INTRODUCCIÓN A LA INTELIGENCIA ARTIFICIAL

SEGUNDA ENTREGA

ANALISIS DE DATOS.

POR:

ANA MARIA ROMERO CARVAJAL

JEAN CARLOS JULIO RODRIGUEZ

PROFESOR:

RAUL RAMOS POLLAN.

UNIVERSIDAD DE ANTIOQUIA.

FACULTAD DE INGENIERÍA.

MEDELLÍN.

2023.

1) ELECCIÓN DE DATOS Y CORRELACION:

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.

La correlación es una medida estadística que cuantifica la fuerza y dirección de la asociación entre dos variables. La matriz de correlación del conjunto de datos de accidentes de carretera del Reino Unido es una tabla que representa la relación entre pares de variables.

La matriz de correlación puede utilizarse para identificar las variables más estrechamente relacionadas entre sí. Si dos variables están altamente correlacionadas, esto puede indicar que existe una relación causal entre ellas 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.

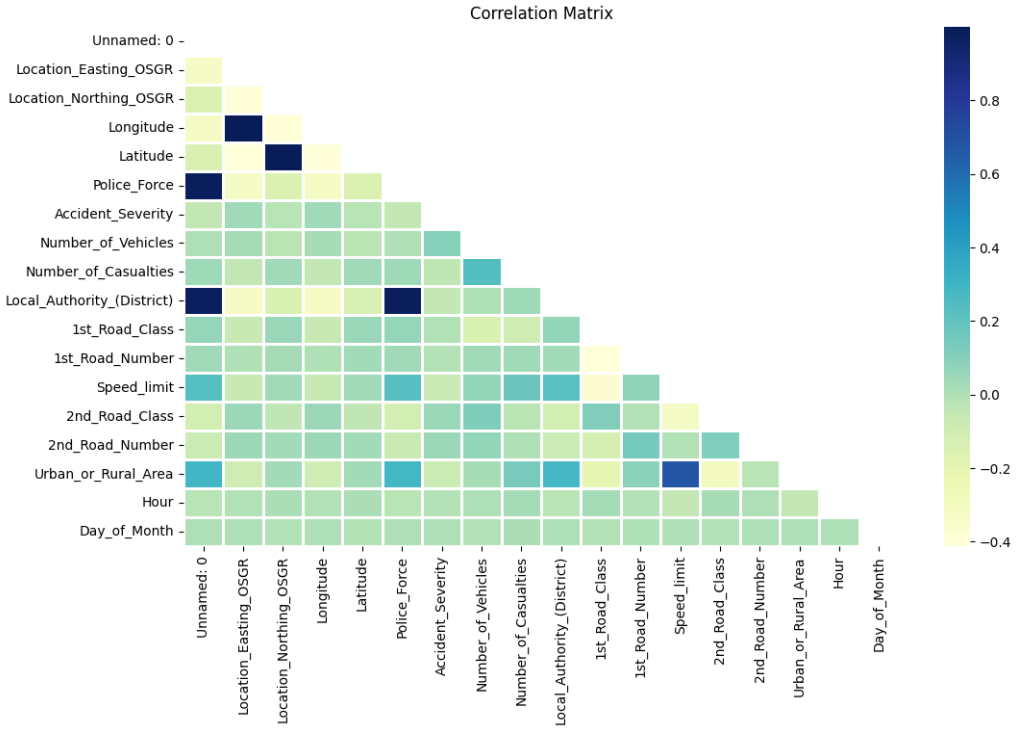

Ilustración 1. matriz de correlación.

2) ANALISIS DE LA VARIABLE OBJETIVO

La variable de interés "Accident_Severity" se emplea con el propósito de pronosticar el nivel de gravedad en incidentes viales. Su relevancia radica en la capacidad que brinda a los investigadores y las autoridades de tránsito para adquirir un conocimiento más profundo de los determinantes asociados a accidentes de gran envergadura, permitiendo la implementación de estrategias preventivas destinadas a disminuir su ocurrencia. Al efectuar un análisis de las particularidades inherentes a los accidentes y sus circunstancias circundantes, los investigadores están en condiciones de desarrollar modelos predictivos que faciliten la anticipación de la probabilidad de ocurrencia de incidentes graves y, consiguientemente, tomar medidas correctivas para su mitigación. Concluyendo, la variable objetivo "Accident_Severity" representa un componente crítico en el ámbito de la prevención de accidentes de tráfico de alta magnitud y en la promoción de la seguridad vial.

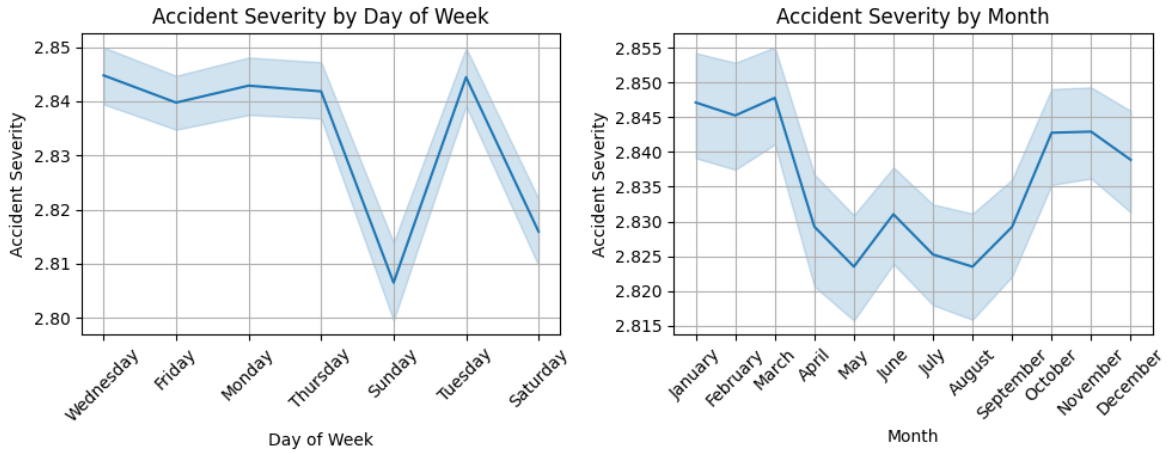

Ilustración 2. Gráfico representativo de la variable objetivo (severidad de accidentes durante meses y días).

