiado el tipo de algunas variables como, por ejemplo, la fecha a tipo **Date**, variables numéricas a **Numeric** y algunas otras a **Factor** u otros tipos según lo necesario.

- *Limpieza calendario_solo_2017.csv*

En este data set, solo era necesario modificar la fecha a tipo Date. Primero, cambiamos los NA's a "No asignado" y después dejamos el tipo correspondiente para la fecha.

```
calendario$Fecha<-as.character(calendario$Fecha)
calendario$Fecha[is.na(calendario$Fecha)]= "NO ASIGNADA"
calendario$Fecha<-as.Date(calendario$Fecha,"%d/%m/%Y")
```

- *Limpieza de 2017_Accidentalidad.csv*

Trabando con este data set, nos encontramos con un fallo a la hora de filtrar aplicando reglas de asociación. Nos daba problemas el hecho de que las variables contenían numerosos espacios al final de cada palabra. Para solucionarlo tuvimos que eliminar los espacios de algunas columnas.

```
datos_accidentes2$LESIVIDAD<- gsub(" ", "", datos_accidentes2$LESIVIDAD)
datos_accidentes2$DISTRITO<- gsub(" ", "", datos_accidentes2$DISTRITO)
datos_accidentes2$Tipo_Vehiculo<- gsub(" ", "", datos_accidentes2$Tipo_Vehiculo)
datos_accidentes2$TIPO.PERSONA<- gsub(" ", "", datos_accidentes2$TIPO.PERSONA)
datos_accidentes2$tipo_accidente<-gsub(" ", "", datos_accidentes2$tipo_accidente)
```

Para no trabajar con rangos de hora amplios, agrupamos las horas según el momento del día: *madrugada* (de 00:00 a 06:59), *mañana* (de 07:00 a 11:59), *mediodía* (12:00 a 15:59), *tarde* (de 16:00 a 19:59) y *noche* (20:00 a 23:59).

```
datos_accidentes2$RANGO.HORARIO[datos_accidentes2$RANGO.HORARIO == "DE 10:00 A 10:59"] = "MAÑANA"
datos_accidentes2$RANGO.HORARIO[datos_accidentes2$RANGO.HORARIO == "DE 11:00 A 11:59"] = "MAÑANA"
datos_accidentes2$RANGO.HORARIO[datos_accidentes2$RANGO.HORARIO == "DE 12:00 A 12:59"] = "MEDIODIA"
datos_accidentes2$RANGO.HORARIO[datos_accidentes2$RANGO.HORARIO == "DE 13:00 A 13:59"] = "MEDIODIA"
datos_accidentes2$RANGO.HORARIO[datos_accidentes2$RANGO.HORARIO == "DE 14:00 A 14:59"] = "MEDIODIA"
datos_accidentes2$RANGO.HORARIO[datos_accidentes2$RANGO.HORARIO == "DE 15:00 A 15:59"] = "MEDIODIA"
datos_accidentes2$RANGO.HORARIO[datos_accidentes2$RANGO.HORARIO == "DE 16:00 A 16:59"] = "TARDE"
datos_accidentes2$RANGO.HORARIO[datos_accidentes2$RANGO.HORARIO == "DE 17:00 A 17:59"] = "TARDE"
datos_accidentes2$RANGO.HORARIO[datos_accidentes2$RANGO.HORARIO == "DE 18:00 A 18:59"] = "TARDE"
```

Al igual que el primer data set, los días de la semana los hemos convertido a tipo numérico siguiendo el mismo proceso.

Modificamos el tipo de algunas variables y eliminamos los NA's.

Transformación de los datos

El primer paso es convertir el fichero en un formato que pueda cargar la herramienta que vamos a usar (R Studio). Se guardan como CSV separados por comas.

Una vez preparado el formato, se cargan los datos en el programa:

```
#CSV utilizados
datos_accidentes2<-read.csv("2017_Accidentalidad.csv", sep = ";", stringsAsFactors = TRUE)
datos_accidentes<-read.csv("Listado_accidentes_2017_filtrado.csv", sep = ";", stringsAsFactors = FALSE)
calendario<- read.csv("calendario_solo_2017.csv", sep = ";")
```

Se transforman los datos para prepararlos para la implementación, se crean data frames, se mezclan los data sets y se añaden o eliminan columnas según sea necesario para realizar la tarea apropiada para el análisis.

Por ejemplo, creamos varios data frame combinando los data sets.

En el calendario venían todos los días del año, mientras que en los listados de accidentes solo venían las fechas de los días que había accidente.

Con todo esto, obtuvimos dos tablas (una para el interior y otra para carreteras exteriores) con las fechas que hubo accidentes, cuántos accidentes hubo en cada fecha y si ese día fue festivo, laboral o fin de semana.

Lo implementamos con funciones como *table* y *merge*.

```
calendario_accidentes<- merge(t, df, by.x = "fecha", by.y = "calendario.Fecha", all.x = TRUE)
```

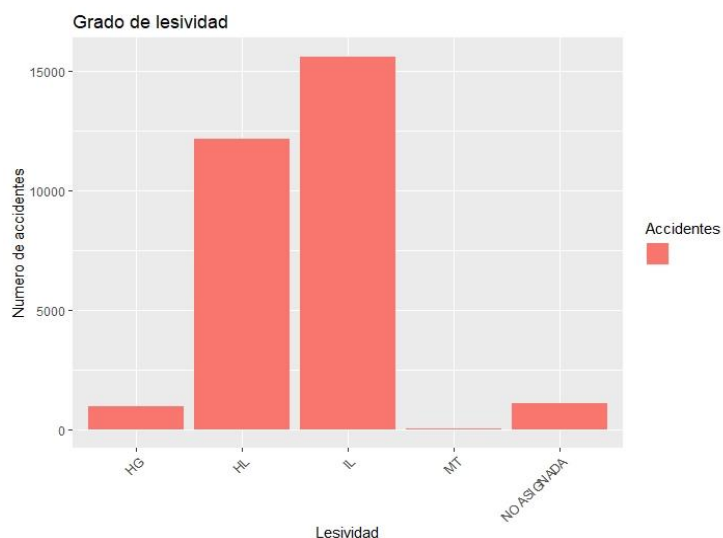
En nuestro estudio, solo empleamos la normalización para calcular los accidentes por día laboral, fin de semana o festivo. Esto se debe a que nuestros datos numéricos se mantienen durante todo el estudio en la misma escala. Además, el único momento en el que realmente tratamos con números es en el clustering (donde las variables no superan las decenas). En el resto del estudio las variables utilizadas son en su mayoría cualitativas.

