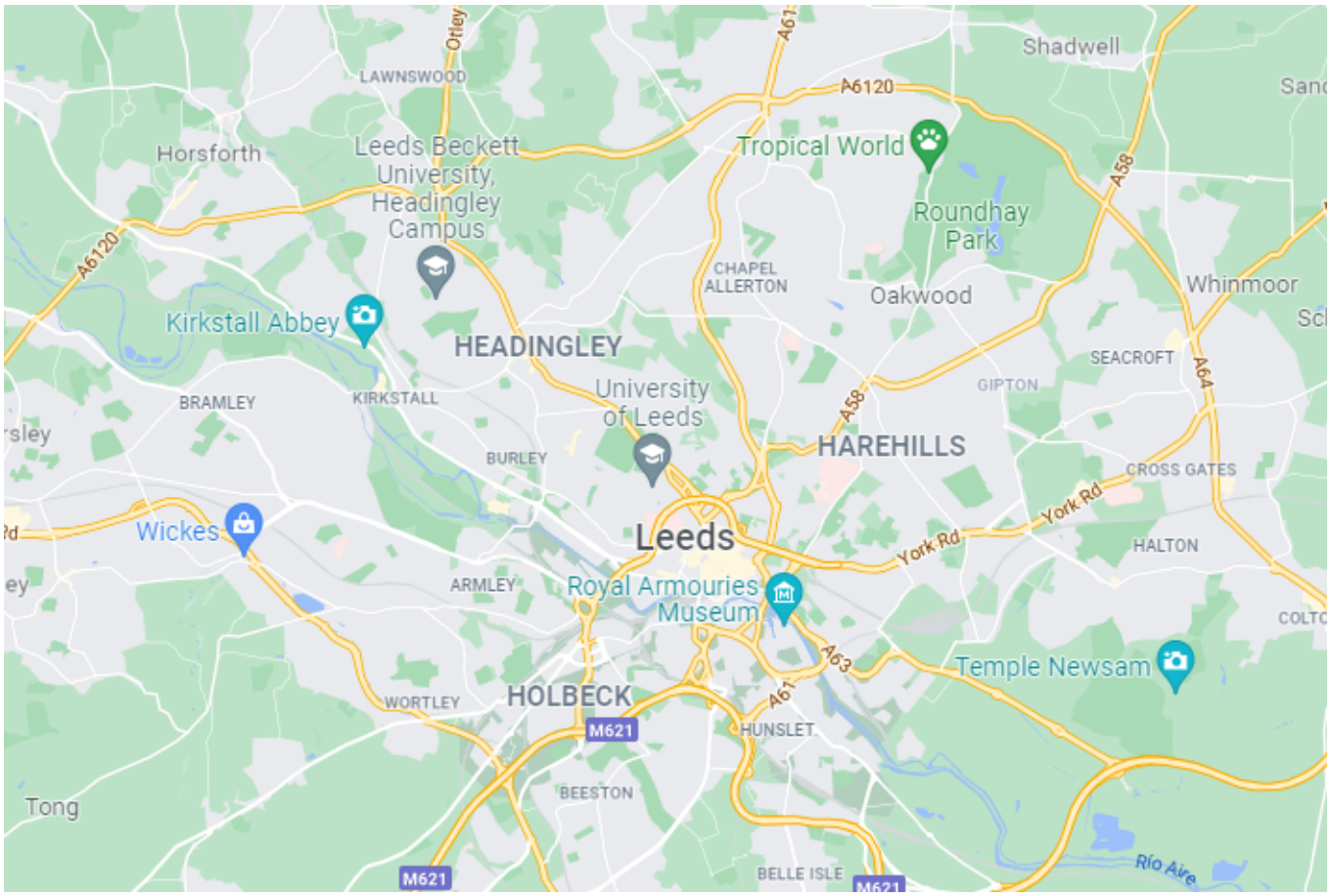
Trabajo Práctico Final

Minería de Datos

Alumno: Navall, Nicolás Uriel. N-1159/2.

