SEGUNDO INFORME INTELIGENCIA ARTIFICIAL.

POR:

BRIAN JOURNEYT RODRIGUEZ VIVAS.

JUAN DANIEL MEJÍA MARTINEZ.

JORGE LUIS GONZALEZ MORELO.

MATERIA:

INTRODUCCIÓN A LA INTELIGENCIA ARTIFICIAL

PROFESOR:

RAUL RAMOS POLLAN.

UNIVERSIDAD DE ANTIOQUIA.

FACULTAD DE INGENIERÍA.

MEDELLÍN.

2023.

INFORME ANALISIS DE DATOS.

1) ELECCIÓN DE DATOS:

En esta sección del informe se detallan los datos más relevantes sobre la problemática de los accidentes en el Reino Unido. En primer lugar, se procedió a la lectura del archivo CSV denominado "UK_Accident.csv". A partir de los datos del año 2013 se realizaron análisis para identificar los principales factores que influyen en los accidentes en el Reino Unido, y su vez predecir los datos del año 2014.

Posteriormente, se convirtieron los datos en formato de cadena de caracteres a datos de tiempo, utilizando la función "To_datetime" de la librería Pandas, que devuelve una indicación de fecha y hora a partir de una cadena de caracteres.

Finalmente, se calculó un resumen de las principales estadísticas de los datos, incluyendo el recuento total, la desviación estándar, los valores máximos y mínimos. Esto se realizó mediante el uso de la función "describe()", la cual permite devolver un resumen estadístico de todas las columnas del DataFrame.

2) MATRIZ DE CORRELACIÓN.

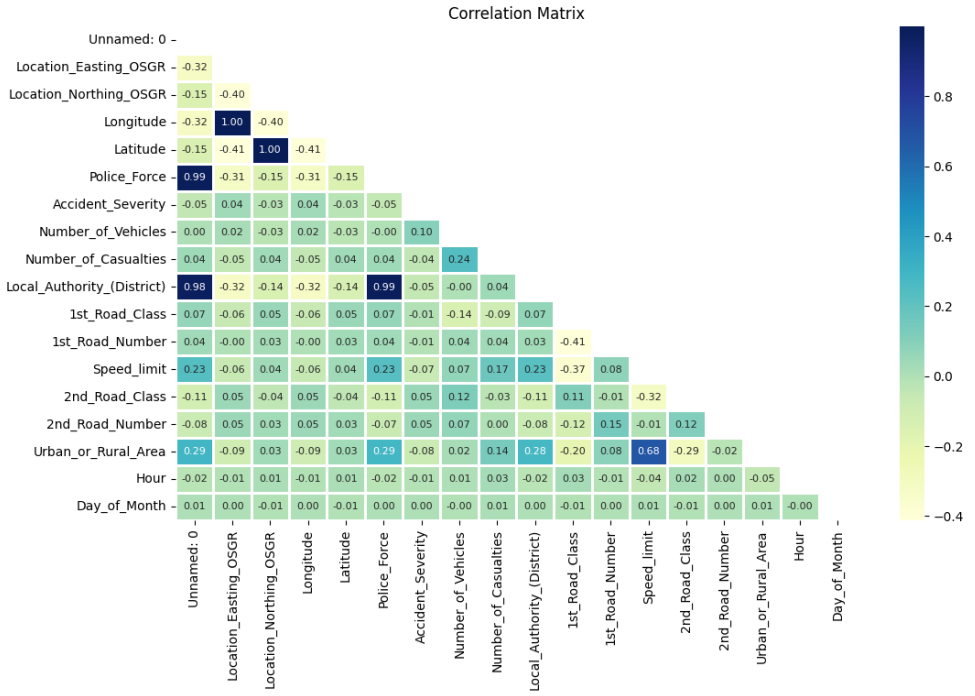

Ilustración 1. Matriz de correlación.

