PREDICCIÓN DE ACCIDENTES AUTOMOVILÍSTICOS EN REINO UNIDO

POR:
Jorge Antonio Franco Vásquez
Felipe Carlos Martínez Mármol
Juan Camilo Tabares Henao

MATERIA:
Introducción a la Inteligencia Artificial

PROFESOR:
Raúl Ramos Pollan

UNIVERSIDAD DE ANTIOQUIA
FACULTAD DE INGENIERÍA
MEDELLÍN 2022

Contenido

Introducción

En cualquier parte del mundo, muchas personas son víctimas cada año en accidentes de tránsito, ya sea por imprudencia o confusiones. Lo cierto es que los datos estadísticos muestran que se han convertido en una gran problemática para los países del mundo, ya que no sólo deben lidiar con la circulación de tránsito, sino también con la salud de las personas afectadas.

La importancia del buen manejo de los datos puede ser vital en situaciones que impliquen riesgo en la seguridad de las personas como por ejemplo los accidentes automovilísticos, mediante este dataset se buscará predecir la accidentalidad teniendo en cuenta las causas y consecuencias, para poder trabajar desde la prevención y reducir la tasa de accidentalidad, así como proveer información clave a las personas de interés cómo hospitales, los cuales podrían beneficiarse de dicha información para estar listos en caso de prever un alto ingreso de pacientes debido a accidentes automovilísticos.

La inteligencia artificial nos ayudará a cumplir este objetivo, ya que se presenta como una gran herramienta para obtener análisis de los datos que se obtienen de las bases de datos y conseguir los resultados deseados.

Dataset

El dataset que se va a utilizar proviene de Kaggle, el cual está compuesto por un archivo CSV con datos de accidentes de tránsito ocurridos en Reino Unido desde el año 2005 al 2014, recolectados por el gobierno del Reino Unido.

El dataset contiene un registro de más de 1.8 millones de accidentes, sin embargo, para el proyecto solo se usarán los datos más recientes, es decir los correspondientes al año 2014, los cuales son 146.322 accidentes en total, y contiene la siguiente información:

Location_Easting_OSGR	Ubicación Este
Location_Northing_OSGR	Ubicación de Norte
Longitude	Longitud del lugar de accidente
Latitude	Latitud del lugar del accidente
Police_Force	No. de Fuerza Policial
Accident_Severity	Severidad del accidente en una escala de 1 a 5
Number_of_Vehicles	Número de vehículos involucrados en el accidente.
Number_of_Casualties	*Número de víctimas (Variable Objetivo)*
Date	Fecha

Day_of_Week	Día de la semana
Time	Hora
Local_Authority_(District)	Autoridad Local (Distrito)
Local_Authority_(Highway)	Autoridad Local (Carretera)
1st_Road_Class	Tipo de la 1ra carretera
1st_Road_Number	Número de la 1ra carretera
Road_Type	Tipo de carretera
Speed_limit	Límite de velocidad
Junction_Control	Control en la intersección
2nd_Road_Class	Tipo de la 2da carretera
2nd_Road_Number	Número de la 2da carretera
Pedestrian_Crossing-Human_Control	Control humano de peatones
Pedestrian_Crossing-Physical_Facilities	Instalaciones físicas para el cruce de peatones
Light_Conditions	Condición de iluminación el día del accidente
Weather_Conditions	Condiciones meteorológicas el día del accidente
Road_Surface_Conditions	Condiciones de la superficie de la carretera en un punto accidental
Special_Conditions_at_Site	Condiciones especiales en el sitio
Carriageway_Hazards	Peligros de la calzada
Urban_or_Rural_Area	Área urbana o Rural
Did_Police_Officer_Attend_Scene_of_Accident	¿El oficial de policía asistió a la escena del accidente?
LSOA_of_Accident_Location	“Lower Layer Super Output Area” es un sustituto para la locación geográfica de longitud y latitud
Year	Año del evento accidental

Métrica

La principal métrica que se utilizará en el modelo de predicción de accidentes de tránsito es la Raíz del Error Cuadrático Medio o Root Mean Square Error (RMSE) y se calcula de la siguiente manera:

$$RMSE = \sqrt{\frac{1}{N}\sum(y_i - \hat{y}_i)^2}$$

Donde RMSE es la raíz cuadrada del promedio de la suma de diferencias cuadradas entre los valores observados en la serie y los esperados según el modelo de tendencia.

Dónde y_i corresponde a la serie observada e $\hat{y}_i$ a la serie estimada, y N el número de datos totales. En los resultados se busca que mientras menor sea el error, más adecuado es el modelo.

Análisis de Datos

Selección de datos

En esta parte del trabajo establecemos los datos más importantes que indican la problemática que se presenta con respecto a los accidentes en Reino Unido. Primero que todo, leemos nuestro archivo .csv con el nombre de 'UK_Accident.csv', y luego tomamos los datos del 2014 como base para analizar los principales problemas que pueden influir en los accidentes en Reino Unido.

Luego, los datos en String lo convertimos en datos de tiempo en el que los pandas puedan reconocer usando la función To_datetime que devuelve una indicación de fecha y hora a partir de un string.