Introducción

Para el trabajo presentado se utilizó un dataset de accidentes de tránsito de la ciudad de Leeds en Yorkshire, Reino Unido. Se eligió dicho dataset porque contiene muchas variables interesantes para analizar que no estaban presentes en otros datasets similares.



El dataset contiene muchas variables discretizadas de forma previa, para ver el significado de estas y sus valores leer “Guidance” del siguiente link: https://data.europa.eu/data/datasets/road-traffic-accidents/. En este link además podemos encontrar los datos usados (en particular utilizaremos los datos del 2019).

Pretratamiento de datos

Lo primero que se hizo fue convertir la variable de fecha en el formato KSP para conservar las distancias entre ellas, y restar sus resultados por 2020 para facilitar su manejo.

Además, genero una variable que transforma la fecha en día de la semana porque al no ser una transformación lineal (ni logarítmica) puede que sea difícil extraer esta información de ser necesaria.

Mantengo el horario en una variable separada en lugar de incluirla dentro del cálculo KSP de la fecha por que interpreto que no son variables relacionadas en el contexto dado (es decir creo que tener la variable “hora” separada es más útil que mezclarla con la fecha en otra variable).

Sacamos la variable Local.Authority por que no varía en todos los datos, por lo que no agregará información.

Elimino las entradas que tienen un valor de 9 en la variable “Weather Conditions”, que representa que no se conoce el clima en el momento del accidente, ya que solo se corresponden a dos entradas.

Se convierte cada valor de la variable “Type.of.Vehicle” en una variable diferente (con excepción de algunos casos que se tenían muy pocas muestras con dicho valor) ya que en la mayoría de los valores no había

ningún sentido de orden. Algunos valores como los asociados a las motocicletas o a los camiones de carga se juntan en una sola variable cada uno, ya que se utiliza el orden de sus valores para reflejar la velocidad máxima de la motocicleta y el peso del camión respectivamente.

Saco la variable “1st Road Class & No” por que tiene tantos valores discretos que no seria practico convertirla en múltiples variables. De igual manera se tiene información de la localización de los accidentes en las variables de coordenadas cartesianas.

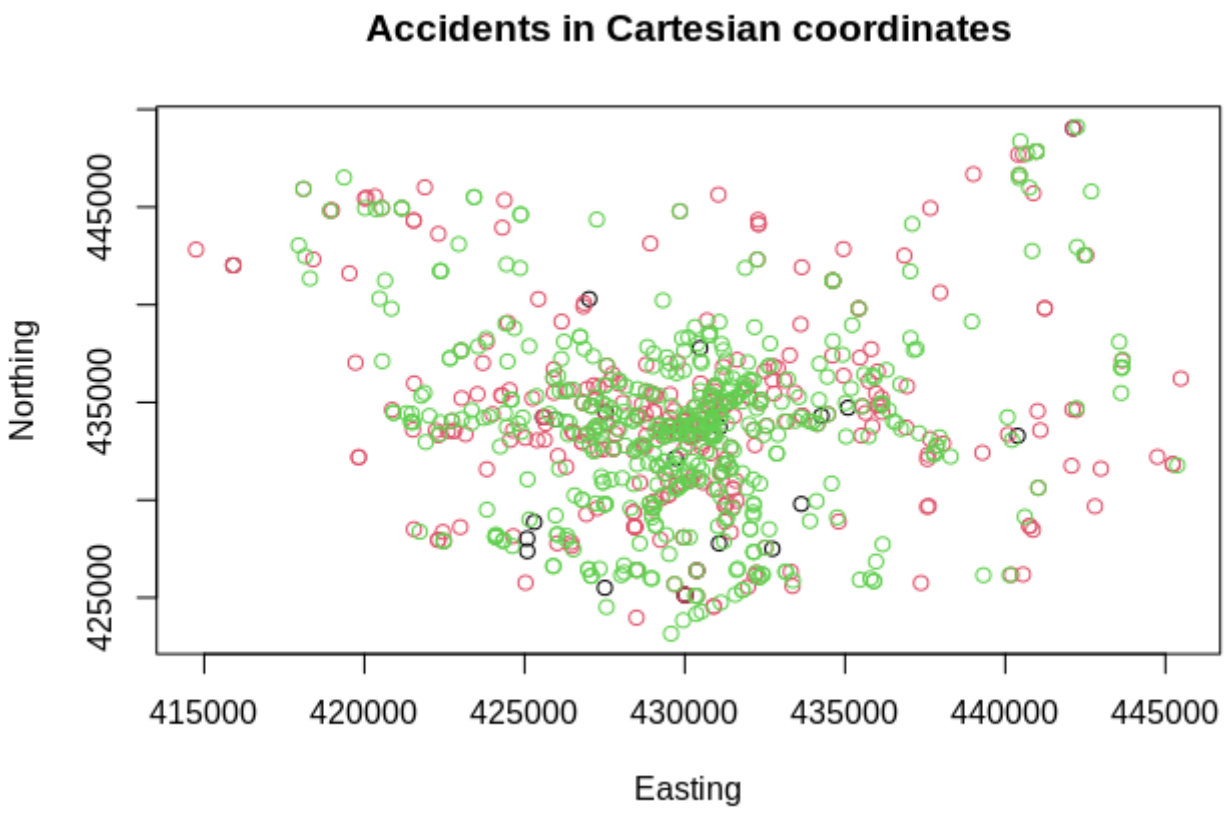
Decido sacar la variable "Number.of.Vehicles" porque está demasiado asociada con la gravedad del accidente y la considero un falso predictor. Además no tendría sentido para un predictor tener información de antemano sobre la cantidad de autos involucrados en un “posible” accidente. Elimino también la variable “Vehicle.Number” por las mismas razones.

