PREDICCIÓN DE ACCIDENTES AUTOMOVILÍSTICOS EN REINO UNIDO.

PRESENTADO POR:

ANDRÉS MAURICIO
GÓMEZ GUAPACHA.

MATERIA:
INTRODUCCIÓN A LA
INTELIGENCIA ARTIFICIAL

PROFESOR:
RAÚL RAMOS POLLAN

UNIVERSIDAD DE ANTIOQUIA
FACULTAD DE INGENIERÍA
MEDELLÍN

2023-2

1.Introducción

En cualquier parte del mundo, los accidentes de tránsito son un grave problema que afecta a numerosas personas cada año. Estos accidentes pueden ocurrir debido a la imprudencia o confusiones de los conductores, y los datos estadísticos revelan que se han convertido en una preocupación significativa para los países. Además de enfrentarse al desafío de gestionar el flujo del tráfico, también deben hacer frente a las repercusiones en la salud de las personas afectadas.

En este contexto, el manejo adecuado de los datos se vuelve fundamental, especialmente en situaciones que implican riesgos para la seguridad de las personas, como los accidentes automovilísticos. Mediante la utilización de un conjunto de datos completo, se pretende predecir la probabilidad de accidentes teniendo en cuenta sus causas y consecuencias. Esto permitirá enfocar los esfuerzos en la prevención y reducción de la tasa de accidentes, así como proporcionar información crucial a las partes interesadas, como los hospitales, quienes podrían beneficiarse de dicha información para prepararse ante un posible aumento en la afluencia de pacientes debido a accidentes de tráfico.

Para lograr este objetivo, la inteligencia artificial se presenta como una herramienta invaluable. Gracias a sus capacidades de análisis de datos, la inteligencia artificial puede extraer información relevante de las bases de datos existentes y obtener los resultados deseados. Con su ayuda, se espera obtener conocimientos profundos sobre las causas subyacentes de los accidentes y las medidas preventivas más efectivas para reducir su incidencia. La combinación de la inteligencia artificial y el análisis de datos promete ser una poderosa alianza en la lucha contra los accidentes de tránsito y la protección de la seguridad vial.

1. Dataset

El conjunto de datos utilizado se obtuvo de Kaggle y consiste en un archivo CSV que contiene información sobre accidentes de tránsito ocurridos en el Reino Unido entre los años 2005 y 2014. Estos datos fueron recolectados por el gobierno del Reino Unido. El conjunto de datos completo contiene más de 1.8 millones de registros de accidentes. Sin embargo, para este proyecto en particular, se ha decidido utilizar únicamente los datos más recientes, correspondientes al año 2014. Estos datos abarcan un total de 146.322 accidentes.

El conjunto de datos proporciona una amplia gama de información relevante sobre cada accidente registrado. Algunos de los atributos incluidos son:

1. Ubicación geográfica: Latitud y longitud del lugar exacto donde ocurrió el accidente.
2. Fecha y hora: Información sobre la fecha y hora en la que se produjo el accidente.
3. Tipo de accidente: Clasificación del tipo de accidente, como colisión de vehículos, atropello, choque con objeto fijo, entre otros.
4. Condiciones climáticas: Descripción de las condiciones climáticas en el momento del accidente.
5. Estado de la carretera: Información sobre el estado de la carretera, como seca, mojada, helada, etc.
6. Factores contribuyentes: Factores que se consideran contribuyentes al accidente, como la velocidad, el consumo de alcohol, el uso del cinturón de seguridad, entre otros.
7. Gravedad del accidente: Indicación de la gravedad del accidente en términos de personas fallecidas, heridas graves o heridas leves.

El análisis de estos datos permitirá obtener información valiosa sobre las causas y consecuencias de los accidentes de tránsito en el Reino Unido en el año 2014. Esto a su vez facilitará la implementación de estrategias de prevención y la toma de decisiones informadas para mejorar la seguridad vial y reducir la tasa de accidentes en el futuro.

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 unpunto 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 parala locación geográfica de longitud y latitud**
Year	**Año del evento accidental**

2. Métrica.

La métrica principal empleada en el modelo de predicción de accidentes de tránsito es el Error Cuadrático Medio (RMSE, por sus siglas en inglés), que se calcula de la siguiente manera:

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

El Error Cuadrático Medio (RMSE) es la métrica principal utilizada en el modelo de predicción de accidentes de tránsito. Se calcula como la raíz cuadrada del promedio de la suma de las diferencias al cuadrado entre los valores observados en la serie y los valores esperados según el modelo de tendencia.

Donde y corresponde a los valores observados en la serie y $\hat{y}$ representa los valores estimados por el modelo. N representa el número total de datos en la serie.

El RMSE se utiliza como una medida de la discrepancia entre los valores observados y los valores estimados. Cuanto menor sea el valor del RMSE, más adecuado será el modelo de predicción, ya que indicará que las predicciones se acercan más a los valores reales.

Al utilizar el RMSE como métrica de evaluación, se busca obtener un modelo de predicción que minimice el error y se ajuste de manera precisa a los datos de accidentes de tránsito. Esto permitirá realizar predicciones más precisas y confiables, lo que a su vez facilitará la toma de decisiones informadas y la implementación de estrategias efectivas para prevenir accidentes de tránsito en el futuro.

3. Análisis de Datos

4.1. Selección de datos

En esta etapa del proyecto, se abordan los datos clave que revelan la problemática relacionada con los accidentes en el Reino Unido. En primer lugar, se realiza la lectura de un archivo CSV llamado '**UK_Accident.csv'.** A continuación, se seleccionan los datos correspondientes al año 2014 como base para analizar los principales factores que pueden influir en los accidentes en el país.

Posteriormente, se realiza una transformación de los datos de tipo cadena (string) a formato de fecha y hora reconocible por la biblioteca Pandas. Esto se logra utilizando la función **to_datetime**, que permite convertir los valores de cadena en una representación de fecha y hora adecuada.

Por último, se realiza un resumen de las principales estadísticas de interés. Esto incluye el recuento total de registros, la desviación estándar, los valores máximos y mínimos de las diferentes variables. Para obtener este resumen, se utiliza la función **describe(),** que proporciona un resumen estadístico de todas las columnas del DataFrame.

Estas acciones permiten obtener una visión general de los datos y establecer una base sólida para el análisis de los problemas más relevantes asociados a los accidentes de tránsito en el Reino Unido. Al comprender las estadísticas clave, se podrán identificar patrones, tendencias y posibles áreas problemáticas, lo que ayudará a orientar las estrategias de prevención y reducción de accidentes de manera efectiva.