Exploración

En este apartado nos hemos centrado en los datos que creíamos que nos aportaban más información y, por lo tanto, conseguir algunas premisas para guiarnos en la aplicación de las técnicas de minería de datos. A continuación, mostramos algunas de las graficas que nos han guiado a lo largo del estudio.

Gráfica 1: Grado de lesividad

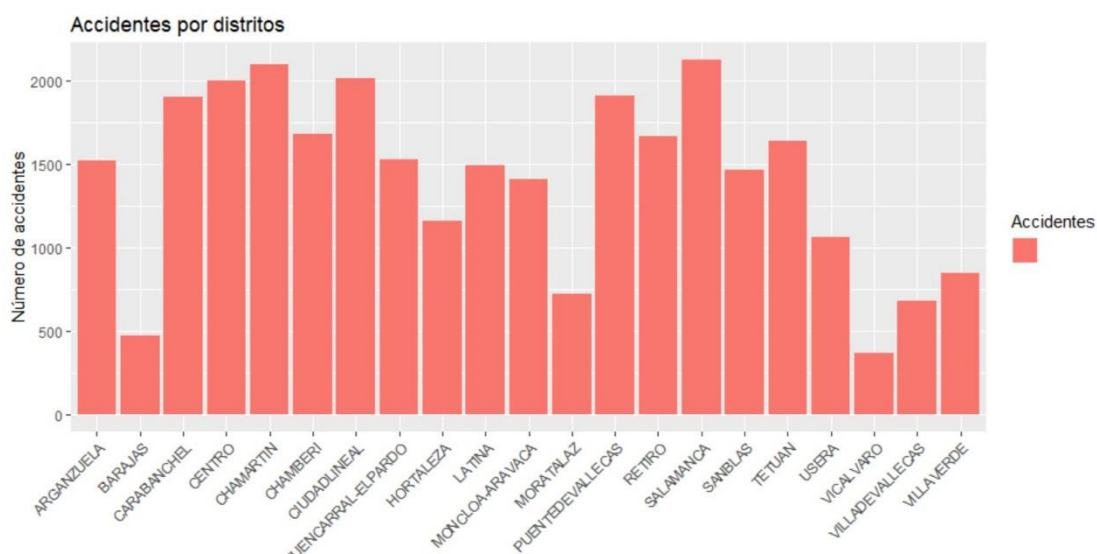
Esta gráfica nos indica que en la mayor parte de los accidentes las personas salen ilesas (**IL**). Como lo que nos interesa saber son los factores que influyen en los accidentes de tráfico con heridos, nos centraremos en la variable de heridos leves (**HL**).



Gráfica 2: Accidentes por distritos

La siguiente gráfica indica la cantidad de accidentes que ha habido por distritos. En un principio, pensamos que un factor clave a la hora de implementar las reglas de asociación sería el saber si en algún distrito en concreto se producían más accidentes para evaluar cual podría ser el motivo.

En cambio, observamos que la cantidad de accidentes es parecida en todos los distritos y, las variaciones se deben a la diferencia de afluencia de personas principalmente. En las zonas donde abunda la circulación de coches y de personas el número de accidentes es mayor, por lo tanto, no es un factor que nos pueda proporcionar mucha información.

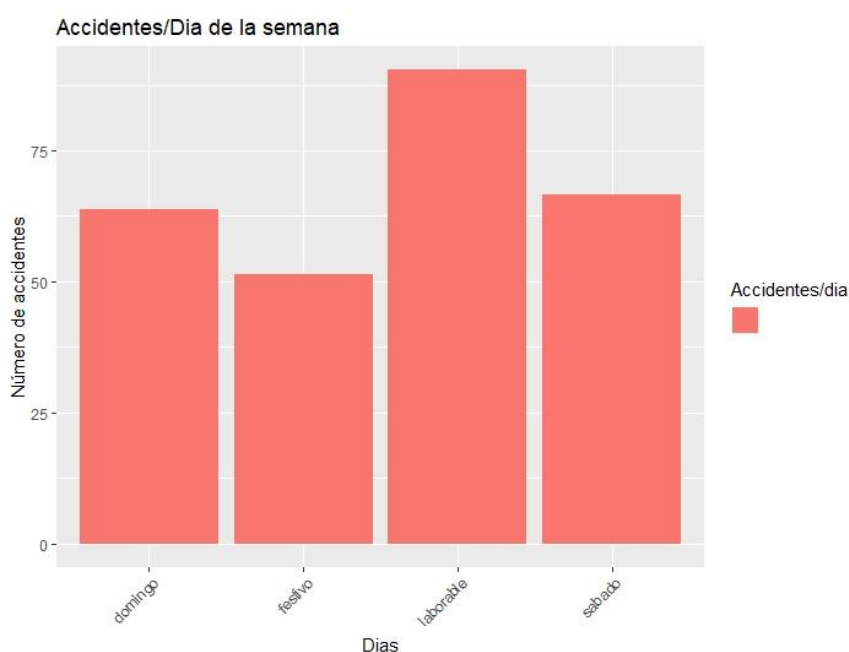


Gráfica 3: Accidentes por día (laborable, sábado, domingo o festivo)

Esta gráfica nos ayuda a saber si que un día sea laborable, festivo o finde de semana influye en la cantidad de accidentes.