Eligo ignorar la variable de “Reference.Number” por que, aunque la información que busca reflejar puede ser útil (edad de los conductores del accidente por ejemplo), no tengo forma de expresar esto si no es reflejando cada persona que pertenece a un mismo accidente en una lista o algo por el estilo. Esto también implica que accidentes con muchas personas tendrán más peso en el dataset procesado.

Aunque “1st Road Class” puede ser de utilidad, lamentablemente tiene la mitad de los datos sin clasificar, por lo que se opto por ignorarla.

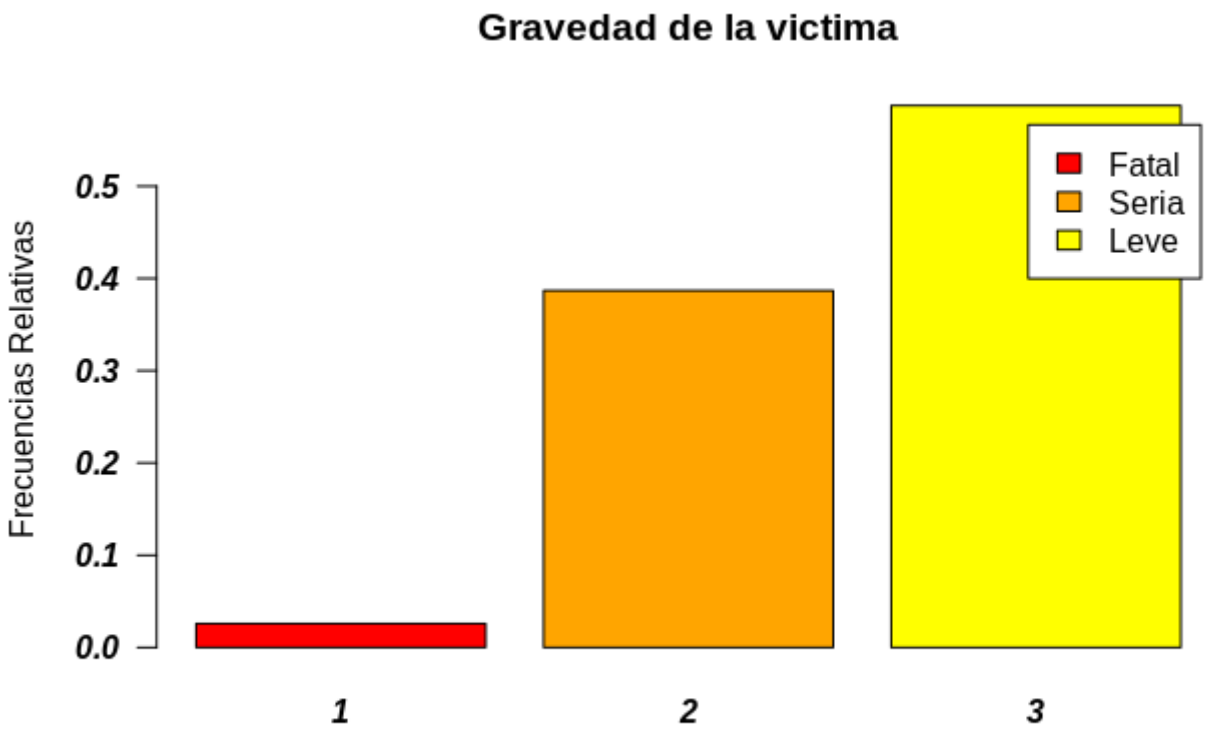
El dataset está desbalanceado, por lo cual fue necesario realizar un subsampleo de la clase mayoritaria.