3.2. Variable objetivo

En este proyecto, la variable objetivo que se desea predecir es **"Number_of_Casualties"** (Número de Víctimas). Esta variable nos proporciona una medida cualitativa de la problemática de los accidentes en el Reino Unido en el futuro. Además, es una variable crucial que nos brinda información sobre el impacto y la magnitud de los accidentes, lo que facilita la identificación de posibles soluciones y enfoques para abordar este problema.

Una vez establecida la variable objetivo, se procederá a realizar un análisis exhaustivo de los datos disponibles. Se evaluarán diversas variables relacionadas con los accidentes de tránsito, como las condiciones climáticas, el tipo de accidente, la ubicación geográfica, la hora del día, entre otras. Estas variables serán consideradas como posibles entradas para el entrenamiento de los algoritmos de predicción.

Durante el análisis, se examinará la relación entre cada variable y el número de víctimas para determinar su relevancia y capacidad predictiva. Se utilizarán técnicas estadísticas y visualizaciones para identificar patrones, correlaciones y posibles factores de riesgo asociados con los accidentes de tránsito.

Posteriormente, se tomará una decisión informada sobre qué variables serán seleccionadas como entradas para el entrenamiento de los algoritmos de predicción. Se buscará encontrar un conjunto de variables que mejor capturen la complejidad y los factores influyentes en la problemática de los accidentes de tránsito en el Reino Unido.

Este enfoque permitirá obtener un modelo de predicción robusto y efectivo, que utilizará las variables seleccionadas como entradas para realizar estimaciones futuras del número de víctimas en los accidentes de tránsito. Al comprender los factores más relevantes que influyen en la problemática, se podrán tomar decisiones informadas y desarrollar estrategias adecuadas para prevenir y reducir los accidentes de tránsito en el Reino Unido.

3.3. Análisis de la variable objetivo

Es de vital importancia describir y analizar el comportamiento de la variable objetivo en el proyecto, que en este caso es "Number_of_Casualties" (Número de Víctimas). Al examinar esta variable, se observa una alta asimetría hacia valores cercanos a 1. Dado que se trata de un valor entero, es necesario verificar la distribución de los datos y asegurarse de que no todos los registros tengan un valor de 1.

Para esto, se procede a analizar los valores únicos presentes en la variable objetivo. Si se confirma que hay una amplia variedad de valores diferentes a 1, se puede realizar una transformación logarítmica. Esta transformación ayuda a visualizar mejor los datos y reducir la asimetría, lo que facilita su interpretación y análisis.

La transformación logarítmica aplicada a la variable objetivo permite observar de manera más clara la distribución y los patrones en los datos. Además, ayuda a reducir la influencia de valores atípicos y proporciona una representación más equilibrada y adecuada para su análisis.

Al comprender el comportamiento de la variable objetivo y aplicar las transformaciones adecuadas, se podrán tomar decisiones informadas y desarrollar modelos de predicción más precisos y confiables. Esto permitirá abordar de manera efectiva la problemática de los accidentes de tránsito y trabajar en la implementación de estrategias preventivas adecuadas en el Reino Unido.

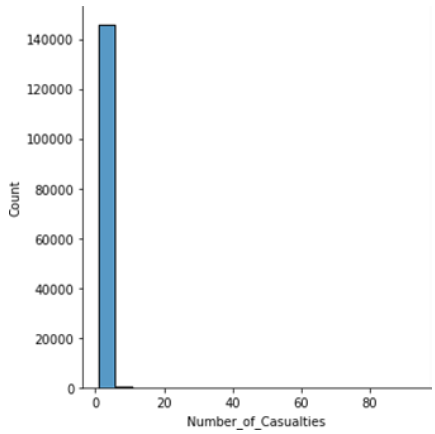


Figura 1: Distribución de la variable objetivo.

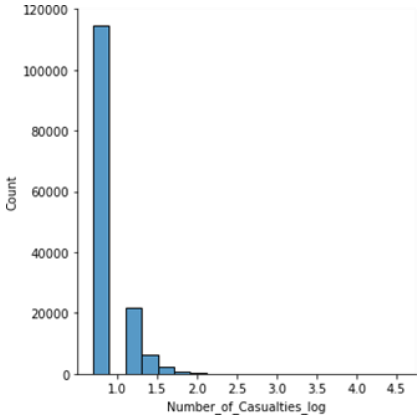


Figura 2: Transformación logarítmica.

En la Figura 2, se puede observar que la distribución de la variable objetivo, después de aplicar la transformación logarítmica, presenta un comportamiento mejorado y más adecuado para su análisis. Esta transformación ha permitido aprovechar más datos que antes se veían afectados por el sesgo presente en ciertos rangos de la gráfica de la Figura 1.

La modificación realizada a la variable objetivo, utilizando la transformación logarítmica, ha mejorado significativamente su distribución. Esto resulta en una mayor cantidad de datos valiosos y relevantes para el análisis y la implementación de pruebas de programación y procesos algorítmicos. Al utilizar la variable objetivo-modificada, se podrá realizar un análisis más preciso y efectivo, aprovechando todos los datos disponibles y evitando el sesgo que existía previamente. Esto permitirá obtener resultados más confiables y tomar decisiones informadas en relación con la problemática de los accidentes de tránsito en el Reino Unido.

4. Exploración de variables

La exploración de variables es un paso crucial para desarrollar el modelo, ya que nos proporciona una visión de cómo se relacionan con la variable objetivo. Esta exploración requiere tener una lista de variables previamente establecida para su análisis. Estas variables se importan y se utilizan para calcular datos estadísticos y crear histogramas que resultarán fundamentales para comprender y describir la problemática de los accidentes en el Reino Unido y sus implicaciones.

La exploración de variables nos permite examinar la relación entre cada variable y la variable objetivo. A través de medidas estadísticas como la media, la desviación estándar y la correlación, obtenemos información valiosa sobre cómo cada variable puede afectar los

accidentes y sus consecuencias. Además, al generar histogramas y visualizaciones, podemos identificar patrones, tendencias y posibles factores de riesgo asociados a los accidentes de tránsito en el Reino Unido.

La lista de variables importadas y analizadas en esta etapa de exploración nos proporciona una base sólida para comprender en profundidad la problemática de los accidentes y las implicaciones asociadas. Al examinar estas variables de manera integral, podremos tomar decisiones informadas y desarrollar estrategias efectivas para abordar la problemática de los accidentes en el Reino Unido.

5.1. 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:

Tras realizar un análisis de la condición de la vía, se ha observado que la mayoría de los accidentes se producen en condiciones de vía seca. Esta condición representa el escenario más frecuente en el que ocurren los accidentes en el dataset analizado. En segundo lugar, se identifica que la vía mojada o húmeda también es una condición en la que se producen un número considerable de accidentes.