Podemos ver que los días en los que hay más accidentes son los días laborables. Esto se debe a que son los días en los que circulan más coches. En los días festivos, sábados y domingos el número de accidentes es prácticamente el mismo.

Conviene resaltar que, aunque es más frecuente que se produzcan accidentes los días laborables, la cantidad de accidentes los fines de semana y festivos es también muy alta.



Gráfica 4: Número de accidentes por franja horaria y tipo de accidente

Esta grafica se hizo para observar de un mismo vistazo la cantidad de accidentes según la franja horaria y la distribución de los accidentes según su tipo.

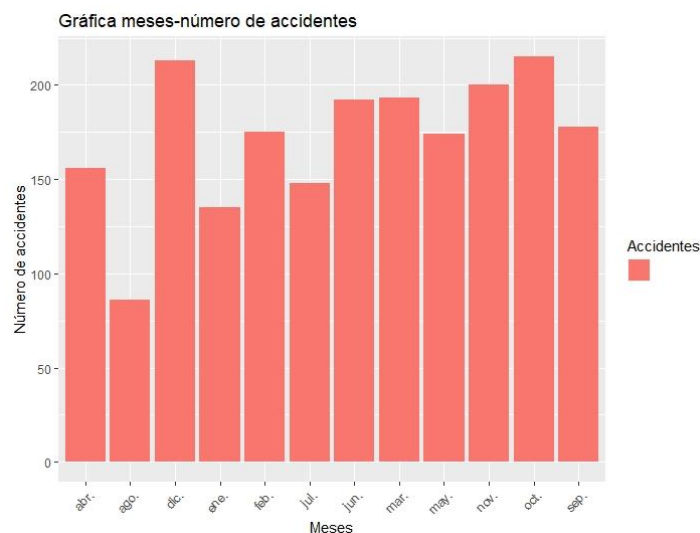
Podemos ver que el número de accidentes según la hora es prácticamente el mismo, excepto en la madrugada cuando el número de coches circulando es menor. Para intentar obtener unos datos más significativos, pensamos en escalarlos al número de vehículos en circulación según la hora, pero no disponíamos de ese dato. Aun así, se puede deducir de manera intuitiva que el número de accidentes en la madrugada es alto en comparación con la cantidad de coches que circulan (hay 1/3 de accidentes respecto al mediodía, en cambio, estimamos que la proporción de vehículos circulando es en torno a 1/5).

Gráfica 6: Accidentes por mes

Esta grafica sirve para darnos una vista general de la distribución de accidentes en el año.

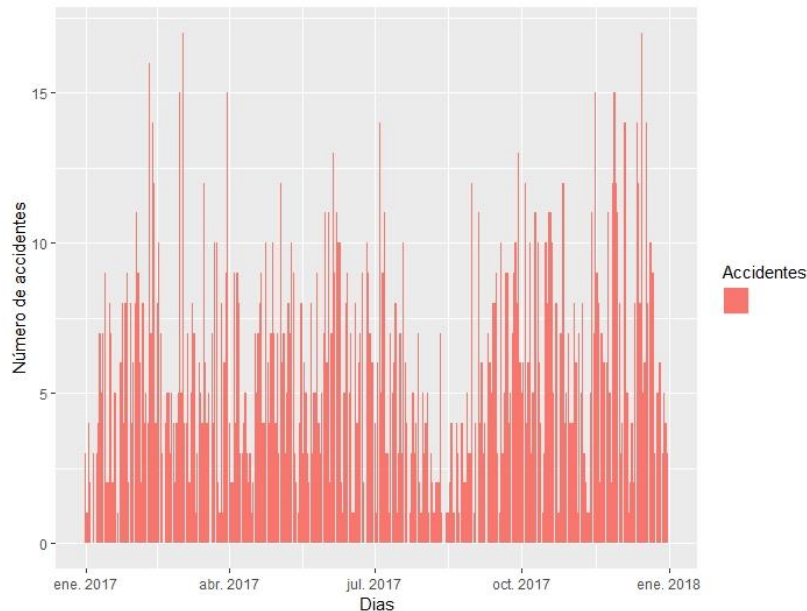
Vemos que la cantidad de accidentes es muy similar todos los meses salvo en diciembre, noviembre y octubre donde la cantidad supera los 200. Esto puede ser porque al bajar las temperaturas la gente tiende a utilizar más el transporte. Además, durante esta época se celebran festividades como Navidad y Año nuevo en las que aumenta la afluencia en la ciudad y hay más distracciones.

Por otro lado, observamos que en agosto la cantidad de accidentes baja drásticamente. Esto puede darse debido a que es periodo de vacaciones y, por lo tanto, los residentes, quienes utilizan más el transporte privado, dejan la ciudad dando paso a los turistas, que se aprovechan más del transporte público y van andando por la ciudad.



Gráfica 7: Accidentes por día

Al igual que la anterior gráfica, esta nos sirve para tener una visión general del año, esta vez día a día. Podemos ver que la media de accidentes al día está en torno a 5. También tenemos valores muy atípicos, con días en los que hay hasta 15 accidentes.



Minería de Datos

Para la implementación de las técnicas de minería de datos, hemos utilizado R, ya que posee una gran cantidad de algoritmos. Tiene facilidad en la manipulación de matrices y vectores (útil en el manejo de bases de datos). A la hora de realizar el análisis estadístico (exploración) es muy preciso y exacto analizando datos. Además, es el lenguaje visto en clase, por lo que era más cómodo en el momento de resolver dudas con el profesor.