Por último, calculamos un resumen de las principales estadísticas como el total contado, la desviación estándar, valores máximos y mínimos. Usando la función describe() que permite devolver este resumen de estadísticas de todas las columnas del DataFrame

Variable objetivo

La variable objetivo que se desea ser objeto de predicción en este proyecto es *Number_of_Casualties*, la cual nos permite saber de forma cualitativa la problemática accidental en el Reino Unido en un futuro, y es también la que más facilita las posibilidades de encontrar soluciones. Luego, se analizarán y posteriormente, se decidirá cuales variables serán las entradas para el entrenamiento de los algoritmos.

Análisis de la variable objetivo

Es fundamental describir y analizar el comportamiento de la variable objetivo, en donde se puede apreciar una alta asimetría hacia valores cercanos a 1, debido a que es un valor entero se procede a verificar los valores únicos de esta variable para verificar que no todos los datos sean 1. Se verifica que existen muchos más datos por lo que se puede aplicar una transformación logarítmica para apreciar mejor los datos.

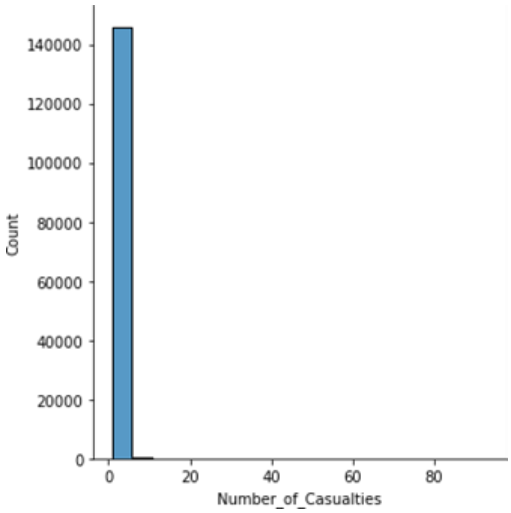


Figura 1: Distribución de la variable objetivo

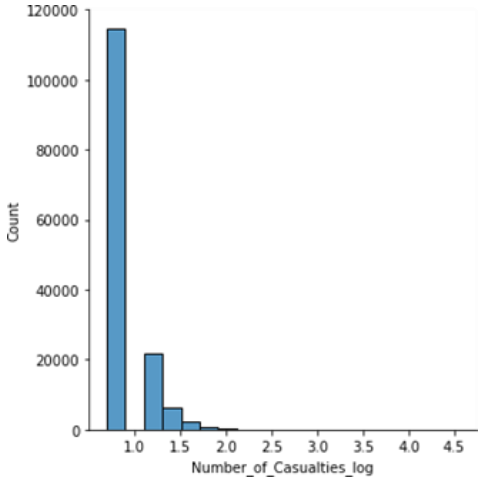


Figura 2: Transformación logarítmica

En la Figura 2, se puede notar que la distribución de la variable objetivo luego de la transformación logarítmica posee tiene un mejor comportamiento justo para procesar un análisis, ya que posee más datos que se despreciaban por el sesgo que existía en ciertos rangos de la gráfica 1. Por lo tanto, esta variable modificada será usada para pruebas de programación y procesos algorítmicos.

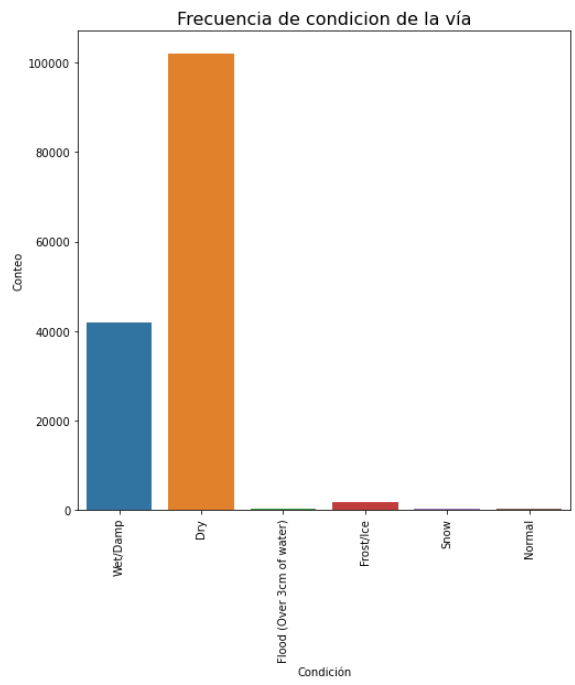
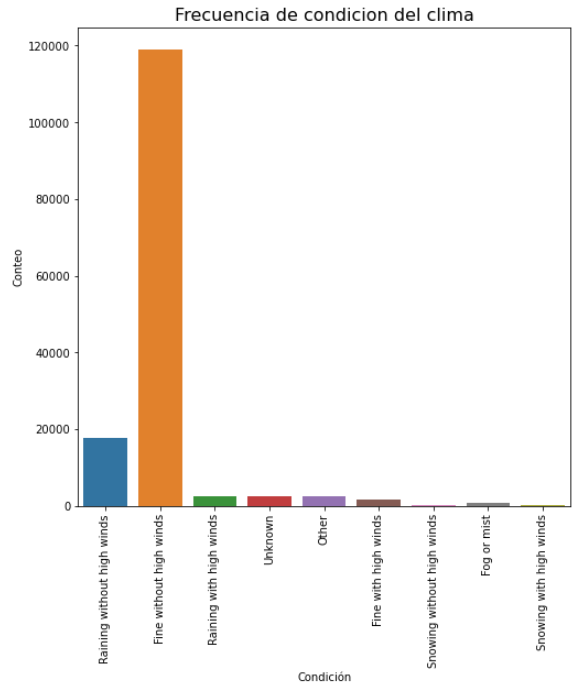
Exploración de variables