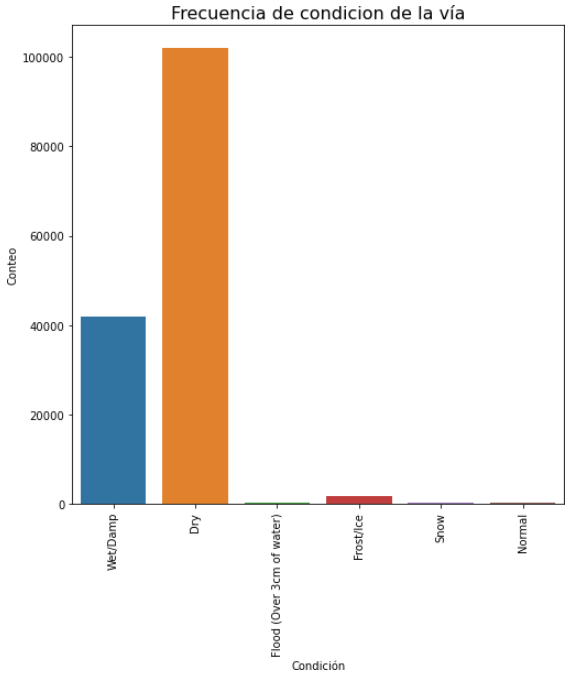
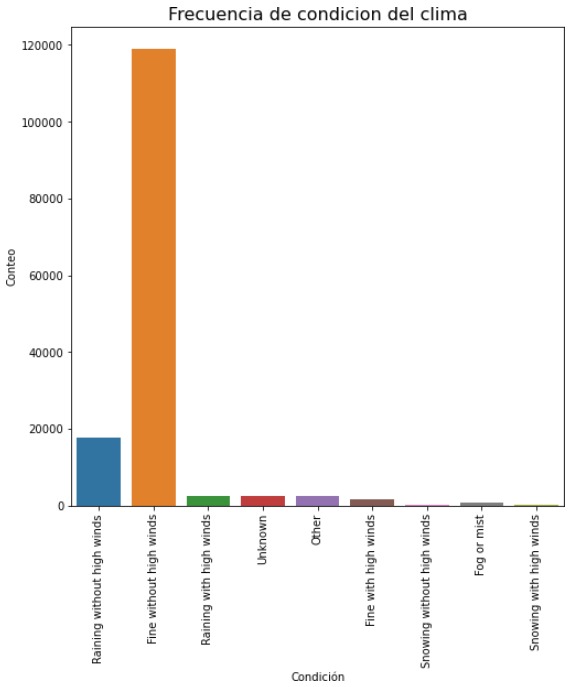


Figura 3: Frecuencia de condición de la vía.

Este hallazgo resalta la importancia de tomar medidas preventivas y de seguridad tanto en condiciones de vía seca como en condiciones de vía mojada o húmeda. La alta incidencia de accidentes en condiciones de vía seca puede estar relacionada con factores como el exceso de velocidad, el incumplimiento de las normas de tráfico y otros comportamientos imprudentes. Por otro lado, las condiciones de vía mojada o húmeda pueden aumentar el riesgo de deslizamientos y pérdida de control del vehículo debido a la disminución de la tracción.

Estos resultados enfatizan la necesidad de implementar medidas de seguridad vial en todas las condiciones de la vía, incluyendo la mejora del mantenimiento de las carreteras y la promoción de conductas responsables por parte de los conductores. Con esta información, las autoridades y los responsables de la seguridad vial pueden dirigir sus esfuerzos hacia la prevención de accidentes en las condiciones más comunes y reducir así los riesgos asociados a ellas.



La mayoría de los accidentes ocurren durante el día, cuando las luces de las calles no están iluminadas. Sin embargo, también se registra un porcentaje significativo de accidentes durante la noche, cuando las luces de la vía están presentes y encendidas, lo que se considera condiciones óptimas.

Ilustración 4. Frecuencia de las condiciones del clima.

De acuerdo con los datos, se observa que la mayoría de los accidentes ocurren en un porcentaje significativamente alto cuando los vehículos circulan a la velocidad permitida en áreas urbanas, que oscila entre 30 y 40 millas por hora. Es importante destacar que 30 millas por hora es la velocidad máxima permitida en áreas urbanas. Además, otro porcentaje de accidentes ocurren en rangos de velocidad entre 60 y 70 millas por hora, que corresponden a las velocidades máximas permitidas en autopistas principales de una y doble calzada, respectivamente.

