SEGUNDO INFORME INTELIGENCIA ARTIFICIAL

POR:

Jorge Antonio Franco Vásquez

Felipe Carlos Martínez Mármol

Juan Camilo Tabares Henao

MATERIA:

Introducción a la Inteligencia Artificial

PROFESOR:

Raul Ramos Pollan

UNIVERSIDAD DE ANTIOQUIA

FACULTAD DE INGENIERÍA

MEDELLÍN 2022

INFORME ANALISIS DE DATOS

Selección de datos:

En esta parte del trabajo establecemos que 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 las 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.

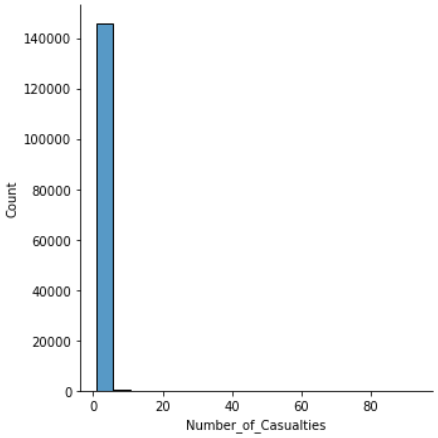


Figura 1. Distribución de la variable objetivo.

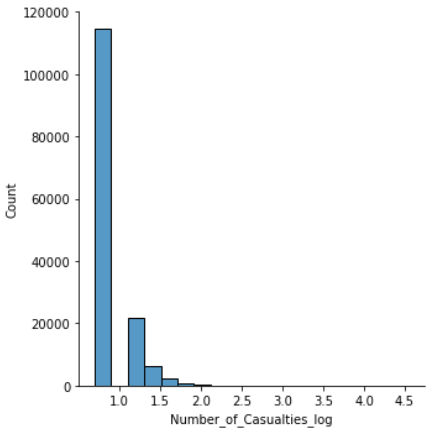


Figura 2. Transformación logaritmica

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 la 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.

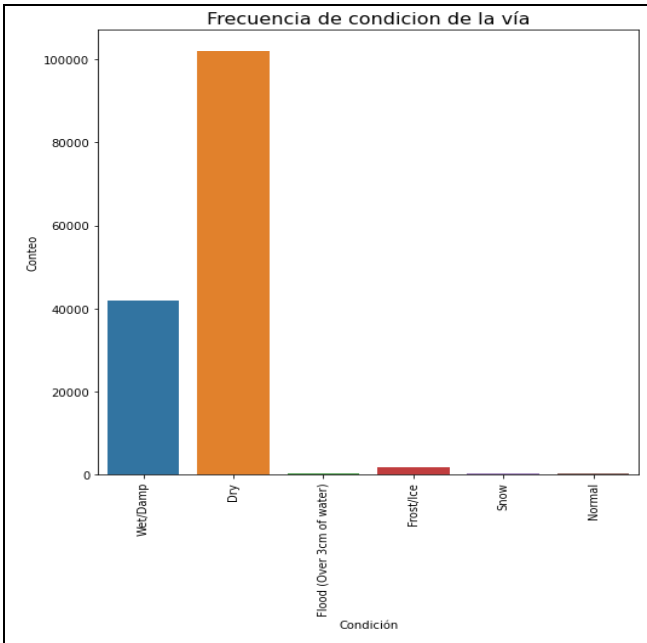
Exploración de variables

La exploración de variables es importante para poder realizar el modelo ya que nos permiten visualizar como 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.