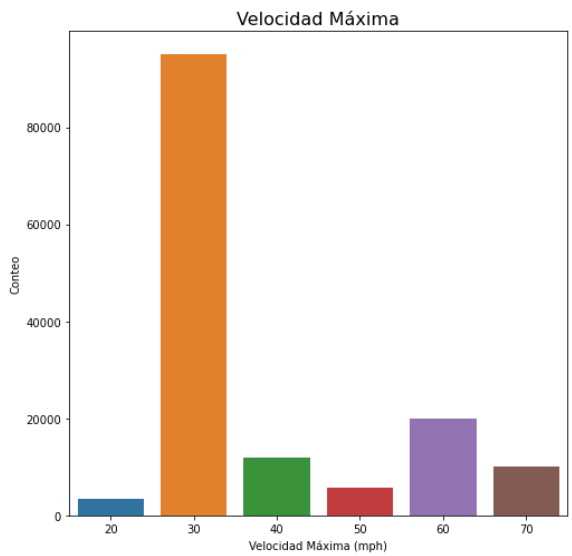
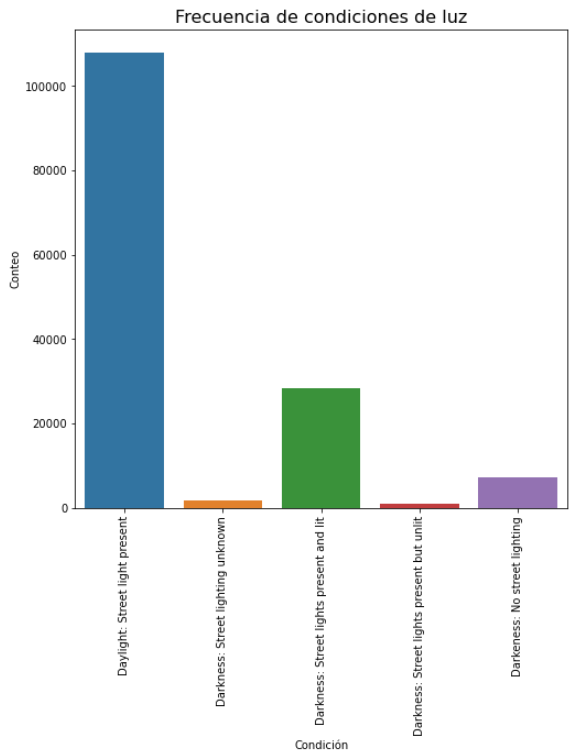
La exploración de variables es importante para poder realizar el modelo ya que nos permiten visualizar cómo se relacionan estas con la variable objetivo, para realizar la exploración se debe tener establecido las variables que se van a analizar. Por tanto, existe una lista de variables que son importadas para poder calcular datos estadísticos e histogramas que serán importantes para leer y describir la problemática, en este caso los accidentes en Reino Unido y sus implicaciones.

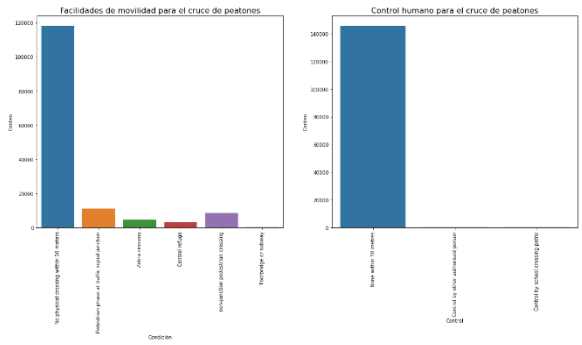
Histogramas

Luego de definir las variables que se van a utilizar, se procede a obtener las gráficas donde podrán apreciar las cifras de las condiciones que influyen en los accidentes.

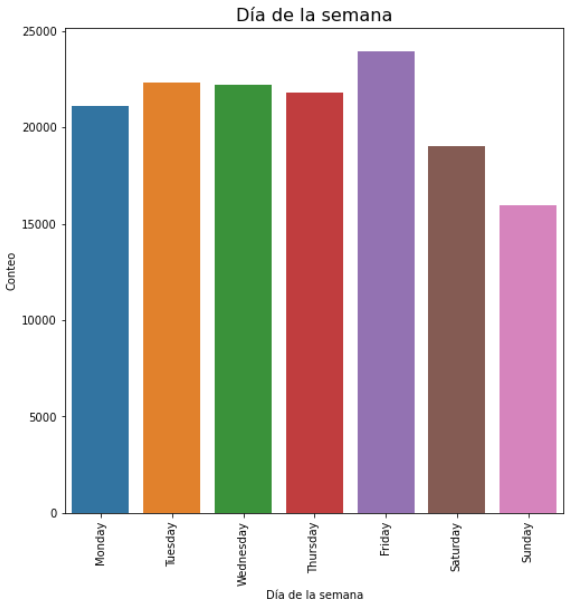
Entre las principales se encuentran:

	Del análisis de la condición de la vía se puede apreciar que la condición donde más se presentan accidentes es el seco, seguido por la vía mojada o húmeda.
	La mayor parte de accidentes del total ocurren durante el día cuando las luces de las calles no están iluminadas, y otro porcentaje apreciable durante la noche con luces presentes y encendidas (condiciones óptimas).

