INTRODUCCIÓN A LA INTELIGENCIA ARTIFICIAL

FINAL DE PROYECTO

ROAD-ACCIDENT-UK

POR:

ANA MARIA ROMERO CARVAJAL

JEAN CARLOS JULIO RODRIGUEZ

PROFESOR:

RAUL RAMOS POLLAN.

UNIVERSIDAD DE ANTIOQUIA.

FACULTAD DE INGENIERÍA.

MEDELLÍN.

2023.

1. INTRODUCCION

En cualquier rincón del planeta, los accidentes de tránsito constituyen una problemática seria que afecta a un elevado número de individuos anualmente. Estos eventos, derivados de la imprudencia o confusiones por parte de los conductores, representan una inquietud significativa a nivel global, tal como indican las estadísticas. Más allá del desafío inherente a la gestión del flujo vehicular, las naciones se ven confrontadas con las consecuencias sanitarias para los afectados.

En este contexto, la gestión eficiente de los datos adquiere una relevancia fundamental, particularmente en situaciones de riesgo para la seguridad pública, como lo son los accidentes automovilísticos. A través de la implementación de un conjunto de datos exhaustivo, se busca anticipar la probabilidad de accidentes considerando sus causas y efectos. Este enfoque tiene como objetivo dirigir los esfuerzos hacia la prevención y reducción de la tasa de accidentes, proporcionando información vital a entidades interesadas, como hospitales, quienes pueden beneficiarse al estar preparados ante un eventual incremento en la demanda de servicios médicos debido a accidentes de tráfico.

Para alcanzar esta meta, la inteligencia artificial se establece como una herramienta inestimable. Gracias a sus aptitudes para analizar datos, la inteligencia artificial puede extraer información relevante de las bases de datos existentes, generando resultados deseados. Se anticipa obtener una comprensión profunda de las causas subyacentes de los accidentes y las estrategias preventivas más eficaces para disminuir su incidencia. La fusión de la inteligencia artificial y el análisis de datos promete constituir una alianza poderosa en la contención de los accidentes de tránsito y la salvaguarda de la seguridad vial.

2. CONJUNTO DE DATOS

El conjunto de datos utilizado en este proyecto, obtenido de Kaggle, se compone de un archivo CSV que detalla accidentes de tráfico en el Reino Unido entre 2005 y 2014, recopilados por el gobierno británico. Aunque la base de datos completa abarca más de 1.8 millones de registros, se ha optado por enfocarse exclusivamente en los datos más recientes del año 2014, totalizando 146,322 accidentes. Este conjunto de datos ofrece información detallada sobre cada accidente, con atributos que abarcan diversos aspectos relevantes para el análisis.

- Ubicación geográfica: Latitud y longitud del lugar exacto donde ocurrió el accidente.
- Fecha y hora: Información sobre la fecha y hora en la que se produjo el accidente.
- Tipo de accidente: Clasificación del tipo de accidente, como colisión de vehículos, atropello, choque con objeto fijo, entre otros.
- Condiciones climáticas: Descripción de las condiciones climáticas en el momento del accidente.
- Estado de la carretera: Información sobre el estado de la carretera, como seca, mojada, helada, etc.

- Factores contribuyentes: Factores que se consideran contribuyentes al accidente, como la velocidad, el consumo de alcohol, el uso del cinturón de seguridad, entre otros.
- 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_CrossingPhysical_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 parala locación geográfica de longitud y latitud
Year	Año del evento accidental

3. MÉTRICA

El Error Cuadrático Medio (RMSE) desempeña un papel central como métrica principal en la evaluación del modelo de predicción de accidentes de tránsito. Su cálculo implica la determinación de la raíz cuadrada del promedio de la suma de las diferencias al cuadrado

entre los valores observados en la serie (y_i) y los valores estimados por el modelo ($\hat{y}_i$), donde N representa el número total de datos en la serie.

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

El RMSE sirve como indicador cuantitativo de la discrepancia entre los valores reales y los proyectados, siendo un criterio de evaluación fundamental. Un menor valor de RMSE refleja una mayor idoneidad del modelo de predicción, indicando una mayor proximidad entre las predicciones y los datos reales. La elección del RMSE como métrica de evaluación subraya la búsqueda de un modelo que minimice el error, permitiendo predicciones más precisas y confiables. Esta precisión en la predicción facilita la toma de decisiones informadas y la implementación eficaz de estrategias orientadas a prevenir accidentes de tránsito en el futuro.