Librerías

Los paquetes y las librerías que hemos utilizado han sido para el uso de funciones y gráficas que se requerían en las técnicas de minería de datos que hemos implementado.

```
#####Paquetes#####
#Reglas de asociación
if (!require(arules)) {
  install.packages("arules",
    dependencies = TRUE)}
if (!require(arulesViz)){
  install.packages("arulesViz",dependencies = TRUE)}
if (!require(ggplot2)) {
  install.packages("ggplot2",
    dependencies = TRUE)}
if (!require(RColorBrewer)){
  install.packages("RColorBrewer",dependencies = TRUE)}
if (!require(tidyverse)) {
  install.packages("tidyverse",
    dependencies = TRUE)}

#####Librerías#####
#Clustering y gráficas
library("FactoMineR")
library("factoextra")
library("cluster")
library("dplyr")
library("fpc")
library("ggplot2")
library("RColorBrewer")
library("tidyverse")
#Reglas de asociación
library("arules")
library("arulesviz")
```

Clustering

Para la implementación de esta técnica hemos utilizado el algoritmo k-means. Este algoritmo trabaja iterativamente, es decir, asigna cada dato a un grupo o clúster basándose en la distancia de cada dato al centro del grupo.

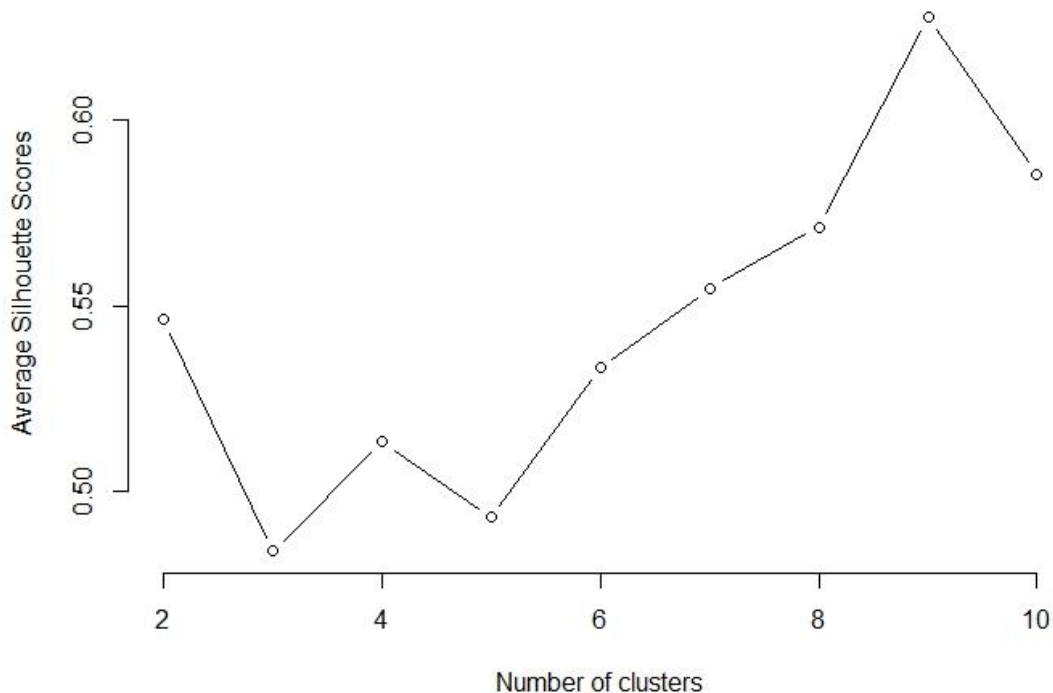
La razón por la cual hemos elegido dicho algoritmo se debe a que es uno de los algoritmos más simples y de los más utilizados en datos numéricos.

Los datos que hemos seleccionado en la implementación de esta técnica son los **días de la semana** y el número de **heridos leves, graves y muertos**. Queremos observar la cantidad de heridos (leves, graves y muertos) que hay según los días de la semana.

Antes de aplicar el algoritmo k-means hay que dar un valor a K (número de clusters que queremos). Para ello, hemos utilizado el método de *silhouette* que nos indica un K óptimo para nuestros datos.

```
silhouette_score <- function(k){  
  km <- kmeans(clustering1, centers = k, nstart=25)  
  ss <- silhouette(km$cluster, dist(clustering1))  
  mean(ss[, 3])  
}  
k <- 2:10  
avg_sil <- sapply(k, silhouette_score)  
plot(k, type='b', avg_sil, xlab='Number of clusters', ylab='Average Silhouette Scores', frame=FALSE)
```

En la siguiente gráfica vemos que la puntuación media es máxima para 9 clusters. Por lo que el valor de K es 9



Teniendo el valor de K, aplicamos la función de **k-means** a los data-frame creados con los heridos leves, graves y muertos.

```
clustering1<-data.frame(datos_accidentes$Dia, datos_accidentes$HL)
kmeans.result1 <- kmeans(clustering1, 9)
```

Para las gráficas hemos hecho uso de la librería *ggplot* aplicando diferentes parámetros para mejorar la comprensión de cada una de ellas:

```
cuenta<-ggplot(d, aes(x=d$Var1,y=d$Freq,fill=""))+ geom_bar(stat="identity")+
  xlab("Cluster")+ylab("Número de elementos")+scale_fill_discrete(name = "Apariciones")
```

Evaluaremos los datos obtenidos en *Resultados y evaluación*.

Reglas de asociación

Antes de la aplicación de las reglas, se descartaron del data frame a utilizar los datos que hicieran referencia a los testigos, ya que estos arrojarían resultados inútiles sobre la lesividad y accidentalidad, que es el objetivo a estudiar.

Para la implementación de esta técnica de minería hemos empleado los conceptos estudiados en clase, utilizando la función a-priori.