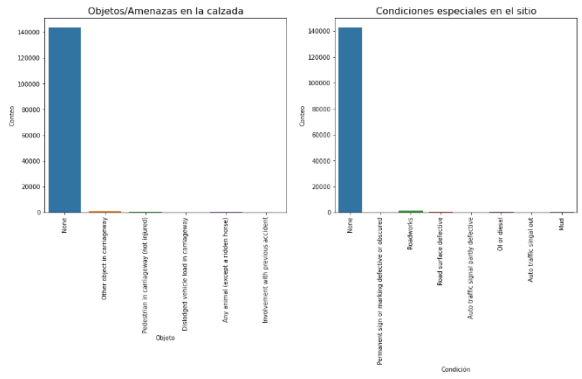
	Según los datos, se muestra que la mayoría de los accidentes en un gran porcentaje significativamente alto se presentan cuando los vehículos o transporte circulan a la velocidad permitida en áreas urbanas, entre 30 y 40 millas por horas (siendo 30 millas la velocidad máxima permitida en áreas urbanas). y otro porcentaje indica que muchos accidentes ocurren entre 60 y 70 millas por horas, que son las velocidades máximas en autopistas principales de una y doble calzada respectivamente.
	La mayor parte de accidentes del total ocurren durante el día cuando las luces de las calles no están iluminadas, y otro porcentaje apreciable durante la noche con luces presentes y encendidas (condiciones óptimas).



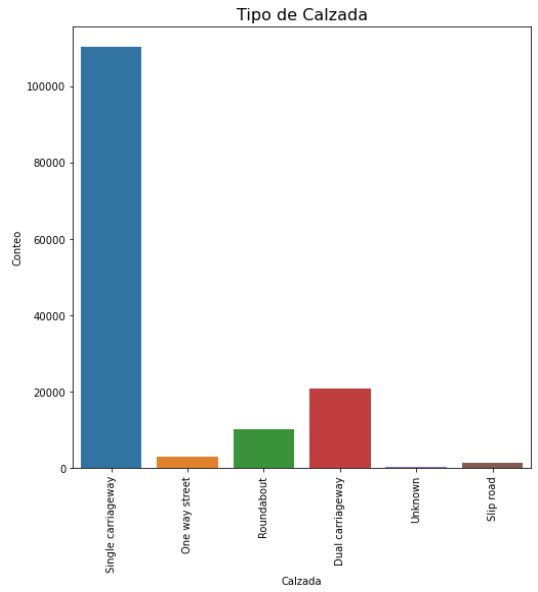
El porcentaje de accidentes es significativamente grande cuando no existen facilidades de movilidad para que crucen los peatones en un radio de 50 metros, mientras que otros dos porcentajes notables se presentan en las intersecciones de varios cruces: 'pedestrian phase at traffic signal juction' y 'non-juction pedestrian crossing'.



Las cantidades de accidentes por día no difieren mucho, pero el día más marcado en siniestros viales o accidentes peatonales es el viernes, y el domingo es el día con menor frecuencia de accidentes.

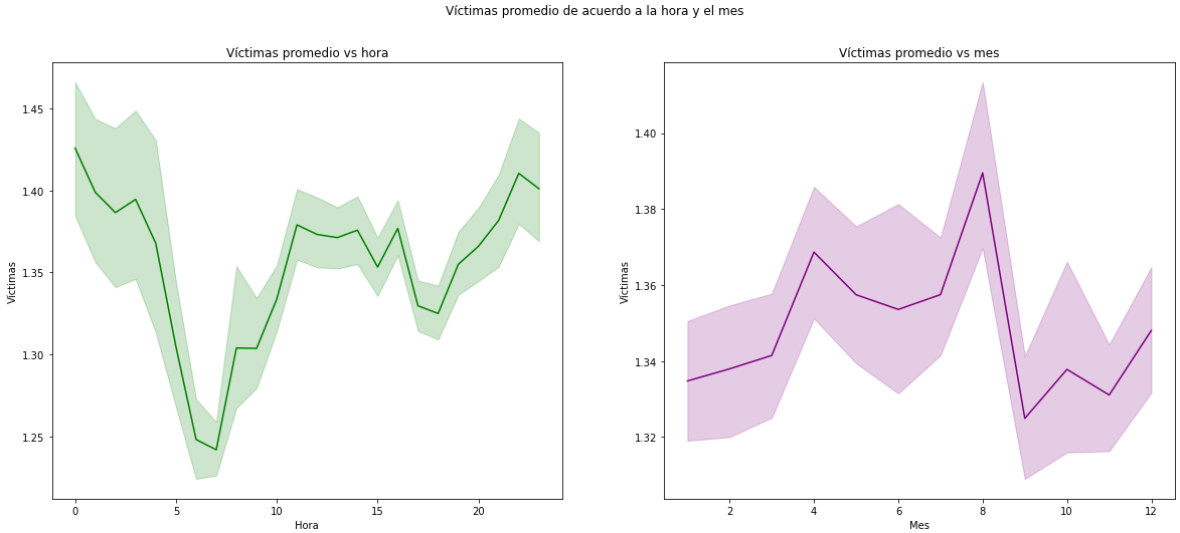


Se puede apreciar que en la mayoría de los accidentes no hay amenazas ni condiciones especiales que se encuentren en la vía durante la ocurrencia de estos