4. ANALISIS DE DATOS

4.1. SELECCIÓN DE DATOS

En esta fase del proyecto, se aborda la problemática de los accidentes en el Reino Unido mediante la lectura del archivo CSV “UK_Accident.csv” y la selección exclusiva de los datos correspondientes al año 2014. Posteriormente, se lleva a cabo una transformación de los datos de tipo cadena a formato de fecha y hora utilizando la función “to_datetime” de la biblioteca Pandas.

La última etapa de este proceso implica la generación de un resumen estadístico clave. Este resumen incluye el recuento total de registros, la desviación estándar, así como los valores máximos y mínimos de las distintas variables presentes en el conjunto de datos. Para realizar esta síntesis estadística, se emplea la función “describe()”, que proporciona una visión general de todas las columnas del DataFrame.

Estas acciones proporcionan una comprensión profunda de los datos y establecen una base sólida para el análisis de los problemas asociados a los accidentes de tránsito en el Reino Unido. Al identificar patrones, tendencias y áreas problemáticas potenciales a través de estadísticas clave, se facilita la formulación de estrategias efectivas para la prevención y reducción de accidentes en el futuro.

4.2. VARIABLE OBJETO

La variable objetivo para predicción es "Number_of_Casualties" (Número de Víctimas), que ofrece una medida cualitativa de la gravedad de los accidentes en el Reino Unido en el futuro. Dada su importancia, esta variable proporciona información clave sobre el impacto y la magnitud de los accidentes, facilitando la identificación de soluciones y enfoques para abordar el problema.

Con la variable objetivo establecida, se llevará a cabo un análisis exhaustivo de los datos disponibles, evaluando diversas variables relacionadas con los accidentes de tránsito, como condiciones climáticas, tipo de accidente, ubicación geográfica y hora del día, entre otras. Estas variables se considerarán como posibles entradas para el entrenamiento de algoritmos de predicción. Durante el análisis, se examinará la relación entre cada variable

y el número de víctimas, utilizando técnicas estadísticas y visualizaciones para identificar patrones, correlaciones y posibles factores de riesgo asociados.

Posteriormente, se tomará una decisión informada sobre qué variables seleccionar como entradas para el entrenamiento de los algoritmos de predicción, buscando un conjunto que capture de manera óptima la complejidad y los factores influyentes en los accidentes de tránsito en el Reino Unido. Este enfoque permitirá desarrollar un modelo de predicción robusto y efectivo, utilizando las variables seleccionadas para realizar estimaciones futuras del número de víctimas en accidentes de tránsito. Al comprender los factores más relevantes, se podrán tomar decisiones informadas y desarrollar estrategias adecuadas para prevenir y reducir los accidentes de tránsito en el Reino Unido.

4.3. ANALISIS DE VARIABLE

La descripción y análisis del comportamiento de la variable objetivo en este proyecto, "Number_of_Casualties" (Número de Víctimas), revela una marcada asimetría hacia valores cercanos a 1. Dado que esta variable toma valores enteros, es esencial examinar la distribución de los datos para garantizar que no todos los registros tengan un valor de 1.

Con el objetivo de abordar esta situación, se realiza un análisis de los valores únicos presentes en la variable objetivo. Si se verifica una diversidad significativa de valores distintos a 1, se considera la aplicación de una transformación logarítmica. Esta transformación, al reducir la asimetría, mejora la visualización de los datos y facilita su interpretación y análisis.

La transformación logarítmica aplicada a la variable objetivo proporciona una representación más clara de la distribución y patrones en los datos, al mismo tiempo que mitiga la influencia de valores atípicos, generando una representación más equilibrada y adecuada para el análisis. Al comprender el comportamiento de la variable objetivo y aplicar transformaciones pertinentes, se facilita la toma de decisiones informadas y el desarrollo de modelos de predicción más precisos. Esto, a su vez, posibilita abordar de manera efectiva la problemática de los accidentes de tránsito y trabajar en la implementación de estrategias preventivas apropiadas en el Reino Unido.