En un comienzo se aplica la regla de a-priori a todo el data frame elegido, filtrando por confianza en busca de alguna regla importante.

```
# Cálculo de las reglas de asociación.
rules <- apriori(trans,
  parameter = list(minlen=2, supp=0.05, conf=0.8))
quality(rules) <- round(quality(rules), digits=3)
rules

#Ordenamiento y visualización de las reglas creadas.
rules.sorted <- sort(rules, by="confidence")
inspect(rules.sorted[1:20])
```

Tras esto, se hace otra iteración del método eliminando la redundancia en las reglas creadas, para así aumentar el rango de exploración y no caer siempre en las mismas normas.

```
#Supresión de las reglas redundantes
subset.matrix <- is.subset(rules.sorted, rules.sorted, sparse=FALSE)
subset.matrix[lower.tri(subset.matrix, diag=T)] <- NA
redundant <- colSums(subset.matrix, na.rm=T) >= 1

#Supresión efectiva
rules.pruned <- rules.sorted[!redundant]
#Ordenamos por confianza
rules.sorted <- sort(rules.pruned, by="confidence")
rules_reducidas<-inspect(rules.sorted[1:50])
```

Después de trabajar sobre todo el conjunto del data frame, se buscan reglas más concretas, especificando la parte derecha e izquierda de la norma y con soportes y confianzas más pequeñas. Algunos ejemplos son los siguientes:

1. Causas directas de la lesividad en accidentes

```
#Se buscan causas directas de la lesividad en los accidentes
rules1 <- apriori(trans,
  parameter = list(minlen=2, supp=0.01, conf=0.8),
  appearance = list(rhs=c("lesividad=HL", "lesividad=HG", "lesividad=MT")))
quality(rules1) <- round(quality(rules1), digits=3)
rules1.sorted <- sort(rules1, by="confidence")
inspect(rules1.sorted)
```

2. Factores de la mortalidad en accidentes

```
#Causas directas de mortalidad
rules2 <- apriori(trans,
  parameter = list(minlen=2, supp=0.00005, conf=0.3),
  appearance = list(rhs=c("lesividad=MT")))
quality(rules2) <- round(quality(rules2), digits=3)
rules2.sorted <- sort(rules2, by="confidence")
inspect(rules2.sorted)
```

3. Lesividad en turismos teniendo en cuenta la hora y el estado de la carretera

```
#Lesividad en turismo dependiendo de la hora y el estado de la carretera
rules1 <- apriori(trans,
  parameter = list(minlen=2, supp=0.00005, conf=0.5),
  appearance = list(rhs=c("lesividad=HL", "lesividad=HG", "lesividad=MT"), lhs=c("tipo_vehículo=TURISMO", "hora=MAÑANA", "hora=TARDE", "hora=NOCHE", "hora=MADRUGADA", "mojada=SI", "mojada=NO"), default="none")))
quality(rules1) <- round(quality(rules1), digits=3)
rules1.sorted <- sort(rules1, by="confidence")
inspect(rules1.sorted)
```

4. Influencia de la hora en los atropellos

```
#Patrones de atropello segun la hora
rules1 <- apriori(trans,
  parameter = list(minlen=2, supp=0.005, conf=0.1),
  appearance = list(rhs=c("hora=MAÑANA", "hora=MEDIODIA", "hora=TARDE", "hora=NOCHE", "hora=MADRUGADA"), lhs=c("tipo_accidente=ATROPELLO"), default="none")))
quality(rules1) <- round(quality(rules1), digits=3)
rules1.sorted <- sort(rules1, by="confidence")
inspect(rules1.sorted)
```

En la parte de Resultados se explicará con detalle cuales fueron los resultados de las reglas y las conclusiones que se pueden sacar de ellas.