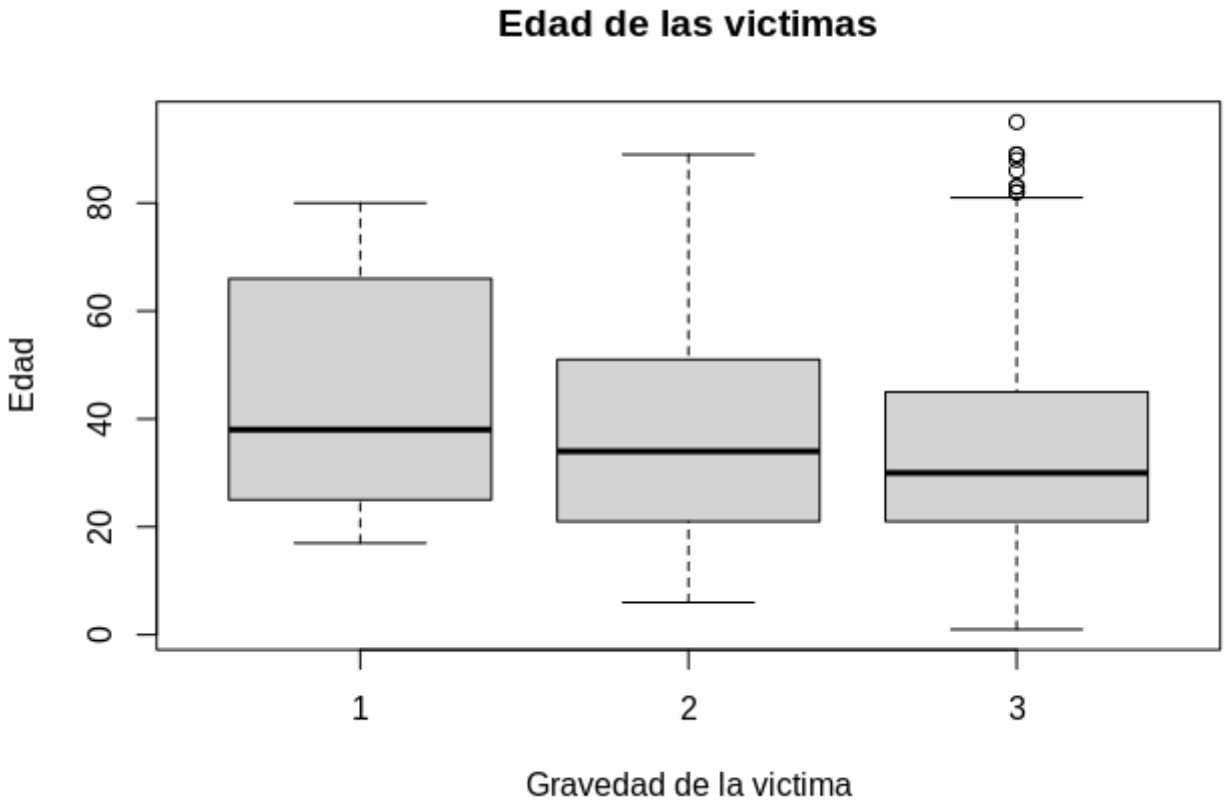
Visualización de los datos

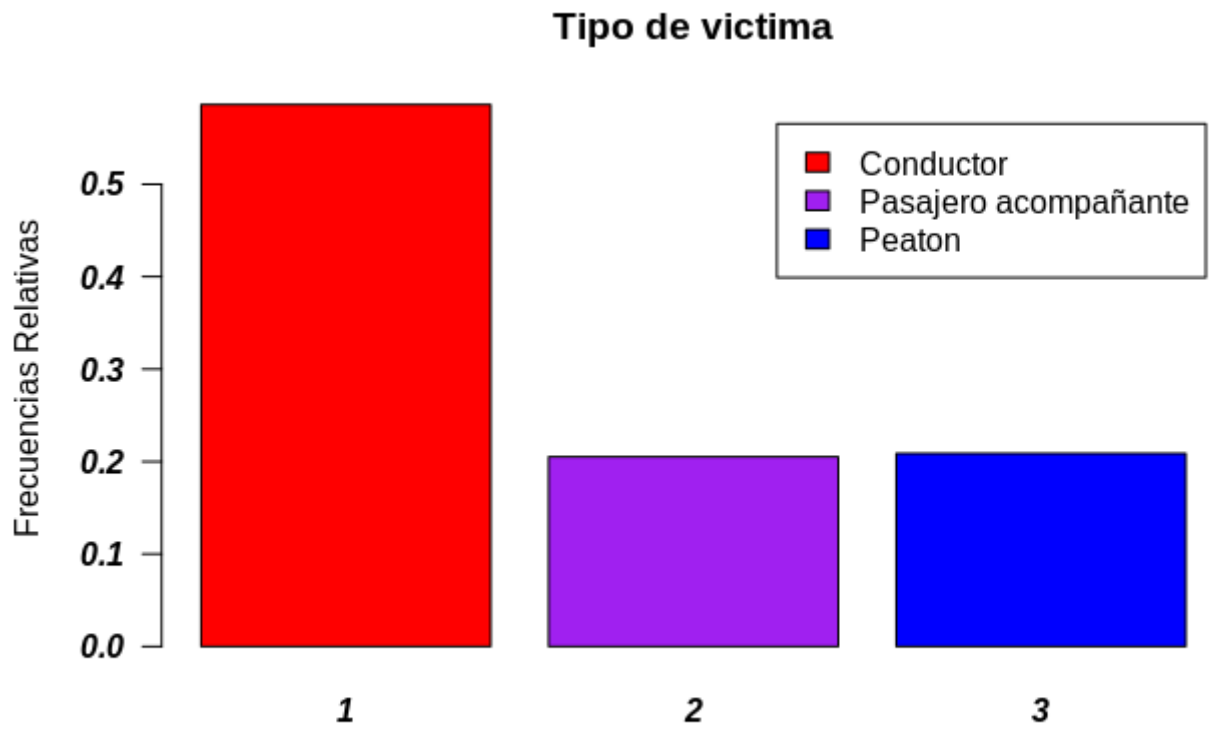


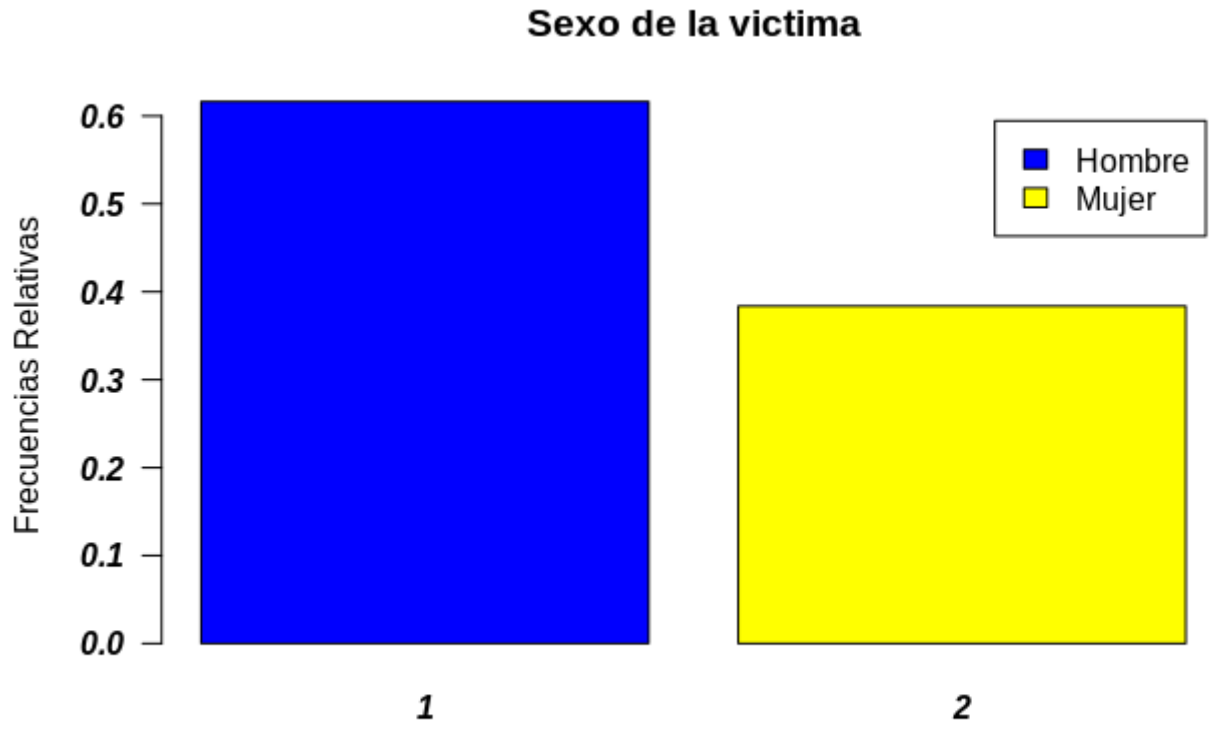
En este gráfico podemos observar la locación de los accidentes. El color indica la gravedad del estado de la víctima, siendo verde usado para “leve”, rojo para “serio” y negro para “fatal”.

Podemos distinguir algunos caminos en este gráfico y lo que interpretamos es el centro de la ciudad.


Gravedad de la victima
Frecuencias Relativas
0.5
0.4
0.3
0.2
0.1
0.0
1
2
3
Fatal
Seria
Leve


Edad de las victimas
Edad
0
20
40
60
80
1
2
3
Gravedad de la victima


Tipo de victima
Frecuencias Relativas
0.0
0.1
0.2
0.3
0.4
0.5
1
2
3
Conductor
Pasajero acompañante
Peaton


Sexo de la victima
Frecuencias Relativas
0.0
0.1
0.2
0.3
0.4
0.5
0.6
1
2
Hombre
Mujer