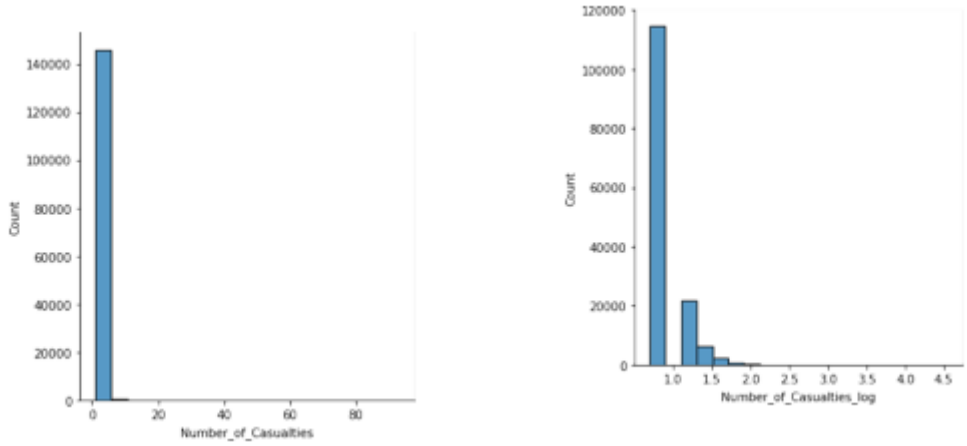


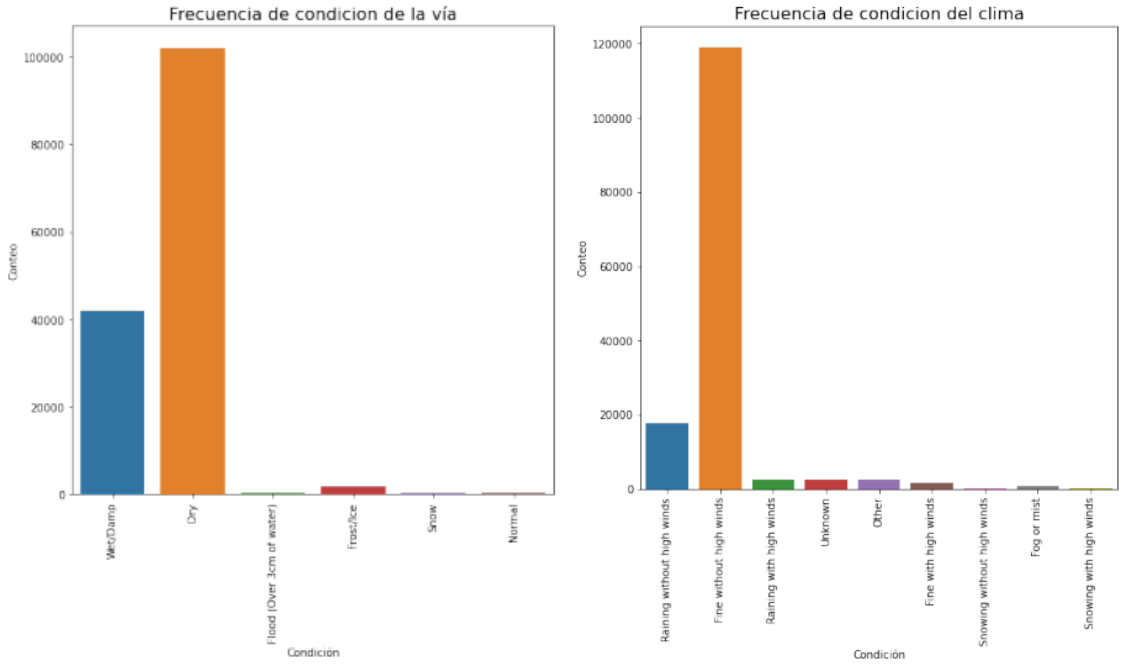
Figura 1. distribución y transformación logarítmica de la variable objetivo

En la Figura 1, se aprecia que la distribución de la variable objetivo, tras la aplicación de la transformación logarítmica, exhibe un comportamiento mejorado y más propicio para el análisis. Esta transformación ha optimizado la utilización de datos previamente afectados por sesgos presentes en ciertos rangos de la gráfica de la figura 1(derecha).

La modificación mediante la transformación logarítmica ha generado una mejora significativa en la distribución de la variable objetivo. Este ajuste se traduce en una mayor cantidad de datos valiosos y pertinentes para análisis, pruebas de programación y procesos algorítmicos. Al emplear la variable objetivo modificada, se posibilita un análisis más preciso y eficiente, aprovechando todos los datos disponibles y eliminando el sesgo previo. Esto resultará en resultados más confiables y en la capacidad de tomar decisiones informadas con respecto a la problemática de los accidentes de tránsito en el Reino Unido.

5. EXPLORACION DE VARIABLES

La exploración de variables constituye un paso esencial en el desarrollo del modelo, brindándonos una perspectiva de cómo estas variables se relacionan con la variable objetivo. Para llevar a cabo esta exploración de manera efectiva, se requiere una lista predefinida de variables que se importan y utilizan para calcular datos estadísticos, así como para crear histogramas que resultan fundamentales en la comprensión y descripción de la problemática de los accidentes en el Reino Unido.