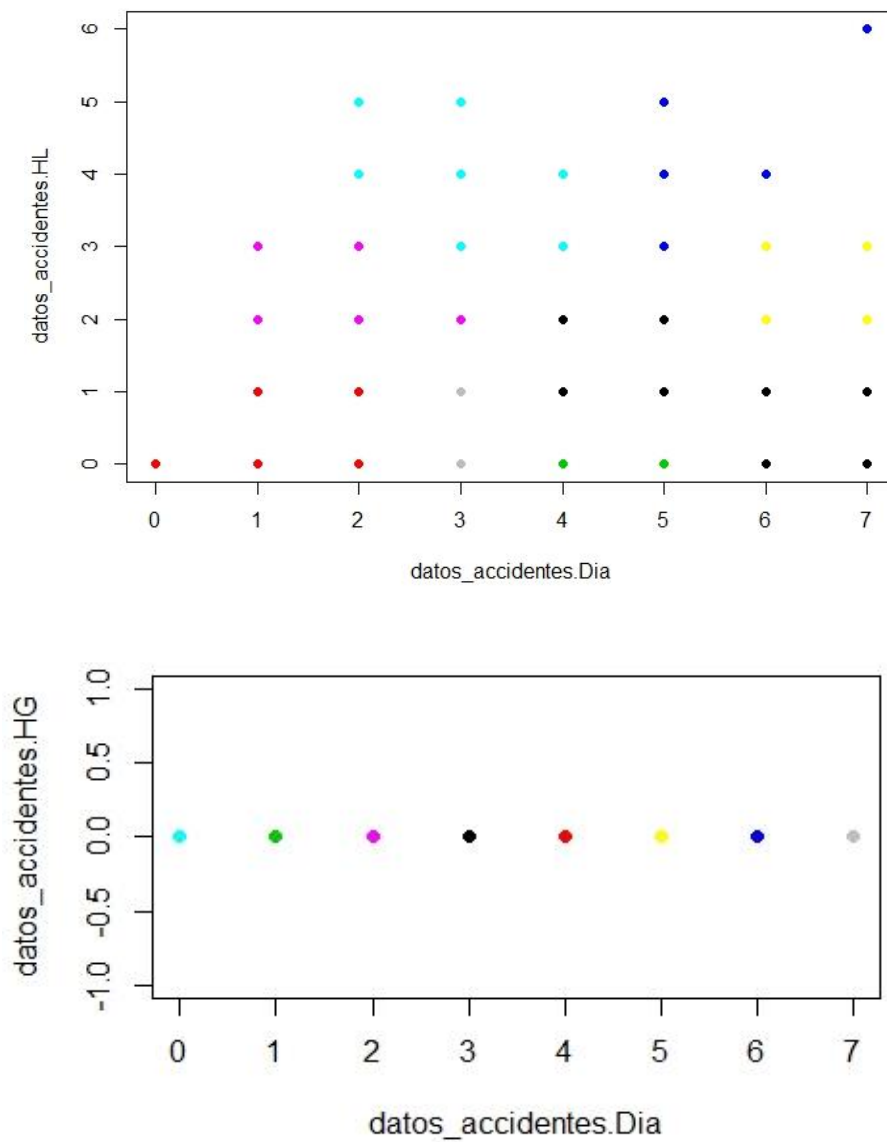
Resultados y Evaluación

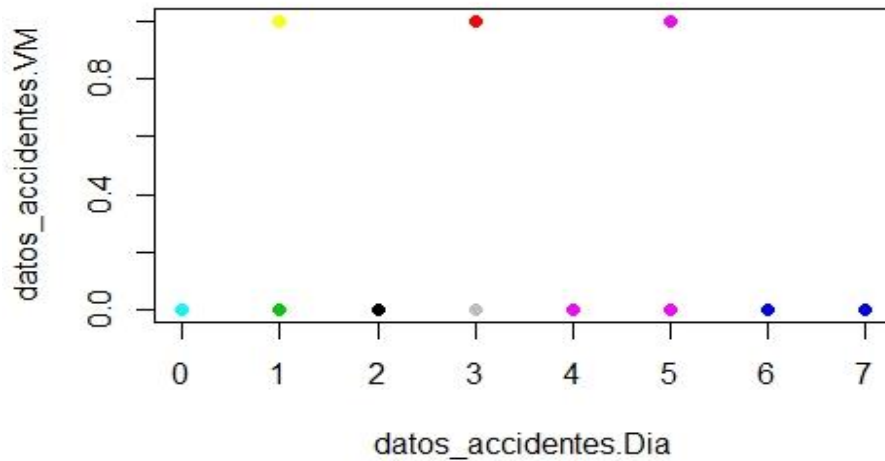
Resultado clustering

Como hemos visto en el apartado de *Exploración*, la única variable que puede darnos información útil es heridos leves, es por eso por lo que hemos trabajado a partir de ella.

Para confirmar lo analizado en la etapa de Exploración, aplicamos el algoritmo k-means en todas las variables.

En estas gráficas vemos el número de heridos en cada clúster con cada uno de los datos que habíamos seleccionado Dia-Heridos leves, Dia-Heridos graves, Dia-Muertes.

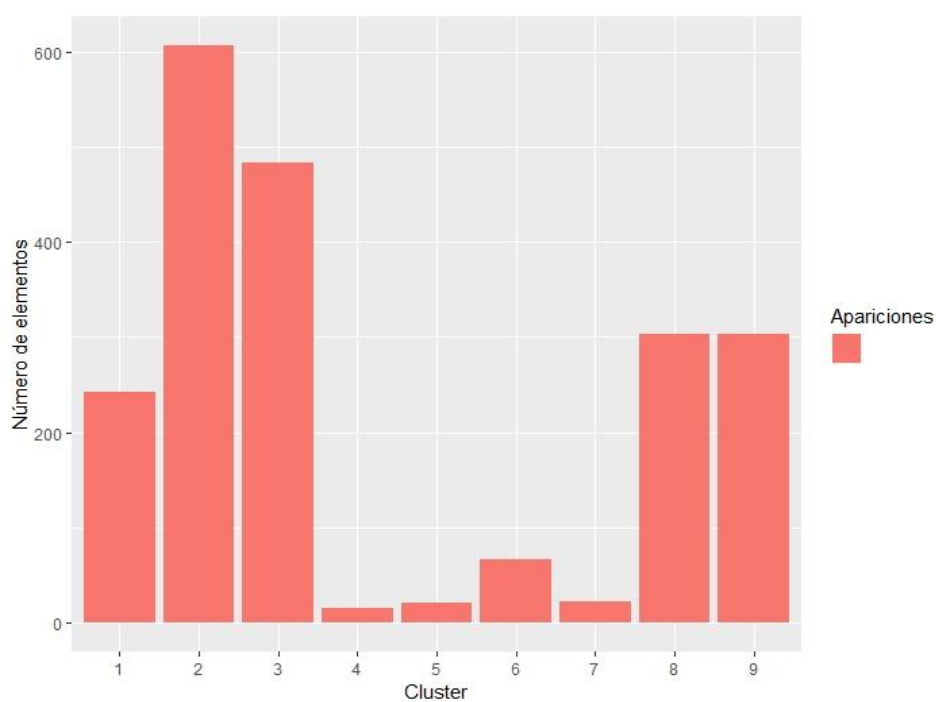




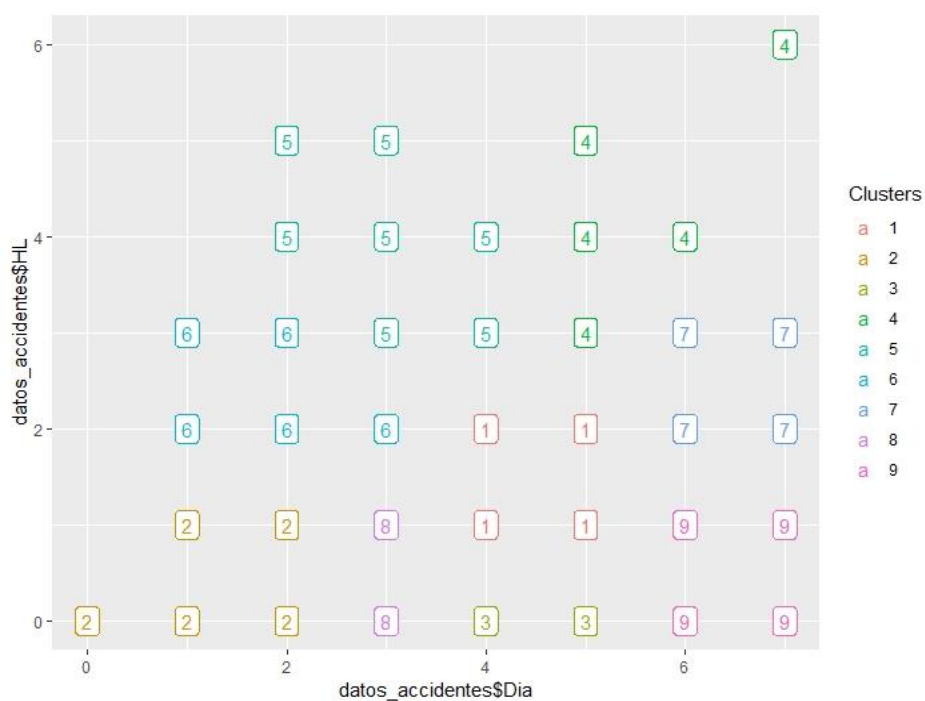
Clustering con días de la semana y heridos leves