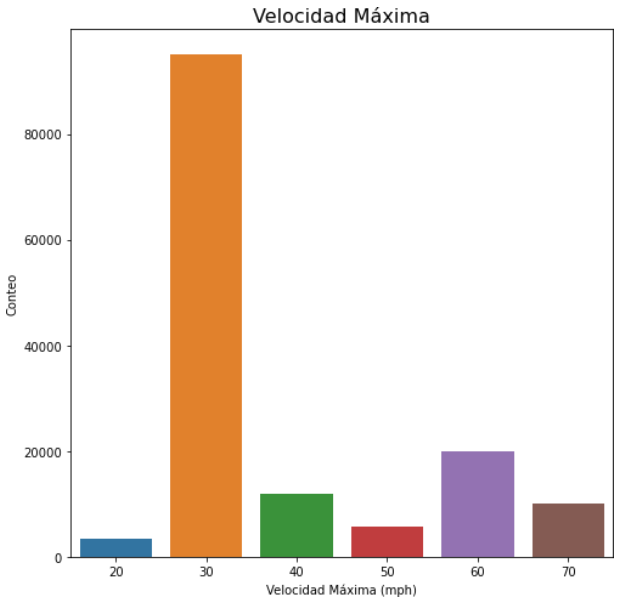
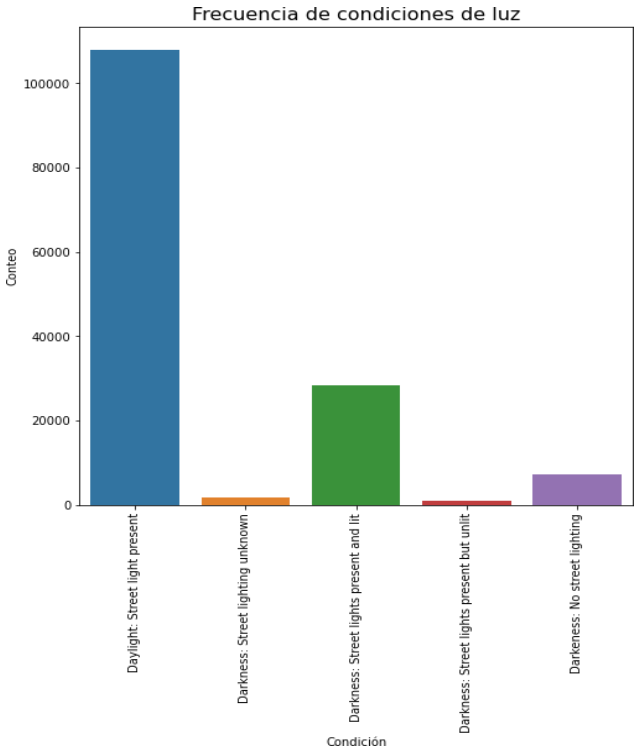
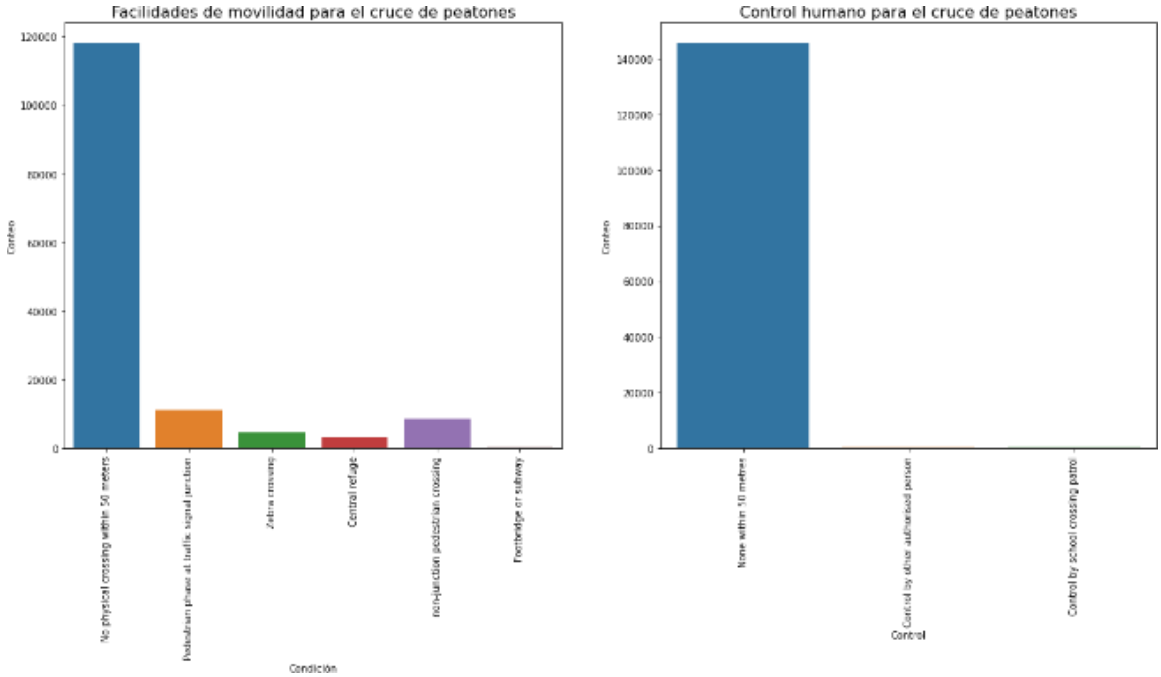


Ilustración 5. Velocidad máxima.



La mayoría de los accidentes ocurren durante el día, cuando las luces de las calles no están iluminadas. Sin embargo, también se registra un porcentaje considerable de accidentes durante la noche, cuando las luces de la vía están presentes y encendidas, lo que se considera condiciones óptimas de visibilidad.

Ilustración 6. Frecuencia de condiciones de luz.



El porcentaje de accidentes es notablemente alto cuando no hay facilidades de movilidad para que los peatones crucen en un radio de 50 metros. Además, se observan dos porcentajes significativos en las intersecciones de múltiples cruces: "pedestrian phase at traffic signal junction" (fase peatonal en intersección con señal de tráfico) y "non-junction pedestrian crossing" (cruce peatonal sin intersección). Estas situaciones representan puntos críticos donde se producen un número considerable de accidentes relacionados con la movilidad de los peatones.

Las cantidades de accidentes por día no varían significativamente, pero se destaca que el viernes es el día con mayor número de siniestros viales o accidentes peatonales. Por otro lado, el domingo es el día con la menor frecuencia de accidentes. Esta información resalta la importancia de estar especialmente alerta los viernes, mientras que los domingos se observa una menor incidencia de accidentes en comparación con los demás días de la semana.