Este proceso nos habilita para examinar la relación entre cada variable y la variable objetivo. A través de medidas estadísticas como la media, la desviación estándar y la correlación, se obtiene información valiosa acerca de cómo cada variable puede influir en los accidentes y sus consecuencias. Asimismo, al generar histogramas y visualizaciones, se pueden 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 proporciona una base sólida para comprender en profundidad la problemática de los accidentes y sus implicaciones. Al examinar integralmente estas variables, se facilita la toma de decisiones informadas y el desarrollo de estrategias efectivas para abordar la problemática de los accidentes en el Reino Unido.

5.1. GRAFICOS E HISTOGRAMAS

Después de la definición de las variables a utilizar, se avanza a la generación de gráficas que permitirán visualizar las cifras asociadas a las condiciones que influyen en los accidentes. Estas representaciones visuales serán fundamentales para comprender de manera más intuitiva y detallada la relación entre las diferentes variables y la incidencia de los accidentes.

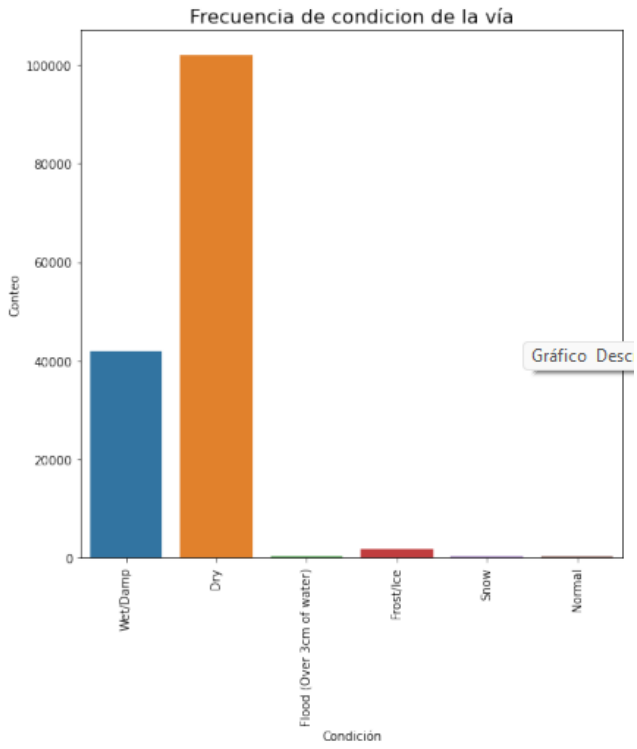


Figura 2. condicion de via

Luego del análisis detallado de las condiciones de la vía, se ha observado que la mayoría de los accidentes ocurren en condiciones de vía seca, siendo este el escenario más frecuente en el conjunto de datos analizado. En segundo lugar, se destaca que la vía mojada o húmeda también representa una condición en la cual se producen un número considerable de accidentes.

Este hallazgo resalta la importancia crítica de implementar medidas preventivas y de seguridad tanto en condiciones de vía seca como en condiciones de vía mojada o húmeda. La elevada incidencia de accidentes en condiciones de vía seca podría vincularse con factores como el exceso de velocidad, el incumplimiento de 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 subrayan la necesidad de implementar medidas de seguridad vial en todas las condiciones de la vía, lo que incluye mejoras en el 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, reduciendo así los riesgos asociados a ellas.