Análisis de variables relevantes

Método	Variables ordenadas
Forward Wrapper RF	Vehicle.Motorcycle, Casualty.Class, Lighting.Conditions, Vehicle.Pedal.Cycle, Road.Surface, Age.of.Casualty, Vehicle.BusOrCoach, Vehicle.GoodsVehicle, Vehicle.Car, Sex.of.Casualty, Weather.Conditions, Weekday, Time..24hr., Vehicle.TaxiPrivateHireCar, Grid.Ref..Easting, Accident.Date, Grid.Ref..Northing
Forward Wrapper LDA	Vehicle.Motorcycle, Age.of.Casualty, Casualty.Class, Vehicle.Pedal.Cycle, Sex.of.Casualty, Accident.Date, Vehicle.BusOrCoach, Vehicle.Car, Grid.Ref..Easting, Weekday, Vehicle.TaxiPrivateHireCar, Time..24hr., Vehicle.GoodsVehicle, Grid.Ref..Northing, Weather.Conditions, Lighting.Conditions, Road.Surface
Backward Wrapper RF	Vehicle.Motorcycle, Casualty.Class, Age.of.Casualty, Vehicle.Car, Vehicle.TaxiPrivateHireCar, Time..24hr., Accident.Date, Lighting.Conditions, Weather.Conditions, Weekday, Vehicle.Pedal.Cycle, Vehicle.BusOrCoach, Sex.of.Casualty, Grid.Ref..Northing, Grid.Ref..Easting, Road.Surface, Vehicle.GoodsVehicle
BackwardWrapper LDA	Vehicle.Car, Vehicle.TaxiPrivateHireCar, Casualty.Class, Age.of.Casualty, Vehicle.GoodsVehicle, Vehicle.BusOrCoach, Vehicle.Pedal.Cycle, Time..24hr., Road.Surface, Sex.of.Casualty, Weather.Conditions, Accident.Date, Grid.Ref..Northing, Grid.Ref..Easting, Lighting.Conditions, Vehicle.Motorcycle, Weekday
Filtro Kruskal-Wallis	Vehicle.Motorcycle, Vehicle.Car, Casualty.Class, Sex.of.Casualty, Age.of.Casualty, Vehicle.Pedal.Cycle, Accident.Date, Weekday, Lighting.Conditions, Grid.Ref..Northing, Time..24hr., Vehicle.TaxiPrivateHireCar, Road.Surface, Vehicle.GoodsVehicle, Vehicle.BusOrCoach, Grid.Ref..Easting, Weather.Conditions
RFE RF	Grid.Ref..Northing, Weather.Conditions, Grid.Ref..Easting, Accident.Date, Sex.of.Casualty, Road.Surface, Time..24hr., Weekday, Vehicle.GoodsVehicle, Vehicle.TaxiPrivateHireCar, Vehicle.Car, Vehicle.BusOrCoach, Lighting.Conditions, Age.of.Casualty, Casualty.Class, Vehicle.Pedal.Cycle, Vehicle.Motorcycle
RFE SVM	Grid.Ref..Easting, Vehicle.GoodsVehicle, Grid.Ref..Northing, Accident.Date, Time..24hr., Weekday, Weather.Conditions, Vehicle.TaxiPrivateHireCar, Sex.of.Casualty, Vehicle.BusOrCoach, Road.Surface, Vehicle.Car, Lighting.Conditions, Vehicle.Pedal.Cycle, Age.of.Casualty, Casualty.Class, Vehicle.Motorcycle

La mayoría de los métodos parecen acordar que la gravedad de las víctimas depende en gran parte si el vehículo es una moto (y la velocidad máxima de esta) o no.

Otras variables que parecen estar bastante altas en los ranking de múltiples métodos son Age.of.Casualty y Casualty.Class.

Búsqueda de clusters

Se busco clusters sobre la variable de

Método	Performance
K-means	0.554
h-clust single	0.578
h-clust average	0.582
h-clust complete	0.481

Aunque a simple vista parece que los primeros 3 métodos dan mejores resultados basta con mirar las clasificaciones para entender que en realidad están tomando a la gran mayoría de los datos como un único cluster, por lo que no son resultados muy útiles. No darse cuenta de problemas como estos podría ser peligroso, dado que las clases más críticas, en nuestro caso los accidentes con víctimas fatales, son muy pocos y clasificarlos mal es más peligroso que clasificar víctimas más leves como posibles víctimas fatales.

También se corrió GAP sobre los datos y no se encontraron clusters, lo que, en conjunto con los resultados obtenidos con los métodos vistos antes, me dan a entender que los datos no son tan clusterizables como pensaba en un principio.

Entrenamiento y aplicación de clasificadores

Para estimar el error de los siguientes clasificadores se utilizó cross validation con 10 folds.

Método	Error
Boosting	0.334
randomForest	0.069