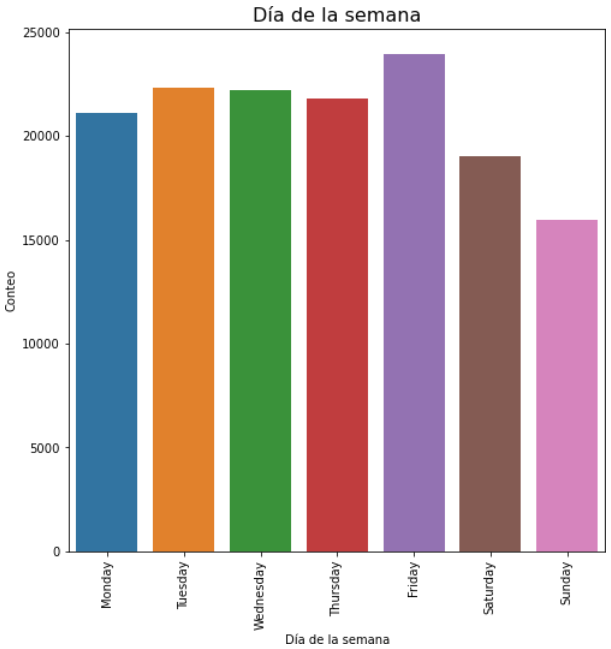
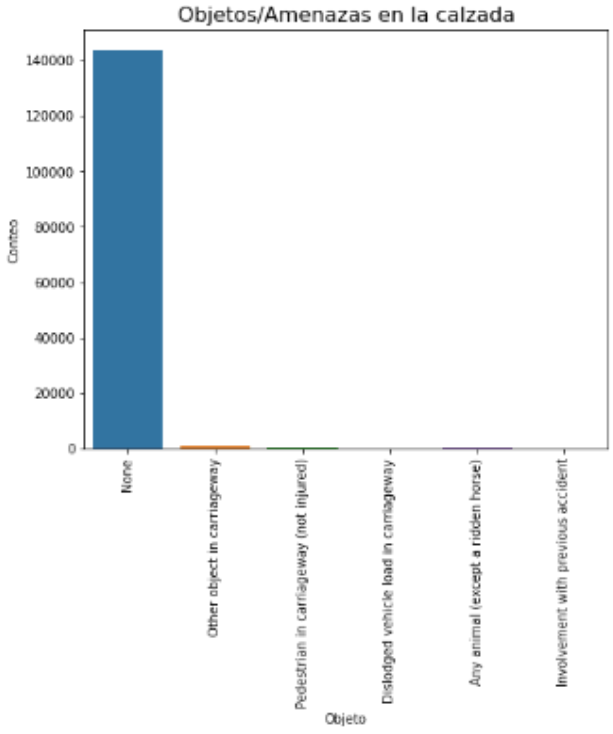
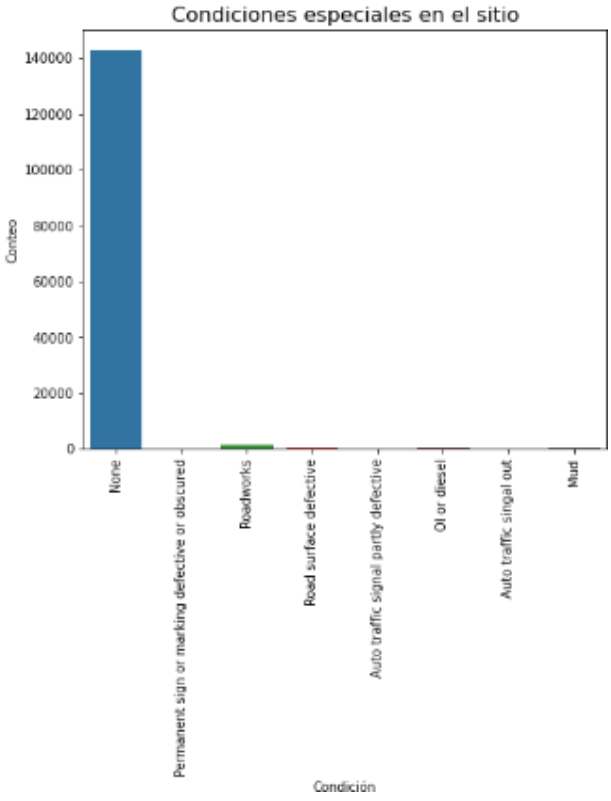
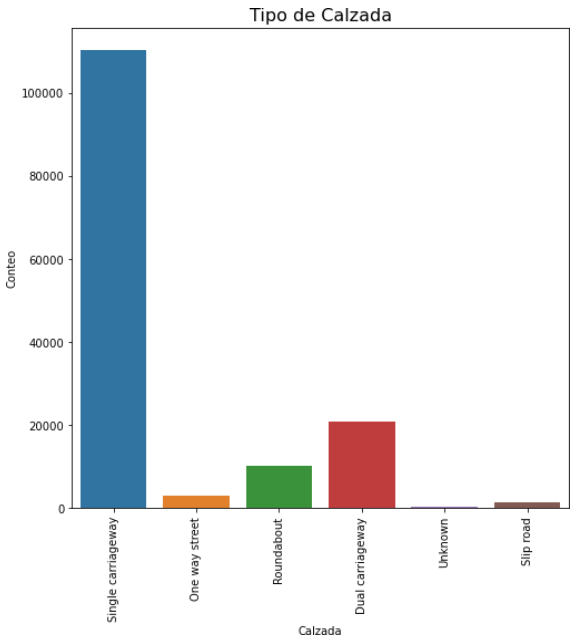


Ilustración 7. Frecuencia de los días





La mayoría de los accidentes ocurren sin la presencia de amenazas o condiciones especiales en la vía. Esto indica que la mayoría de los accidentes no están relacionados con factores externos o situaciones de riesgo específicas. Sin embargo, es importante destacar que existen otros factores y variables que pueden contribuir a la ocurrencia de los accidentes, como el comportamiento del conductor, el estado de la vía, las condiciones climáticas, entre otros. Es fundamental analizar en detalle estos factores adicionales para comprender mejor las causas de los accidentes y tomar medidas preventivas adecuadas.



La mayoría de los accidentes se producen en calzadas de un solo carril y en calzadas de doble carril. Esta observación tiene sentido ya que estas vías suelen tener una alta densidad de tráfico, con numerosos vehículos circulando y transitando en ellas. La mayor concentración de carros en estas vías aumenta la probabilidad de colisiones y otros tipos de accidentes. Es crucial tener en cuenta este patrón al diseñar estrategias de seguridad vial y medidas de prevención que se centren en la gestión del tráfico y la reducción de los riesgos asociados con estas vías de mayor circulación.

Ilustración 8. Tipo de calzada en la que se desplaza los carros.

4.2. Accidentes por hora y mes.

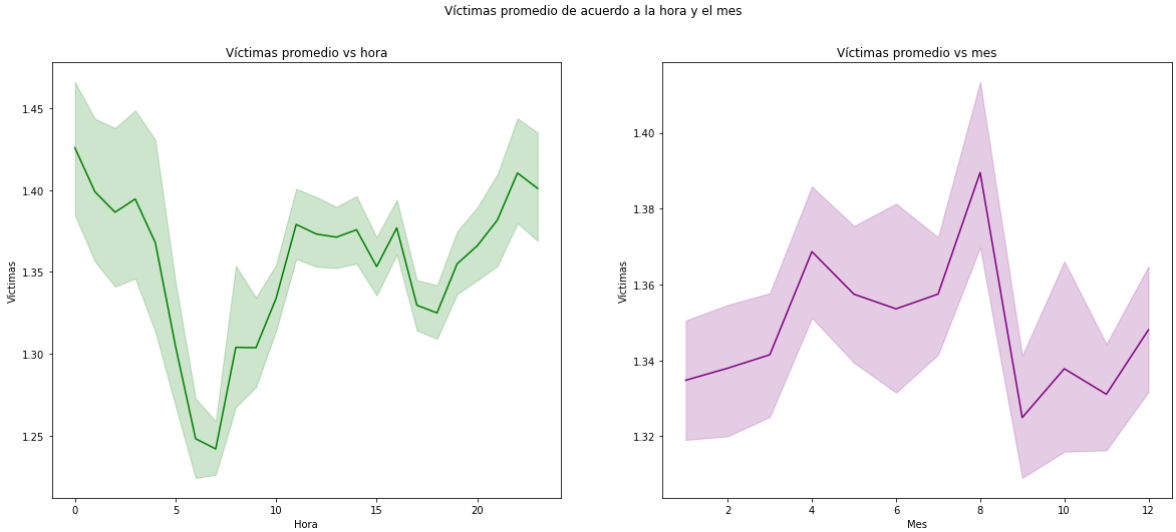


Ilustración 9. Correlación de los accidentes por hora y meses.