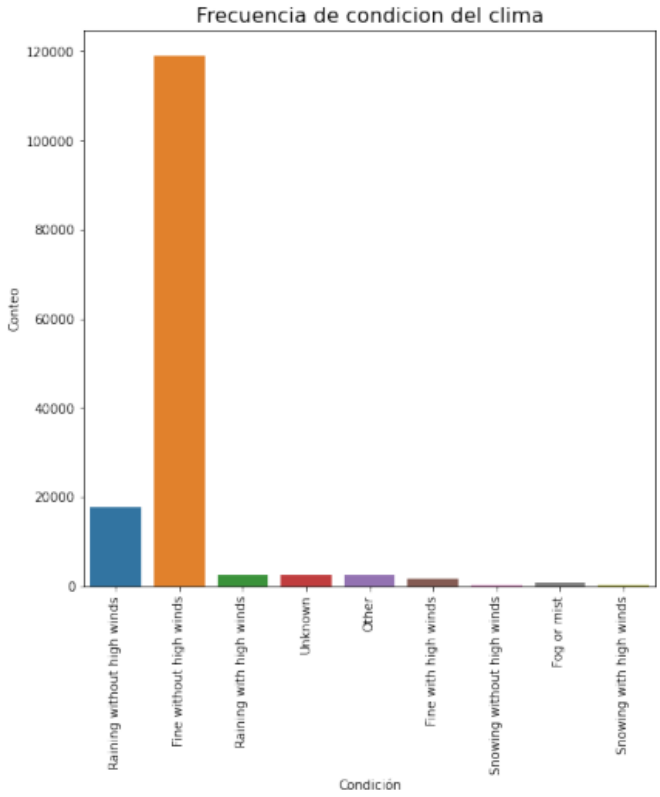


Figura 3. frecuencia de las condiciones del clima

Es notable que la mayoría de los accidentes ocurren durante el día, cuando las luces de las calles no están iluminadas. No obstante, también se registra un porcentaje significativo de accidentes durante la noche, incluso cuando las luces de la vía están presentes y encendidas, condiciones que se consideran óptimas. Este hallazgo sugiere que la visibilidad por sí sola no garantiza la ausencia de accidentes, y otros factores pueden contribuir a la ocurrencia de eventos no deseados en condiciones de iluminación adecuada.

Según 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 crucial destacar que 30 millas por hora es la velocidad máxima permitida en áreas urbanas. Además, se registra otro porcentaje de accidentes 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. Este patrón resalta la importancia de abordar la seguridad vial en áreas urbanas y en autopistas,

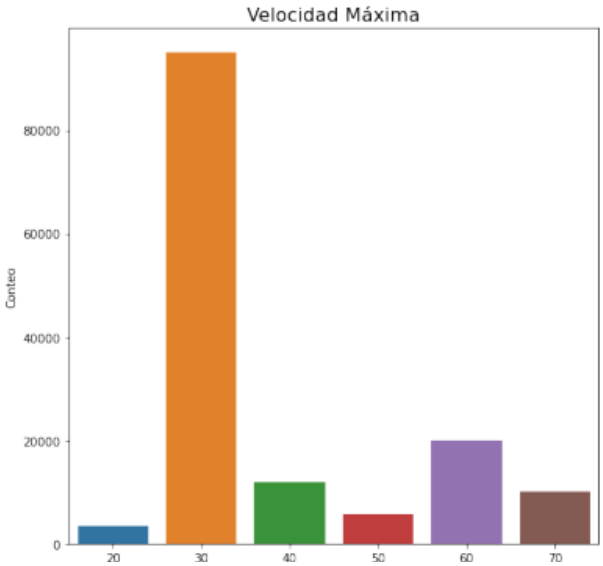


Figura 4. velocidad máxima

ajustando estrategias y medidas preventivas para adaptarse a las velocidades predominantes en cada entorno.

Es notorio que la mayoría de los accidentes tienen lugar durante el día, en situaciones en las cuales las luces de las calles no están iluminadas. No obstante, se observa también un porcentaje considerable de accidentes durante la noche, incluso cuando las luces de la vía están presentes y encendidas, condiciones que generalmente se consideran óptimas para la visibilidad. Este hallazgo sugiere que la mera presencia de iluminación no garantiza la ausencia de accidentes, y otros factores pueden influir en la seguridad vial durante las horas nocturnas.