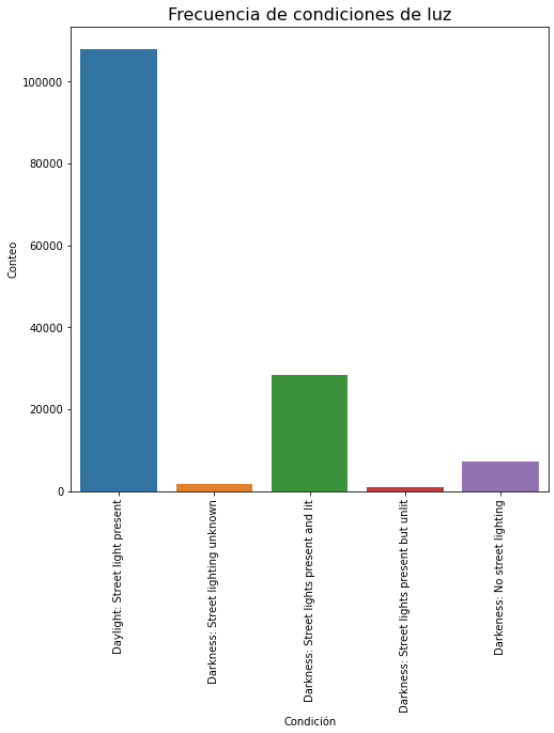
Histogramas.

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:

Condiciones del medio:

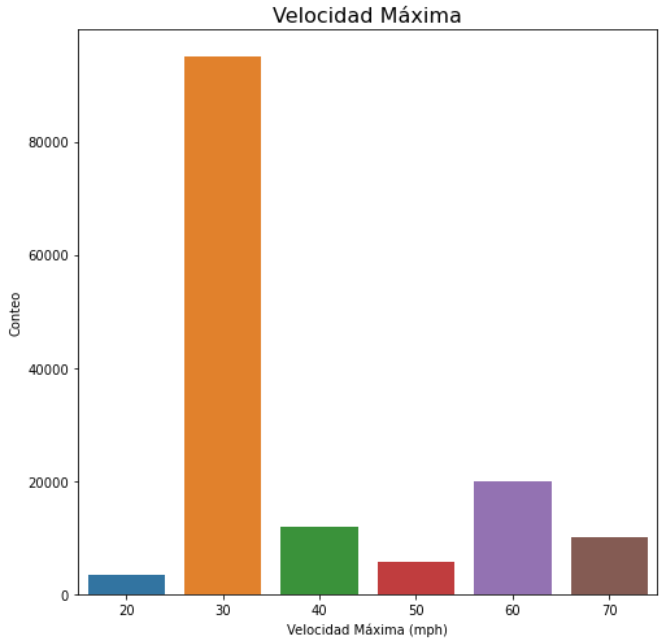


Del análisis de la condición de la vía se puede apreciar que la condición donde más se presentan accidentes es el seco, seguido por la vía mojada o húmeda.



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).

Velocidad Máxima:



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.

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 el 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 más columnas, en este caso se escogieron:

- Police_Force
- Road_Type
- Number_of_Vehicles

Tratamiento de datos

Rellenar datos faltantes:

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.

Para las columnas Number_of_Vehicles y Police_Force se decidió usar la moda a partir de un análisis estadístico.

Para el tipo de vía se pueden agrupar todos los datos faltantes en la clase 'Unknown'.

Eliminación de variables no relevantes para el modelo

Se eliminaron variables que consideramos tenían poca información relevante, era información duplicada o que tenia poca correlación con la variable objetivo según el análisis realizado.

Añadir variables que pueden ser relevantes y convertir variables categóricas a numéricas

Se añadieron 3 variables que podrían ser relevantes en un accidente de tránsito, estas son la estación del año, si es de día o de noche y si la zona se encontraba iluminada durante el accidente. Una vez realizado ese proceso se convirtieron todas las variables categóricas a variables numéricas asignándoles números con la función *LabelEncoder*.

Primer modelo

Para generar el primer modelo se dividió primero el dataset en una matriz X (las variables) e y (la variable objetivo), y luego se hizo una división para train y test, finalmente se probó un modelo de '*Decision Tree*' con dos niveles de complejidad (3 y 5), y se le calculó el error al modelo, el cual dio valores muy cercanos: 0.75 y 0.74 respectivamente.