La matriz de correlación en el conjunto de datos de accidentes de carretera del Reino Unido disponible en Kaggle indica la relación entre las diferentes variables presentes en el conjunto de datos. Esta matriz es una tabla cuadrada que muestra la correlación entre pares de variables. La correlación es una medida estadística que indica la fuerza y dirección de la relación entre dos variables.

La matriz de correlación puede ser utilizada para identificar qué variables están más fuertemente correlacionadas entre sí. Si dos variables están altamente correlacionadas, esto puede indicar que hay una relación causal o que ambas variables están influenciadas por una tercera variable. La matriz de correlación también puede ayudar a identificar patrones y tendencias en los datos y puede ser útil en la selección de variables para modelos de análisis predictivo.

3) VARIABLE OBJETIVO:

La variable objetivo "Accident_Severity" se utiliza para predecir la gravedad de los accidentes. Esta variable es importante ya que permite a los investigadores y las autoridades de tránsito comprender mejor los factores que contribuyen a los accidentes graves y tomar medidas preventivas para reducir su frecuencia. Al analizar las características de los accidentes y los factores que los rodean, los investigadores pueden desarrollar modelos predictivos que ayuden a prever la probabilidad de un accidente grave y tomar medidas para evitarlo. En resumen, la variable objetivo "Accident_Severity" es una herramienta clave en la prevención de accidentes de tráfico graves y la promoción de la seguridad vial.

4) ANÁLISIS DE LA VARIABLE OBJETIVO:

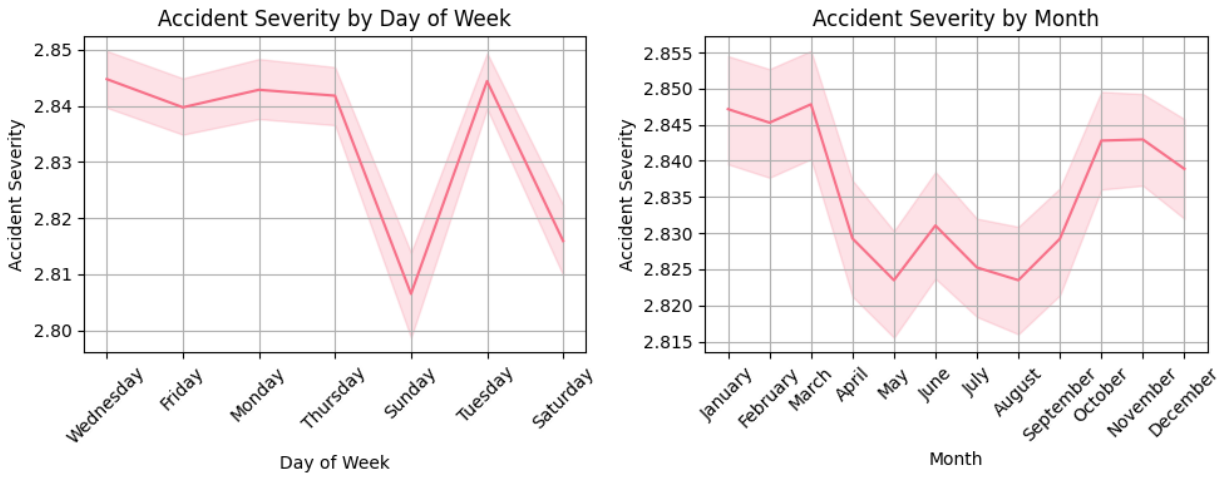

Ilustración 2. Severidad de accidentes en días y meses.

La matriz de correlación puede ser útil para la predicción de la severidad de los accidentes de carretera en el conjunto de datos del Reino Unido, ya que puede ayudar a identificar qué variables están más fuertemente correlacionadas con la severidad de los accidentes. Por ejemplo, si se encuentra una alta correlación entre la velocidad del vehículo y la gravedad del accidente, esto puede indicar que la velocidad es un factor importante en la predicción de la gravedad del accidente. De esta manera, la matriz de correlación puede ser utilizada para seleccionar las variables más relevantes para incluir en un modelo de análisis predictivo de la severidad de los accidentes.