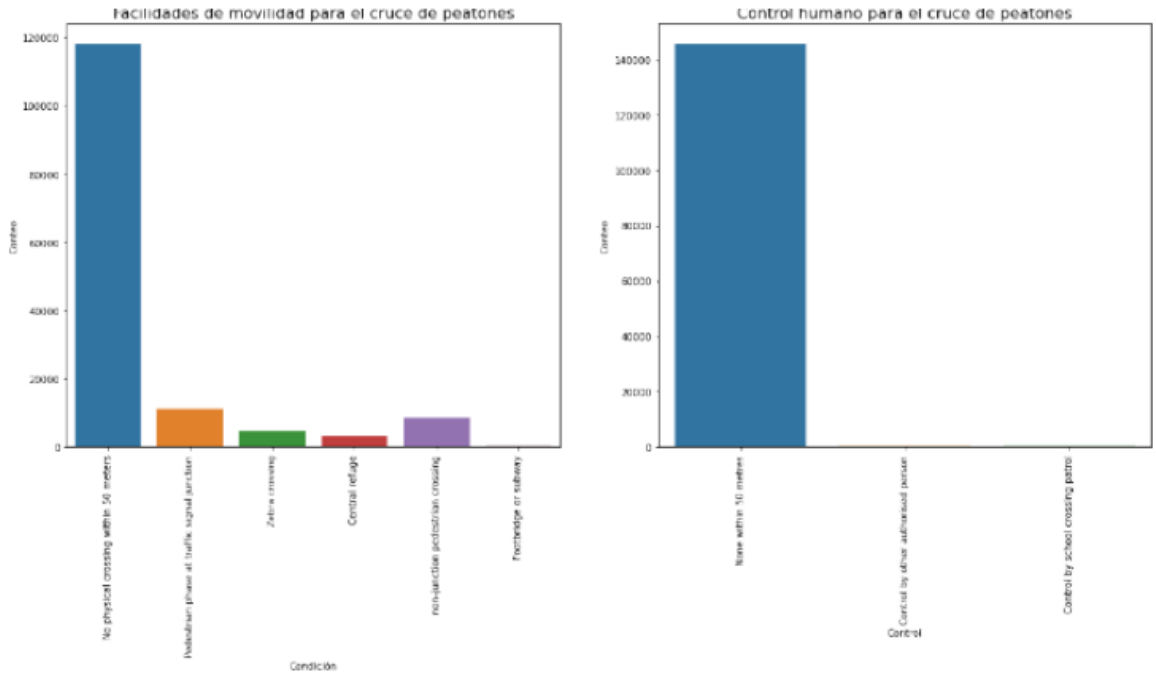


Figura 5. frecuencia de las condiciones de luz

El porcentaje de accidentes es notablemente alto en áreas donde no existen facilidades de movilidad para que los peatones crucen en un radio de 50 metros. Además, se destacan 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 produce un número considerable de accidentes relacionados con la movilidad de los peatones.

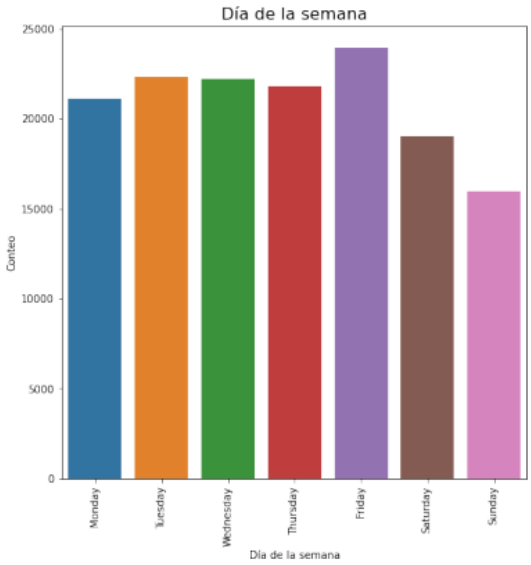
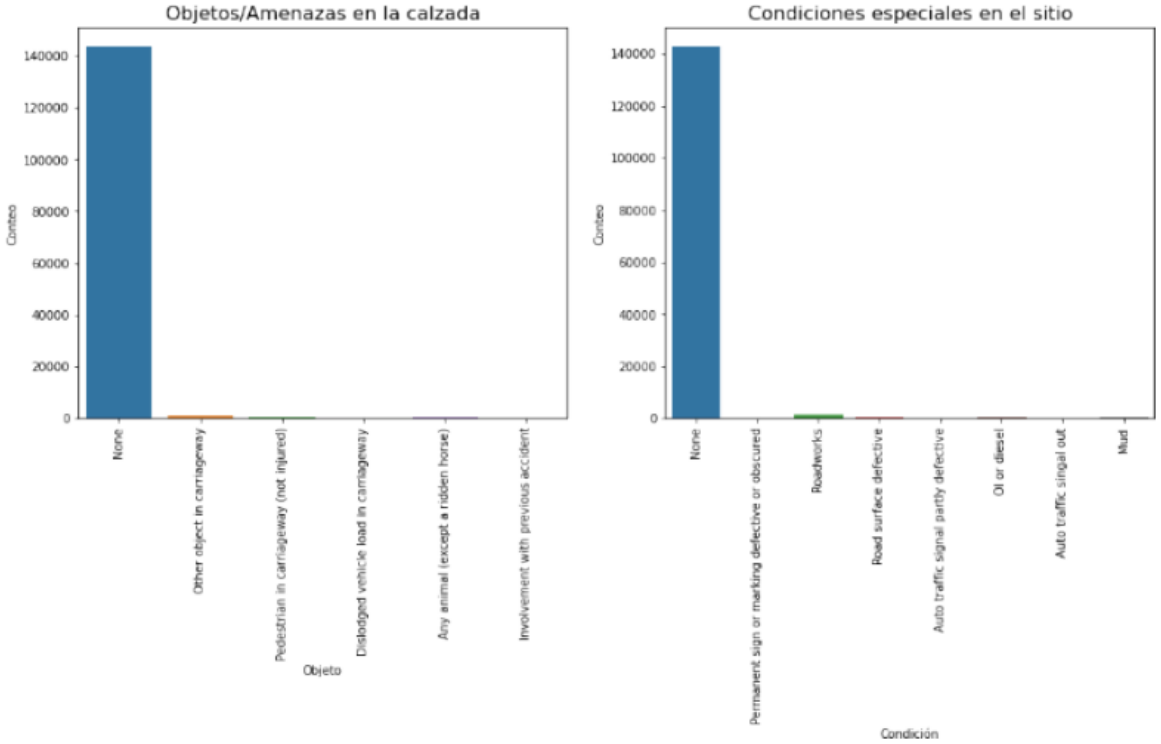