Al analizar el número de víctimas promedio por hora, se observa una disminución significativa entre las 5 y las 8 de la mañana. Esta reducción puede atribuirse a varios factores. En primer lugar, durante esas horas, muchas personas se encuentran en el inicio de sus jornadas laborales y no hay un flujo de tráfico tan intenso como en otros momentos del día. Además, es posible que las condiciones de iluminación y visibilidad sean mejores en comparación con las horas nocturnas, lo que podría contribuir a una disminución en los accidentes. Además, es probable que, durante esas horas, las personas estén más alertas y despiertas, lo que puede llevar a una conducción más segura y a una menor probabilidad de accidentes.

En cuanto al análisis del número de víctimas promedio por mes, se destaca un pico notable en el mes de agosto y un pequeño pico en el mes de abril. Estos patrones pueden estar influenciados por diversos factores. Por ejemplo, en el mes de agosto, es común que muchas personas estén de vacaciones o disfrutando del verano, lo que puede resultar en un aumento en el tráfico y, por lo tanto, en un mayor número de accidentes. Además, el clima puede ser un factor relevante, ya que en algunos lugares abril puede marcar el inicio de la primavera, lo que podría llevar a un aumento en la actividad y el movimiento en las vías, así como a condiciones climáticas variables que podrían contribuir a un incremento en los accidentes.

Estos patrones observados en los datos son importantes para comprender las tendencias y variaciones en la ocurrencia de accidentes a lo largo del día y del año. Esta información puede ser utilizada para implementar medidas preventivas y estrategias de seguridad vial específicas en momentos y lugares clave, con el objetivo de reducir la incidencia de accidentes y proteger la vida de las personas en las vías.

5.3. Identificación Área Urbana/ Rural.

En la columna "Urban_or_Rural_Area" del dataset, se observa que solo existen dos valores posibles: 1 y 2. Aunque el dataset no proporciona información explícita sobre qué valor corresponde a un área urbana o rural, es posible inferirlo utilizando otros datos disponibles, como la longitud y latitud, junto con el conocimiento del mapa del Reino Unido.

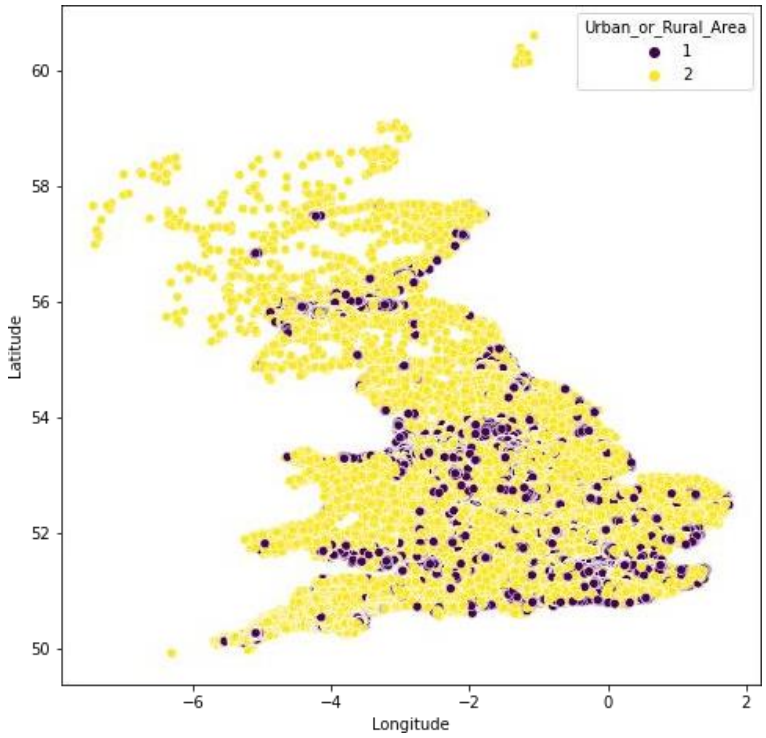


Ilustración 10. Identificación de las zonas rurales y

En la columna "Urban_or_Rural_Area" del dataset, se observa que solo existen dos valores posibles: 1 y 2. Aunque el dataset no proporciona información explícita sobre qué valor corresponde a un área urbana o rural, es posible inferirlo utilizando otros datos disponibles, como la longitud y latitud, junto con el conocimiento del mapa del Reino Unido.

A partir del análisis de la ubicación geográfica de los accidentes y teniendo en cuenta las características típicas de las áreas urbanas y rurales en el Reino Unido, se puede deducir que el valor 1 en la columna "Urban_or_Rural_Area" corresponde a áreas urbanas, mientras que el valor 2 corresponde a áreas rurales.

Esta deducción se basa en el entendimiento de que en áreas urbanas, como ciudades y zonas densamente pobladas, es más probable encontrar accidentes, y la presencia de infraestructuras viales más complejas y una mayor concentración de tráfico contribuyen a esta tendencia. Por otro lado, en áreas rurales, donde hay menos densidad de población y la infraestructura vial es menos compleja, se espera que los accidentes sean menos frecuentes.

Es importante tener en cuenta esta inferencia al analizar y realizar estudios basados en el dataset, ya que nos permite comprender mejor la distribución de los accidentes entre áreas urbanas y rurales y considerar las implicaciones específicas de cada tipo de área en relación con la seguridad vial y la implementación de medidas preventivas adecuadas.

5.4. 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.