Se puede observar que la mayor parte de los accidentes en total se producen en calzadas únicas y calzadas dobles, esto tiene lógica ya que son estas vías donde circulan y transitan muchos carros con alta intensidad.

Accidentes por mes y por hora

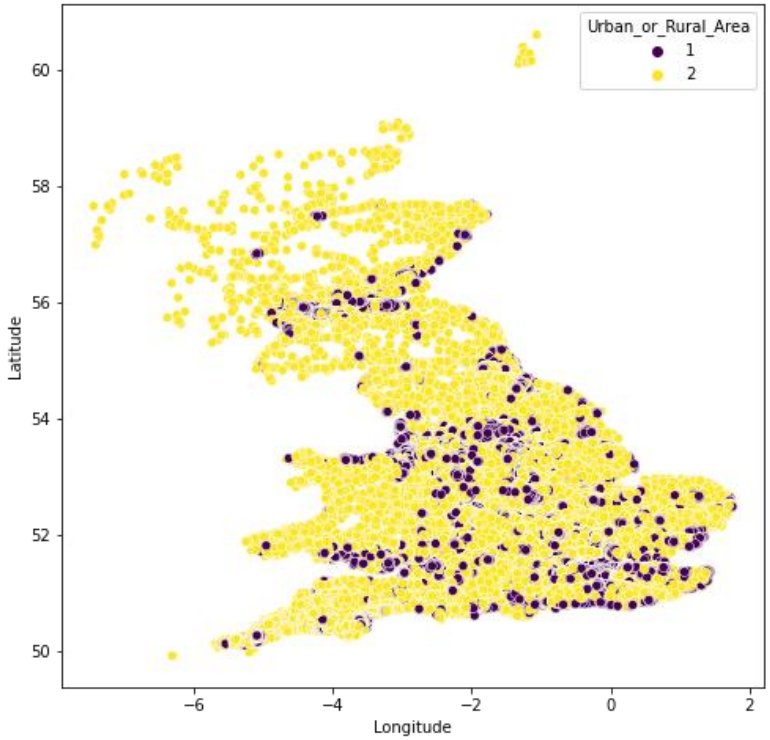


Se puede observar que los meses que el número de víctimas promedio por hora tiene una disminución bastante notable entre las 5 y 8 de la mañana.

De las víctimas promedio por mes se aprecia un pico en el mes de agosto y un pequeño pico en el mes de abril.

Identificación Área Urbana/ Rural

Se puede apreciar que en la columna de Area rural o urbana solo hay dos valores, 1 y 2. Sin embargo el dataset no nos indica a cuál hace referencia 1 y 2 en la columna Urban_or_Rural_Area, pero se puede deducir a partir de los datos de la longitud y latitud y el mapa del Reino Unido, donde se puede deducir que el valor de 1 corresponde a Urbano y 2 a Rural.