Figura 6. frecuencia de los días

En relación con la frecuencia de accidentes por día, aunque las cantidades diarias no varían significativamente, es importante señalar que los viernes presentan el mayor número de siniestros viales o accidentes peatonales. Contrariamente, los domingos muestran la menor frecuencia de accidentes. Esta información resalta la importancia de estar especialmente alerta los viernes, mientras que los domingos exhiben una menor incidencia de accidentes en comparación con los demás días de la semana. Estos patrones temporales son esenciales para orientar estrategias de seguridad vial y la asignación de recursos preventivos de manera efectiva.

La mayoría de los accidentes ocurren sin la presencia de amenazas o condiciones especiales en la vía, indicando que la mayoría de los siniestros no están relacionados con factores externos o situaciones de riesgo específicas. No obstante, es importante destacar que existen otros factores y variables que pueden contribuir a la ocurrencia de 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.



Adicionalmente, 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 vehículos 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.

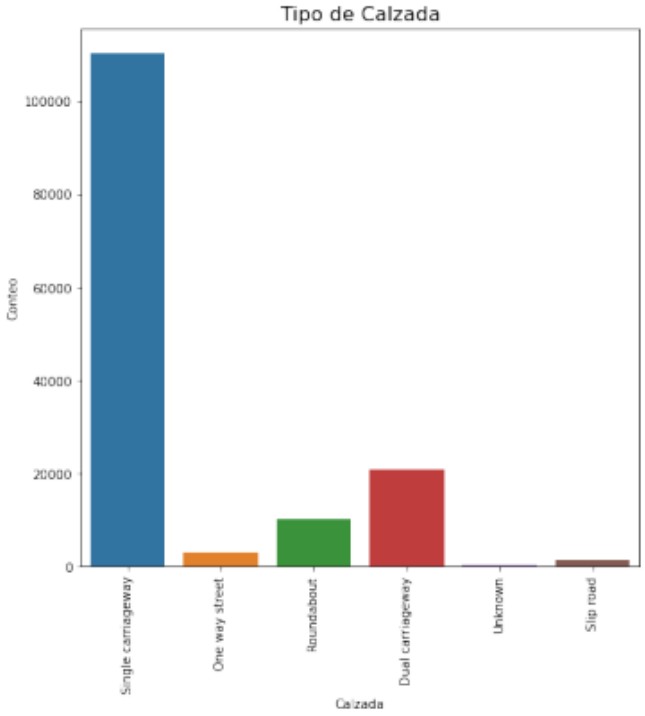


Figura 7. Tipo de calzada vehicular.

5.2. ACCIDENTES DE POR HORA Y MES

Al analizar el número promedio de víctimas por hora, se destaca una significativa disminución entre las 5 y las 8 de la mañana. Esta reducción puede atribuirse a varios factores. En primer lugar, durante esas horas, muchas personas están en el inicio de sus jornadas laborales, lo que se traduce en un tráfico menos intenso en comparación con otros momentos del día. Además, es plausible que las condiciones de iluminación y visibilidad sean mejores en comparación con las horas nocturnas, contribuyendo así a una disminución en los accidentes. Asimismo, es probable que, durante estas horas matutinas, las personas estén más alertas y despiertas, favoreciendo una conducción más segura y una menor probabilidad de accidentes.

En relación con el análisis del número promedio de víctimas 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 agosto, es común que muchas personas estén de vacaciones o disfrutando del verano, lo que podría resultar en un aumento del tráfico y, por ende, en un mayor número de accidentes. Además, las condiciones climáticas pueden desempeñar un papel relevante, ya que en algunos lugares abril puede marcar el inicio de la primavera, generando un aumento en la actividad y el movimiento en las vías, así como condiciones climáticas variables que podrían contribuir a un incremento en los accidentes.