5.5. Distribución de las variables numéricas.

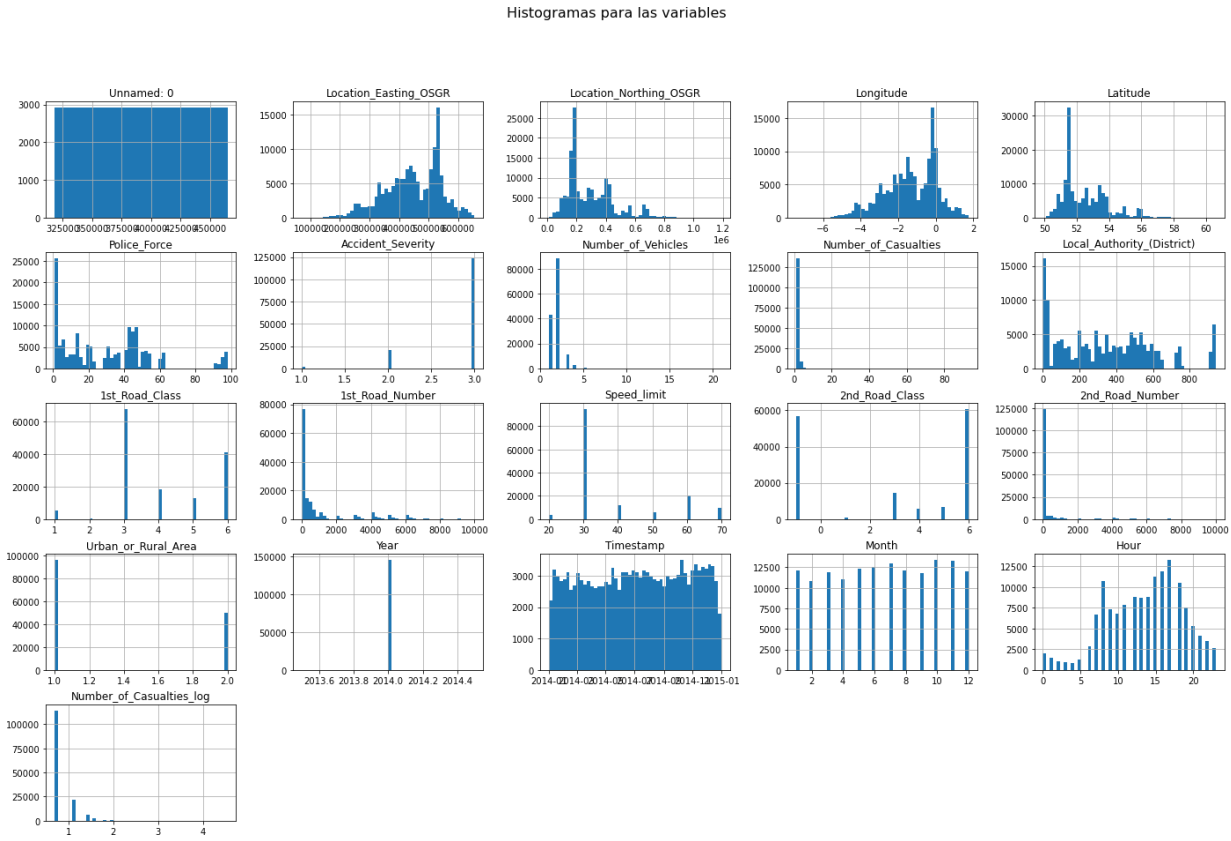


Ilustración 11. Distribución de las variables.

5. Simulación de datos faltantes.

Para cumplir con los requisitos del proyecto, es necesario que el dataset contenga al menos un 5% de datos faltantes en al menos tres columnas. Actualmente, el dataset presenta datos faltantes en una columna, que es LSOA_of_Accident_Location. Con el fin de simular la falta de datos en dos columnas adicionales, se han seleccionado Road_Type, Police_Force y Number_of_Vehicles. De esta manera, los datos faltantes se distribuyen 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

6. Tratamiento de datos.

Al analizar la columna LSOA_of_Accident_Location, se observa que alrededor del 6% de los datos están faltantes. Estos datos corresponden a una notación única para cada zona del Reino Unido, pero su utilidad es limitada y rellenar los datos faltantes sería una tarea tediosa. Por lo tanto, se recomienda eliminar esta columna del dataset.

En cuanto a las columnas Police_Force y Number_of_Vehicles, es posible analizar la distribución de los datos disponibles. Al observar la distribución y calcular la mediana y la moda, se puede proceder a rellenar los datos faltantes utilizando la moda, que representa el valor más frecuente en los datos existentes.

Para la columna Road_Type, donde hay datos faltantes, se sugiere agrupar todos los valores faltantes bajo la categoría 'Unknown'. Esto permitirá identificar claramente los casos donde no se dispone de información precisa sobre el tipo de vía en el momento del accidente.

De esta manera, se aborda la gestión de los datos faltantes de manera eficiente y se mantiene la integridad del dataset para su posterior análisis.

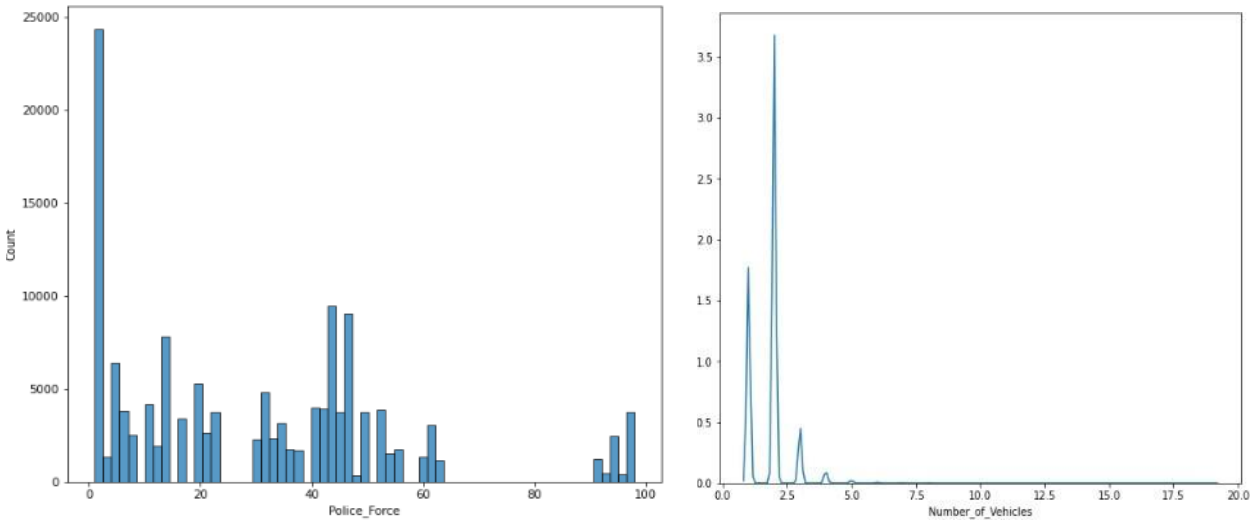


Ilustración 12. Graficas del número de policías utilizados y el número de vehículos utilizados

7.1. Eliminación de variables no relevantes.

Se recomienda eliminar las siguientes variables del dataset debido a que no aportan información relevante o presentan duplicación de información:

• Location_Easting_OSGR y Location_Northing_OSGR: Estos campos son información duplicada de las columnas Longitude y Latitude, por lo tanto, se pueden eliminar para evitar redundancia.
• Year: Todos los datos del dataset corresponden al año 2014, por lo tanto, la variable "Year" no aporta variabilidad y puede ser eliminada.
• Number_of_Casualties_log: Esta columna fue creada con el propósito de visualización de datos, pero no es esencial para el análisis y modelado. Por lo tanto, se puede eliminar sin afectar la información principal.
• Unnamed: 0 y Accident_Index: Estos campos son identificadores únicos de cada accidente, pero no aportan información relevante para el análisis. Por lo tanto, se pueden eliminar sin perder información valiosa.
• Carriageway_Hazards y Special_Conditions_at_Site: Estas variables contienen principalmente datos nulos, lo que indica que la información disponible es limitada. Por lo tanto, se sugiere eliminar estas variables para mantener un dataset más limpio y con información más completa.