- Primero, vamos a fijarnos en el grupo aparentemente menos importante ya que el número de accidentes es menor en el clúster 4 y 5. Esto se debe a que es muy improbable que los días martes, miércoles y jueves (**clúster 5**) y viernes, sábado y domingo (**clúster 9**) haya accidentes con múltiples heridos leves (entre 3 y 6).
- Ahora nos centramos en los clústeres 2 y 3 que son en los que hay mayor número de accidentes. En este caso, al contrario que los clústeres anteriormente explicados, es muy probable que los días lunes y martes (**clúster 2**) y los viernes y jueves (**clúster 3**) haya accidentes con entre 0 y 1 heridos leves.
- Por último, podemos fijarnos en los clústeres 8 y 9. Estos dos clústeres cuentan con el mismo número de miembros, sin embargo, el **clúster 8** pertenece a un solo día, el miércoles, y el **clúster 9** corresponde a dos días, sábado y domingo. Lo que quiere decir que se concentran más accidentes de ese tipo el miércoles.

En esta gráfica podemos observar el número de accidentes en cada clúster.



Gráfica días-heridos leves con la distribución de cada clúster.



Con este análisis podemos asumir que en la mayoría de los accidentes hay pocos heridos y que el día de la semana no es un factor clave en el número de accidentes en los que hay heridos leves.

Resultado reglas de asociación

Siguiendo el orden en el que aplicamos las reglas procedemos a revisar los resultados.

En primer lugar, al aplicar la función a priori a todo el data frame y ordenarlo por confianza los resultados conseguidos no son muy prometedores.

	lhs	rhs	support	confidence	lift	count
[1]	{lesividad=HL,tipo_vehículo=NO ASIGNADO,tipo_accidente=ATROPELLO}	=> {tipo_persona=PEATON}	0.050	1.000	15.469	1306
[2]	{tipo_persona=PEATON}	=> {tipo_vehículo=NO ASIGNADO}	0.065	0.999	15.323	1685
[3]	{tipo_persona=PEATON,tipo_accidente=ATROPELLO}	=> {tipo_vehículo=NO ASIGNADO}	0.062	0.999	15.323	1620
[4]	{tipo_vehículo=NO ASIGNADO,tipo_accidente=ATROPELLO}	=> {tipo_persona=PEATON}	0.062	0.999	15.459	1620
[5]	{lesividad=HL,tipo_persona=PEATON}	=> {tipo_vehículo=NO ASIGNADO}	0.052	0.999	15.330	1349
[6]	{mojada=NO,tipo_persona=PEATON}	=> {tipo_vehículo=NO ASIGNADO}	0.059	0.999	15.322	1547
[7]	{lesividad=HL,tipo_persona=PEATON,tipo_accidente=ATROPELLO}	=> {tipo_vehículo=NO ASIGNADO}	0.050	0.999	15.330	1306

Las reglas que obtenemos tienen una gran confianza, pero en cambio ofrecen información inútil.

Al aplicar el filtro por redundancia, los resultados obtenidos son ligeramente mejores, pero siguen sin ofrecer resultados útiles.

	lhs	rhs	support	confidence	lift	count
[1]	{lesividad=HL,tipo_vehículo=NO ASIGNADO,tipo_accidente=ATROPELLO}	=> {tipo_persona=PEATON}	0.050	1.000	15.469	1306
[2]	{tipo_persona=PEATON}	=> {tipo_vehículo=NO ASIGNADO}	0.065	0.999	15.323	1685
[3]	{lesividad=HL,tipo_persona=PEATON}	=> {tipo_accidente=ATROPELLO}	0.050	0.968	7.403	1307
[4]	{lesividad=HL,tipo_vehículo=NO ASIGNADO}	=> {tipo_accidente=ATROPELLO}	0.050	0.963	7.364	1306

A partir de este momento, comenzamos con los resultados que nos arrojaron los estudios más exhaustivos y específicos que realizamos.

A la hora de crear las normas nos centramos en las variables que nos mostraban mayor importancia durante la etapa de la exploración. Estas serían por ejemplo los accidentes de motocicleta, los accidentes de turismo (imposible determinar reglas útiles debido a la variedad de los datos), el estado de la carretera, los atropellos y las colisiones dobles.

1. Causas directas de la lesividad en accidentes

Intentamos encontrar causas directas de la lesividad en los accidentes, y encontramos una regla que nos ofrecía información interesante.

```
[7] {tipo_vehículo=MOTOCICLETA} => {lesividad=HL} 0.139 0.832 1.788 3639
```

Los accidentes con moto casi siempre suponen heridos leves. Esto nos hizo darnos cuenta de que era importante tratar de reducir este tipo de accidentes y que era necesario buscar las causas de estos.