La matriz de correlación se presenta como una herramienta de considerable utilidad en el contexto de la predicción de la severidad de los incidentes viales en el conjunto de datos del Reino Unido. Su valor radica en la capacidad para discernir las relaciones de mayor magnitud entre las diversas variables y la severidad de los accidentes. Por ejemplo, una correlación significativa entre la velocidad del vehículo y la gravedad del accidente podría insinuar que la velocidad representa un factor influyente en la anticipación de la gravedad del incidente. De esta manera, la matriz de correlación se convierte en una herramienta de ayuda para la selección de variables pertinentes a incluir en un modelo analítico con fines predictivos en torno a la severidad de los accidentes.

Adicionalmente, la matriz de correlación se emplea para la detección de multicolinealidad, es decir, la presencia de una alta correlación entre las variables predictoras. La multicolinealidad puede ejercer un impacto adverso en el desempeño de un modelo predictivo, reduciendo la precisión de las estimaciones de los coeficientes de las variables independientes. Por consiguiente, la matriz de correlación cumple un rol fundamental en la identificación y posterior eliminación de variables que mantienen una correlación sustancial, lo cual conlleva a una mejora en la precisión del modelo predictivo.

Es relevante destacar que en los gráficos previamente expuestos se presenta una representación de la severidad de los accidentes a lo largo del tiempo, considerando tanto los días como los meses.

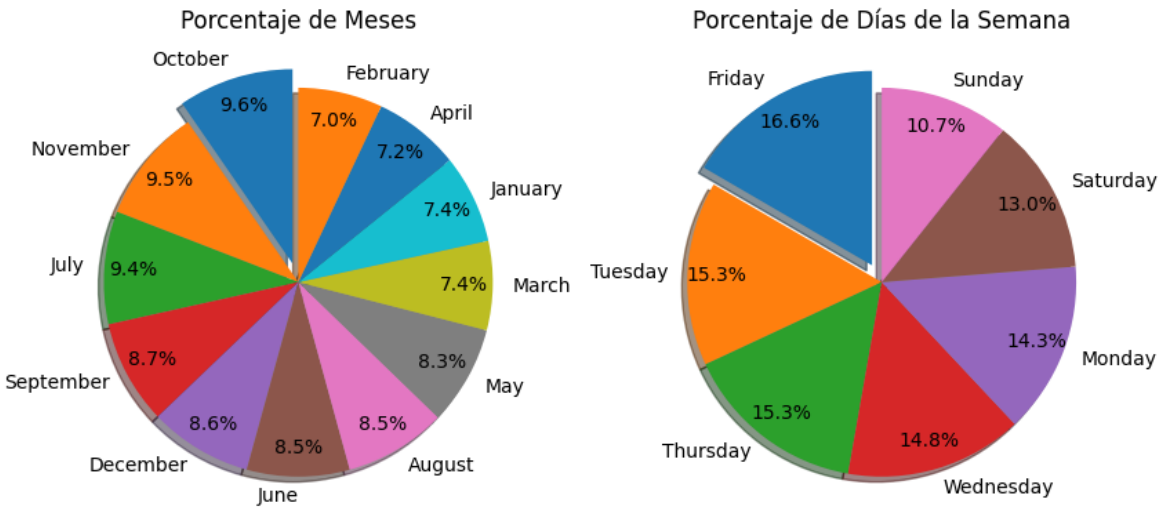

Ilustración 3. porcentaje de severidad de accidentes por meses y días de la semana.

3)EXPLORANDO OTROS DATOS

La exploración de variables también desempeña un papel fundamental en la construcción de un modelo, ya que nos brinda la capacidad de visualizar las relaciones existentes entre dichas variables y la variable objetivo. Para llevar a cabo esta exploración de manera efectiva, es exigente contar con un conjunto predefinido de variables que serán objeto de análisis. Por consiguiente, se hace necesario disponer de una lista de variables que serán importadas, con el fin de calcular datos estadísticos y generar histogramas que servirán como herramientas esenciales para comprender y describir la problemática abordada, en este caso, los accidentes en el Reino Unido y sus consecuencias.

4) DATOS COMPLEMENTARIOS

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

5) TRATAMIENTO DE DATOS:

5.1 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".

5.2 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.

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

Se han incorporado tres variables adicionales que potencialmente desempeñan un papel relevante en la dinámica de un accidente de tráfico: la estación del año, el período del día en el que se produjo el accidente y la condición de iluminación de la zona donde ocurrió. Tras esta incorporación, se ha procedido a la transformación de todas las variables categóricas en variables numéricas empleando la función LabelEncoder, asignándoles valores numéricos correspondientes. Este proceso de conversión facilita la manipulación y análisis de los datos, permitiendo su inclusión en modelos analíticos y contribuyendo a una comprensión más profunda de la relación entre estas variables y la severidad de los accidentes de tráfico.