Además, la matriz de correlación también puede ser utilizada para detectar multicolinealidad, es decir, la presencia de alta correlación entre las variables predictoras. La multicolinealidad puede afectar negativamente el rendimiento de un modelo predictivo al reducir la precisión de las estimaciones de los coeficientes de las variables predictoras. Por lo tanto, la matriz de correlación puede ser utilizada para identificar y eliminar variables altamente correlacionadas, mejorando así la precisión del modelo predictivo. En las gráficas anteriores se da a conocer la severidad de los accidentes teniendo como longitud de tiempo días y meses.

5) 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 Unidoy sus implicaciones.

6) DIAGRAMAS CIRCULARES:

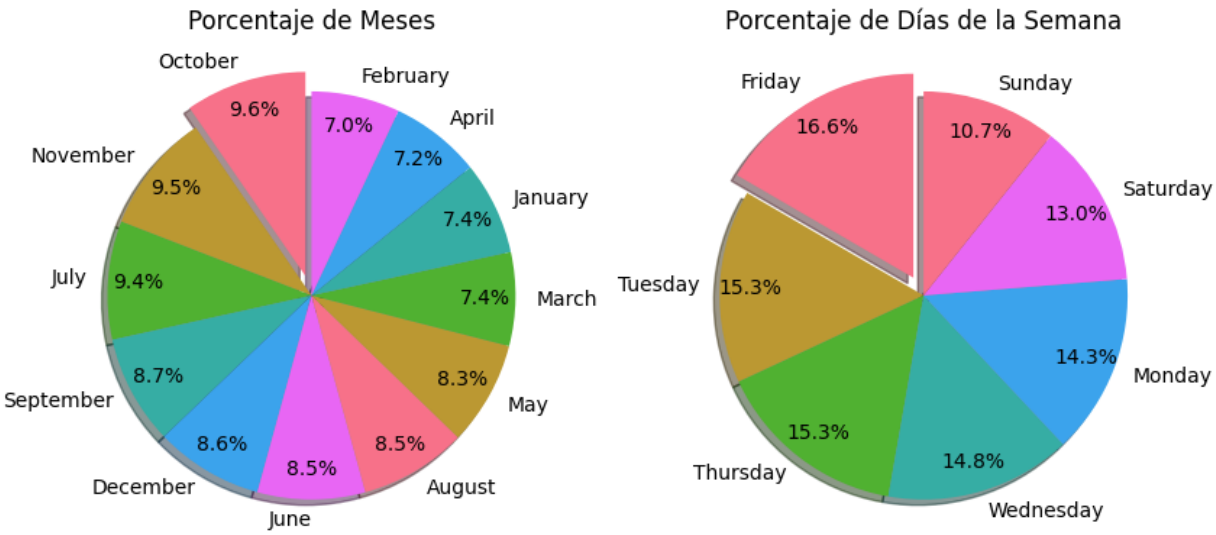

Ilustración 3. Diagramas circulares con porcentajes de los datos obtenidos.

7) 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 las 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

8) TRATAMIENTO DE DATOS:

8.1 Rellenar datos faltantes

En cuanto a las columnas "Number_of_Vehicles" y "Police_Force", se ha optado por utilizar la moda como método estadístico para llenar los valores faltantes. En el caso de la columna que indica el tipo de vía, se ha agrupado todos los datos faltantes en la categoría "Unknown".

8.2 Eliminación de variables no relevantes para el modelo

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

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

Se han incluido tres nuevas variables que podrían ser relevantes en un accidente de tráfico: la estación del año, si el accidente ocurrió de día o de noche, y si la zona en la que tuvo lugar estaba iluminada. Después de esta adición, se han convertido todas las variables categóricas en variables numéricas mediante la función LabelEncoder, asignándoles un número correspondiente.