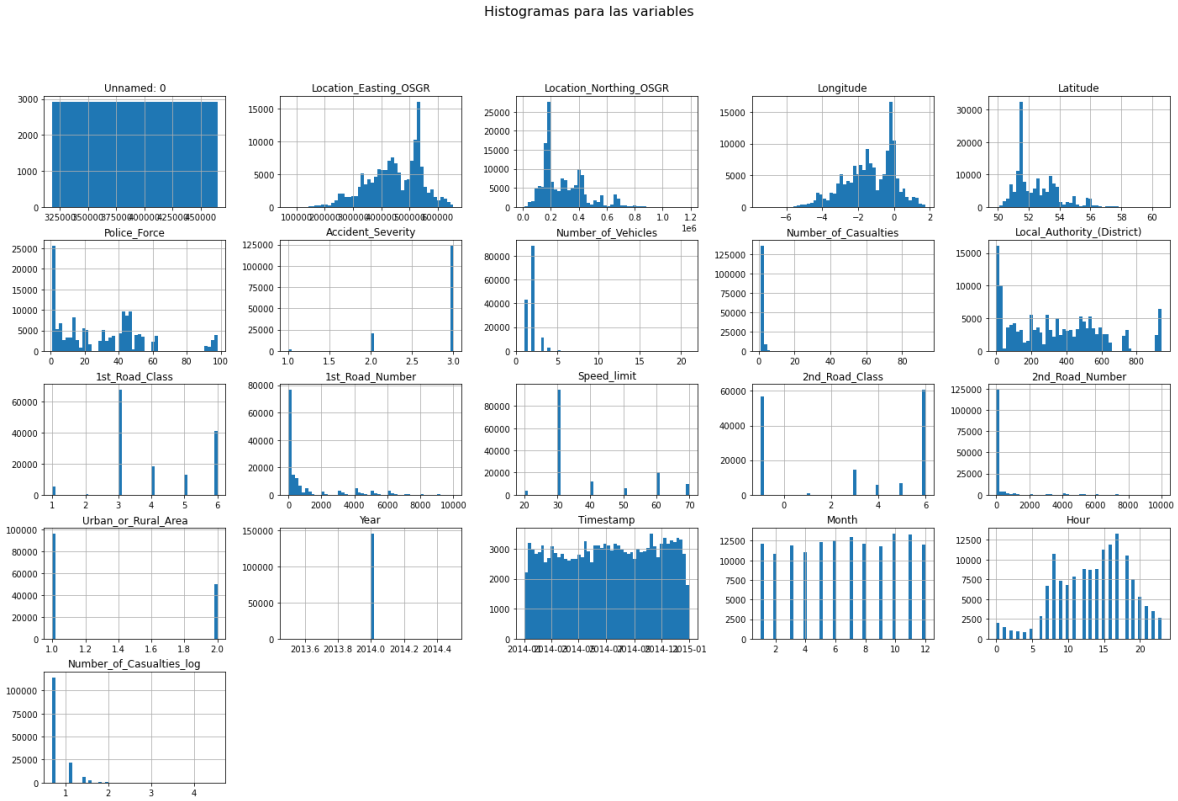


Correlación entre parámetros y variable objetivo

	Number_of_Casualties
Number_of_Casualties	1.000000
Number_of_Casualties_log	0.904197
Number_of_Vehicles	0.229829
Speed_limit	0.138503
Urban_or_Rural_Area	0.114192
Unnamed: 0	0.031783
Latitude	0.029246
Location_Northing_OSGR	0.029116
Local_Authority_(District)	0.020365
Hour	0.015797
Police_Force	0.013969
1st_Road_Number	0.005484
2nd_Road_Number	0.000482
Month	-0.001141
2nd_Road_Class	-0.034233
Longitude	-0.034669
Location_Easting_OSGR	-0.035971
Accident_Severity	-0.058472
1st_Road_Class	-0.079708
Year	NaN

Se puede observar que, para la variable objetivo, los parámetros que más se relacionan con estas son el número de vehículos involucrados, la velocidad límite de la zona y si se trata de un área urbana o rural.

Distribución de las variables numéricas



Simulación de datos faltantes

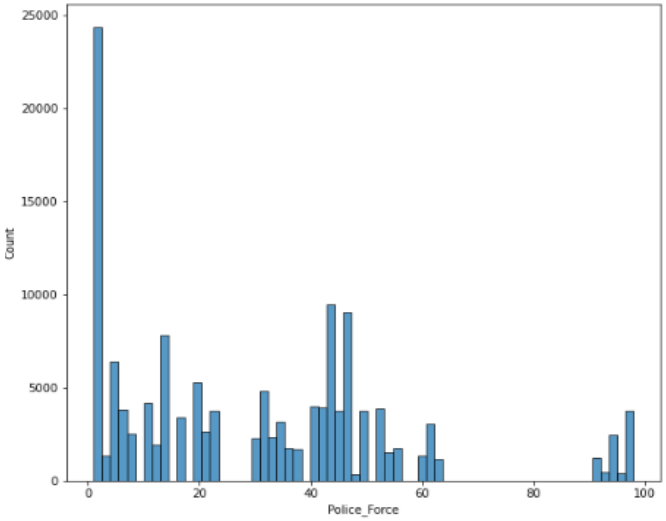
Teniendo en cuenta los requisitos del proyecto, el dataset al menos ha de tener un 5% de datos faltantes en al menos 3 columnas, el dataset actualmente contiene datos faltantes en una columna, la cual es LSOA_of_Accident_Location, por lo que es necesario simular la falta de datos de al menos dos columnas más, en este caso se escogieron Road_Type, Police_Force y Number_of_Vehicles, por lo que los datos faltantes quedaron distribuidos de la siguiente manera:

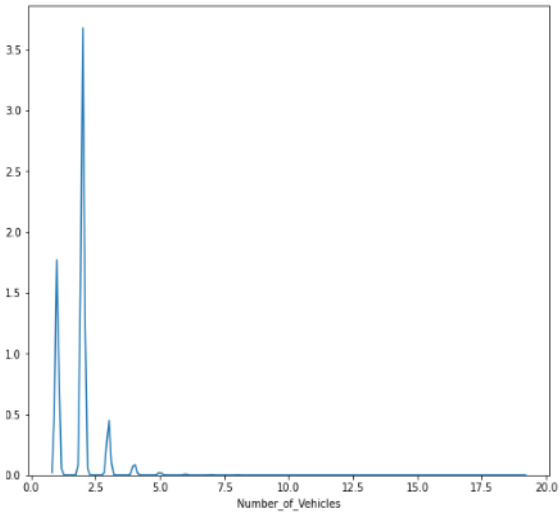
	Total	Percent
LSOA_of_Accident_Location	9277	6.340127
Road_Type	7316	4.999932
Police_Force	7316	4.999932
Number_of_Vehicles	7316	4.999932

Tratamiento de datos

Es posible ver que en la columna LSOA_of_Accident_Location faltan alrededor del 6% de los datos, estos corresponden a una notación o nomenclatura y cuyos datos son únicos para cada zona del Reino Unido, estos datos nos dan poca información y supondría una

tarea tediosa rellenar los datos faltantes con los correctos, por lo que es mejor eliminar esa columna. En cuanto a los policías que atendieron la zona, y el número de vehículos se puede analizar primero la distribución que siguen estos datos, la cual es la siguiente.