2. Tipo de accidente-motocicleta

Procedemos a investigar mas a fondo en que tipo de accidentes se ven implicados las motocicletas. Debíamos de conocer si las motos tenían accidentes por errores suyos o problemas en la calzada (caídas de motocicleta) o si se debía a otros factores diferentes.

	lhs	rhs	support	confidence	lift	count
[1]	{tipo_vehiculo=MOTOCICLETA}	=> {tipo_accidente=COLISIÓN DOBLE}	0.109	0.651	1.085	2848
[2]	{tipo_vehiculo=MOTOCICLETA}	=> {tipo_accidente=CAÍDA MOTOCICLETA}	0.044	0.260	4.594	1138
[3]	{tipo_vehiculo=MOTOCICLETA}	=> {tipo_accidente=COLISIÓN MÚLTIPLE}	0.006	0.035	0.284	154
[4]	{tipo_vehiculo=MOTOCICLETA}	=> {tipo_accidente=ATROPELLO}	0.005	0.032	0.243	139

Como podemos ver la mayoría de los accidentes de motocicleta son colisiones dobles (muy probablemente entre coche y motocicleta). Esto nos demuestra que la principal causa de los accidentes de motocicleta es un factor externo, los turismos.

3. Estado de la carretera-motocicleta

Examinamos si el estado de la carretera afecta en los accidentes de moto.

	lhs	rhs	support	confidence	lift	count
[1]	{tipo_vehiculo=MOTOCICLETA}	=> {mojada=NO}	0.150	0.894	0.980	3910
[2]	{tipo_vehiculo=MOTOCICLETA}	=> {mojada=SI}	0.018	0.106	1.213	464

Aunque hay que tener en cuenta que hay muchos menos días en los que la carretera esta mojada comparados con los que no lo está, queda claro que el estado de la carretera no es un factor diferencial en Madrid.

4. Estado de la carretera-atropello

A la hora de encontrar las causas de los atropellos pensamos que, si la carretera esta mojada la distancia de frenada aumenta, siendo el estado de la carretera un factor importante.

	lhs	rhs	support	confidence	lift	count
[1]	{tipo_accidente=ATROPELLO}	=> {mojada=NO}	0.12	0.917	1.005	3131

Nuevamente, volvemos a ver que el estado de la carretera no es un factor diferencial. Podemos deducir por tanto que los atropellos se deben a errores del peatón o el conductor.

Conclusión

Finalmente, no podemos sacar un gran número de conclusiones después del estudio, debido a la variedad y complejidad tanto del tema como de los datos. Aun así, si conseguimos buenas conclusiones que pueden ayudar a mejorar la seguridad en la carretera y así evitar accidentes.

A continuación, procedemos a enumerar las principales conclusiones que sacamos del estudio, y damos algunas ideas de como enfocar el problema.

1. La motocicleta es un transporte vulnerable en la carretera, ya que no tiene la robustez de otro tipo de transportes. La mayoría de los accidentados salen heridos y por ello es importante tomar medidas. Como hemos visto los accidentes más comunes en los que se ven involucradas son los choques con otros vehículos en los que la persona peor parada siempre es el motociclista. De este modo, debemos proporcionar una alternativa en la que los motociclistas estén realmente seguros y no dependan de factores externos.

Nosotros proponemos vías de utilización única para motocicletas. De esta manera se reduciría en sobremanera los choques con los demás vehículos.

2. Los atropellos en las zonas urbanas son un problema presente en la capital. Ya que el peatón es el individuo más vulnerable en la vía urbana, hay que abordar este problema de inmediato. Tras comprobar que el estado de la carretera no es la causa del gran número de atropellos, llegamos a la conclusión que este es un problema bilateral peatón-conductor. Estos accidentes se producen por imprudencias y despistes en ambos, por lo tanto, proponemos aumentar la señalización y seguridad tanto en pasos de cebra como en semáforos. Otra medida, que ya se está llevando a cabo y puede haber mejorado el problema en estos últimos años, es la limitación de velocidad a 30 km/h en las zonas urbanas.

3. La mayor parte de los accidentes suceden a mitad de semana, cuando la gente está acabando la semana de trabajo. Podemos deducir que esto se debe, además de a la gran afluencia de vehículos, al cansancio de los conductores. El cansancio lleva a despistes o a cometer errores en la conducción. Como evitar el cansancio es algo difícil, lo más lógico es promover el uso de otras alternativas de transporte durante la jornada laboral como el transporte público.

En cuanto a conclusiones sobre las técnicas de minería de datos, nos hemos dado cuenta de que hay muchos factores que influyen en la tarea. Las fuentes, la limpieza, la cantidad de datos son una parte muy importante a la hora de realizar un estudio. Por ejemplo, para hacer un estudio realmente útil y certero en nuestro ámbito haría falta la recolección de muchas más bases de datos, obteniendo así todos los flancos del problema.

También la correcta aplicación de las técnicas no es una tarea fácil, ya que hay que tener claro sobre que partes de los datos aplicar cada una. Una buena exploración previa y el conocimiento de las funciones y técnicas es crucial para no emplear tiempo “minando donde no hay para picar”.

Nos ha dado la sensación de que, aunque teníamos herramientas para conseguir algunos resultados, nuestro conocimiento sobre la minería de datos era demasiado poco para las tareas que teníamos en mente.

De todas formas, con este trabajo hemos asentado las principales bases sobre la minería de datos. Quedándonos claras sus partes y lo importante que es en el mundo actual.

Bibliografía

<https://mnrva.io/kdd-platform.html>

<http://fcojlanda.me/es/ciencia-de-los-datos/kdd-y-mineria-de-datos-espanol/>

<https://techlandia.com/mejores-10-algoritmos-mineria-datos-info-295108/>

<https://datos.madrid.es/portal/site/egob/menuitem.9e1e2f6404558187cf35cf3584f1a5a0/?vgnextoid=374512b9ace9f310VgnVCM100000171f5a0aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD&vgnextfmt=default>

Apuntes del Campus Virtual de Minería de Datos.

<https://www.rdocumentation.org/>