• Pedestrian_Crossing-Physical_Facilities y Pedestrian_Crossing-Human_Control: La mayoría de los accidentes no presentan facilidades de movilidad o control peatonal en un radio de 50 metros. Por lo tanto, estas variables no aportan información significativa y pueden ser eliminadas. Al eliminar estas variables, se simplifica el dataset y se concentra en aquellas variables que son más relevantes para el análisis de los accidentes de tránsito en Reino Unido durante el año 2014.

7.2. 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.

8. 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: 0.1012485556363758
 En train: 0.10230442207776677
```

9. Métodos no supervisados

Para los métodos no supervisados se procedió a realizar un PCA, el cual es una función que permite obtener los datos más 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:

```python
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')
```

10. 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:

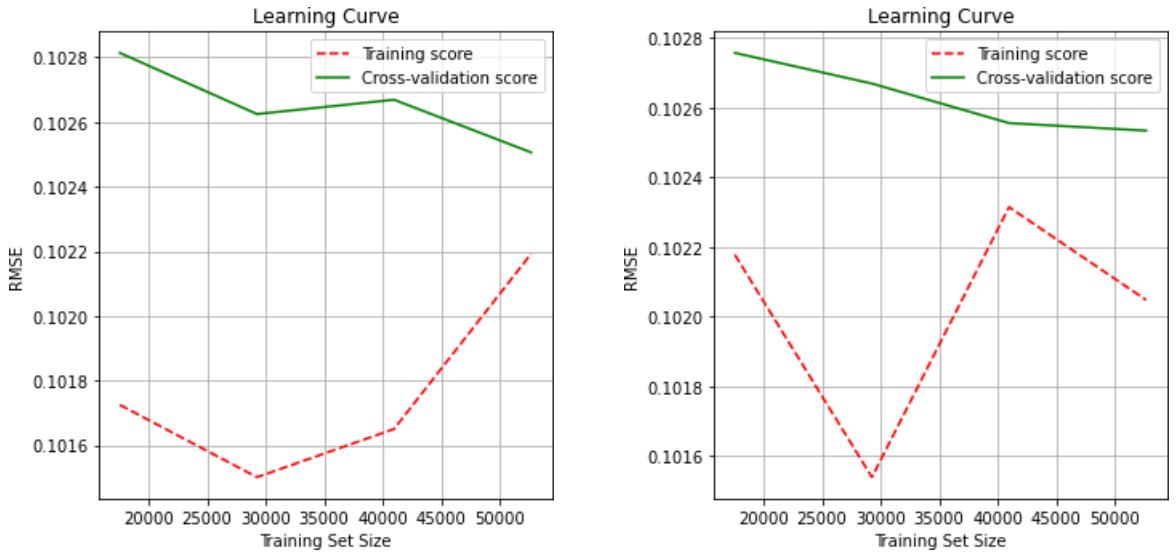


Figura 13: Curvas de aprendizaje, (De derecha a izquierda: Decision tree, Decision tree + PCA)

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.

11. Retos y condiciones de despliegue del modelo.

El modelo de predicción de víctimas en accidentes automovilísticos enfrenta diversos desafíos, especialmente en términos de recolección de datos significativos. Aunque el dataset actual contiene una gran cantidad de información sobre los accidentes, no logra proporcionar un desempeño óptimo para la predicción. Esto plantea la necesidad de considerar la adición de nuevas columnas al dataset, lo cual puede implicar costos adicionales.

Si se busca desarrollar un modelo adecuado para su implementación en producción, es necesario abordar los siguientes desafíos:

Recolección de datos adicionales: Se requeriría recopilar más información sobre los accidentes para enriquecer el dataset. Esto podría incluir variables como condiciones climáticas, estado de la vía, presencia de señales de tránsito, entre otros. Sin embargo, esta tarea implica consideraciones logísticas y costos asociados.

Evaluación con los profesionales de la salud y servicios de emergencia: Sería necesario establecer una colaboración activa con centros de salud, ambulancias y paramédicos para evaluar la viabilidad del modelo. Se debe determinar si la implementación del modelo proporcionaría una mejora significativa en la eficiencia de la atención a los accidentes automovilísticos.

12. Conclusiones.

Es altamente recomendable obtener más datos representativos para el dataset con el fin de mejorar su rendimiento y reducir el sesgo presente en los datos actuales. Obtener datos adicionales permitirá tener una muestra más completa y diversa, lo que contribuirá a que los modelos puedan capturar mejor la variabilidad de los accidentes de tránsito. Además, al tener más datos, se tendrán más ejemplos para el entrenamiento de los modelos, lo que aumentará su capacidad de generalización.

En cuanto a la selección del modelo, es importante tener en cuenta que los resultados pueden ser muy similares entre los tres modelos evaluados inicialmente. Sin embargo, es recomendable explorar y analizar otros modelos disponibles en el campo de la predicción de accidentes de tránsito. Cada modelo tiene sus propias fortalezas y debilidades, y probar diferentes enfoques permitirá tener una visión más completa y robusta de las posibles soluciones.

El sesgo en los datos también puede estar influenciado por la propia naturaleza de los accidentes de tránsito. Si existe una acumulación significativa de valores cercanos a 1 en los datos, esto puede plantear desafíos para que los modelos "aprendan" de manera efectiva. Es posible que los modelos tengan dificultades para capturar y generalizar patrones en los datos debido a esta falta de variabilidad. Por lo tanto, es fundamental considerar técnicas de preprocesamiento y manejo de datos que aborden este sesgo, como la transformación logarítmica mencionada anteriormente, con el fin de mejorar la capacidad de los modelos para aprender y hacer predicciones más precisas.

13. Referencias.

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