	Mediana	Moda
Police_Force	30.0	1.0
Number_of_Vehicles	2.0	2.0

Al observar la distribución que siguen los datos y los datos de mediana y moda se pueden rellenar los datos faltantes con la moda, Para el tipo de vía se pueden agrupar todos los datos faltantes en la clase 'Unknown'.

Eliminación de variables no relevantes

Se pueden eliminar las siguientes variables del dataset ya que no representan información relevante o es información duplicada:

- Location_Easting_OSGR y Location_Northing_OSGR es información duplicada de Longitude y Latitude.
- Year ya que todos los datos son del 2014.
- Number_of_Casualties_log ya que es una columna que se creó para mejor visualización de datos
- Unnamed: 0 y Accident_Index ya que solo son un identificador del accidente
- Carriageway_Hazards y Special_Conditions_at_Site ya que la mayoria de datos son nulos
- Pedestrian_Crossing-Physical_Facilities y Pedestrian_Crossing-Human_Control debido a que en la mayoría de los accidentes no hay ninguna facilidad de movilidad en al menos 50 metros.

Creación de variables

Adicional a las variables que ya tenemos se crearon variables que pueden ayudar a describir mejor la situación durante los accidentes, para ello se establecieron tres variables nuevas, las cuales son:

- Una variable que nos indique si es de día o de noche en el momento del accidente
- Una variable que nos indique la estación del año en el momento del accidente
- Una variable binaria que nos indique si está bien iluminada la zona o n

Además, se optó por crear una variable que clasifique de manera categórica las víctimas, de la siguiente manera:

- Para los accidentes con menos de 5 víctimas se clasificaron cómo accidentes leves.
- Para los accidentes con menos de 10 víctimas se clasificaron como accidentes moderados.
- Para los accidentes con más de 10 víctimas se clasificaron como accidentes graves.

Con esta clasificación se busca obtener una gravedad del accidente y un número de víctimas estimado, en lugar de un número exacto de estas, pues la información relevante es si se va a presentar un número alto o bajo de estas.

Métodos supervisados

Para los métodos no supervisados se utilizaron cómo modelos el random forest classifier, el decision tree classifier y el SVC. Mediante el uso del cross validation y por medio de una adaptación del código proporcionado de ejemplo se eligió el mejor modelo para los datos, sin embargo, cabe resaltar que la diferencia en los errores RSME de los tres modelos varia muy poco y también podrían generar modelos efectivos, no obstante, se decidió trabajar únicamente con el clasificador “Decision tree”.

```
---------
RMSE Test:  0.10299 (± 0.00221915 )
RMSE  Train:  0.10176 (± 0.00167175 )
---------
RMSE Test:  0.10177 (± 0.00248090 )
RMSE  Train:  0.10266 (± 0.00174965 )
---------
RMSE Test:  0.10235 (± 0.00172781 )
RMSE  Train:  0.10225 (± 0.00129624 )
Seleccionado:  1

 Mejor modelo:
DecisionTreeClassifier(max_depth=3)
```

En la imagen se pueden observar los errores obtenidos para cada modelo, siendo el random forest classifier, el decision tree classifier y el SVC respectivamente.

A continuación, se procede a encontrar los mejores hiperparámetros para el modelo, esto se realiza a través del GridSearchCV, la cual es una herramienta del Scikit Learn para realizar un cross validation utilizando diferentes parámetros especificados antes de ejecutar el código, se obtuvieron los siguientes resultados:

```
Fitting 5 folds for each of 5 candidates, totalling 25 fits
Mejores parámetros para el estimador Decision Tree:  {'max_depth': 2}
```

```
Modelo_selec = DecisionTreeClassifier(max_depth=2)
Modelo_selec.fit(Xtv, ytv)

print('El error RSME del modelo de Decision Tree Classifier es\n En test: '+str(RMSE(yts,  Modelo_selec.predict(Xts)))+
      '\n En train: '+str(RMSE(ytv,  Modelo_selec.predict(Xtv))))
```

```
El error RSME del modelo de Decision Tree Classifier es
 En test: 0.10124855563637 58
 En train: 0.10230442207776677
```

Métodos no supervisados

Para los métodos no supervisados se procedió a realizar un PCA, el cual es una función que permite obtener los datos mas representativos del dataset con el fin de realizarles una transformación y obtener mejores resultados con el modelo del decision tree. El análisis se realizó a través del siguiente código:

```
from sklearn.decomposition import PCA
components = [1,3,5]
test_size = 0.3
val_size  = test_size/(1-test_size)
perf = [] #desempeños de los modelos
Dec_tree = DecisionTreeClassifier(max_depth = 15)
for i in components:
    pca = PCA(n_components = i)
    X_t = pca.fit_transform(X)

    Xtv, Xts, ytv, yts = train_test_split(X_t, y, test_size=test_size)
    print (Xtv.shape, Xts.shape)

    Dec_tree.fit(Xtv, ytv)
    perf.append(RMSE(yts , Dec_tree.predict(Xts)))
    print('RMSE del modelo con ', i ,'elementos: ',"{:.5f}".format(RMSE(yts , Dec_tree.predict(Xts))))
    print('------------')

print('Mejor RMSE: ',"{:.5f}".format(np.min(perf)),'  ; obtenido con ',components[np.argmin(perf)],' componentes para PCA')
```

Curvas de aprendizaje

Las curvas de aprendizaje representan cómo se comportaría el modelo en caso al momento de ir agregando más datos a este a lo largo del tiempo, se presentaron las siguientes graficas:

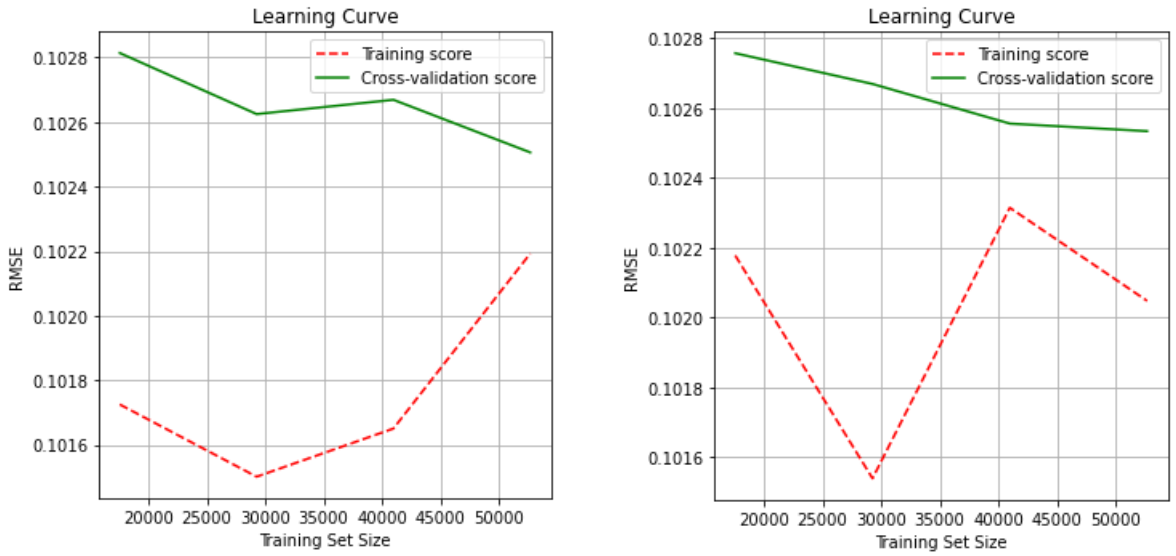


Figura 3: Curvas de aprendizaje, (De derecha a izquierda: Decision tree, Decision tree + PCA)

De ambas graficas se puede observar una tendencia al Bias, es decir a un sesgo, lo cual implica que o los modelos son muy simples o hay datos mezclados y se necesitan más columnas en el dataset. Cabe recalcar que en el caso específico del modelo con PCA, se observa un comportamiento errático de los datos de entrenamiento, además de no presentar una mejora significativa en cuanto a la métrica con respecto al modelo supervisado, que podría surgir de usar una reducción tan grande de columnas con el PCA.

Retos y condiciones de despliegue del modelo

Los principales retos que representa este modelo con el fin de predecir las víctimas en los posibles accidentes automovilísticos es la recolección de datos más significativos para estos, debido a que, aunque el dataset presenta mucha información de cada uno los accidentes y recopila una cantidad considerablemente grande de ellos, estos no logran desempeñarse de una buena manera. Por lo que implicaría plantearse la posibilidad de agregar más columnas al dataset, y evaluar los posibles costos de realizar dicha tarea, por lo que se considera que el modelo no se encuentra apto para desplegarse en producción.

Sin embargo, si se deseara realizarse un modelo apto para la producción se debe tener en cuenta los siguientes retos: La tarea de recopilar mas columnas de los accidentes, y como ya se había mencionado antes, los costos de realizar dicha tarea. Evaluar de manera más activa con los centros de salud, ambulancias y paramédicos la viabilidad del modelo y si este les permite obtener una mejora significativa en la eficiencia a la hora atender estos accidentes automovilísticos. Y por último, se debe tener en cuenta el reto de obtener la información para el modelo a tiempo real para que este funcione de manera correcta.

Conclusiones

- Se sugiere obtener más datos representativos para el dataset si se quiere que funcione de la manera más optima posible y así reducir el sesgo de estos.

- Al realizar la selección del modelo se puede obtener resultados muy similares entre los tres modelos, por lo que se sugiere realizar el análisis con los otros modelos.

- El sesgo de los datos también podría deberse a la propia naturaleza de los datos, puesto que estos se acumulan mucho en valores cercanos a 1 y darían problemas a los modelos para “aprender”.

Referencias

Road Accident (United Kingdom (UK)) Dataset. (2022, May 28). Kaggle.https://www.kaggle.com/datasets/devansodariya/road-accident-united-kingdom-uk-dataset.