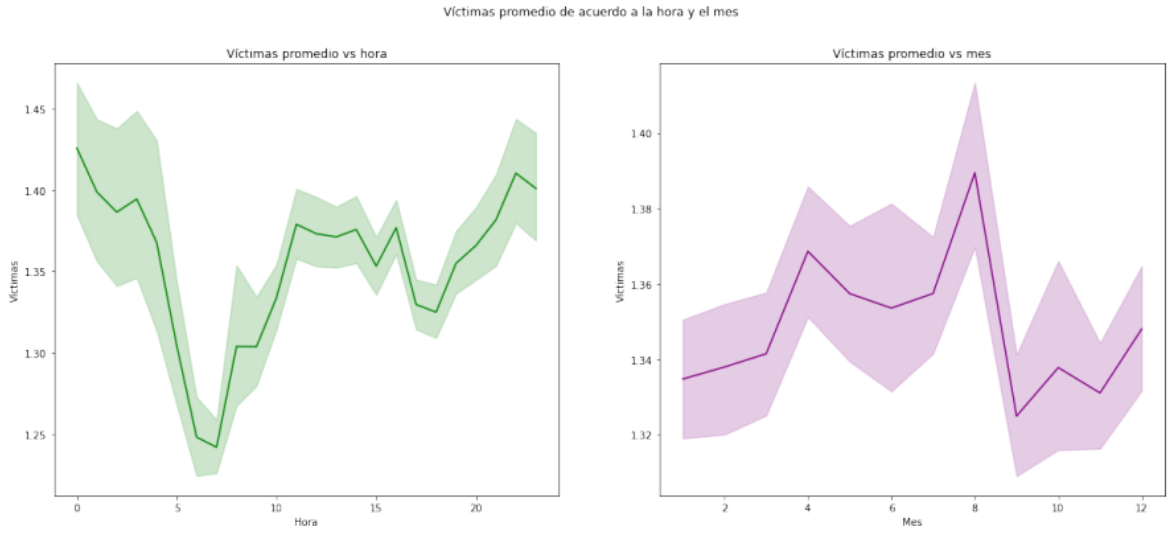


Figura 8. correlación accidentes hora mes

5.3. INDETIFICACION DE AREAS

La columna "Urban_or_Rural_Area" del dataset presenta únicamente dos valores posibles: 1 y 2. Aunque el dataset no brinda información explícita sobre qué valor corresponde a un área urbana o rural, es posible inferirlo mediante el análisis de otros datos disponibles, como la longitud y latitud, en conjunción con el conocimiento del mapa del Reino Unido. Este enfoque permitirá asignar con mayor precisión las categorías de área urbana y rural a los respectivos valores en la columna mencionada, optimizando así la interpretación y utilidad de estos datos para futuros análisis en el contexto de la seguridad vial.

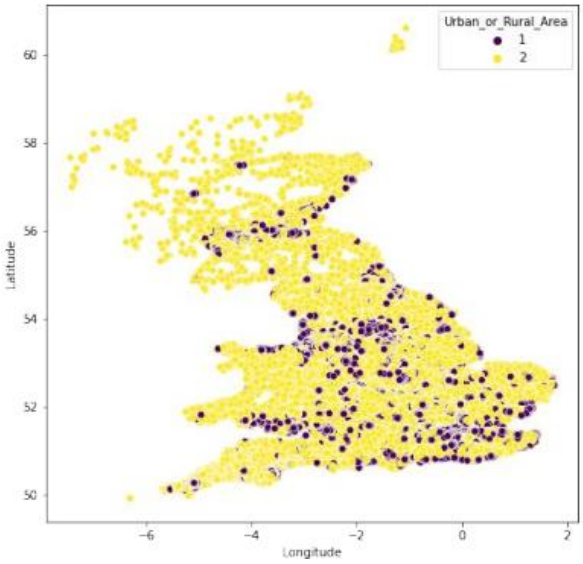


Figura 9. identificaciones de las zonas urbanas y rurales

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.

5.4. CORRELACION ENTRE VARIABLE OBJETIVO Y PARAMETROS

Se evidencia que, para la variable objetivo, los parámetros más correlacionados son el número de vehículos involucrados, la velocidad límite de la zona y si se trata de un área urbana o rural. Esta observación sugiere que la cantidad de vehículos en un incidente, las condiciones de velocidad específicas y el entorno urbano o rural son factores clave que influyen significativamente en la gravedad de los incidentes. Estos hallazgos destacan la importancia de considerar estos factores al desarrollar estrategias de seguridad vial y al implementar medidas preventivas para reducir el impacto de los accidentes de tránsito.

	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

6. APENDICE DE DATOS

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 como se muestra en la ilustración de la derecha

	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

7. TRATAMIENTO DE DATOS

En el análisis de la columna LSOA_of_Accident_Location, se ha identificado una tasa de aproximadamente el 6% de datos faltantes. Estos corresponden a una notación singular para cada zona del Reino Unido, sin embargo, su utilidad limitada y la complejidad de su relleno aconsejan la eliminación de esta columna del dataset.

Para las columnas Police_Force y Number_of_Vehicles, se propone realizar un análisis de la distribución de los datos disponibles. La mediana y la moda se calcularán para determinar

la estrategia de imputación más apropiada. La imputación se llevará a cabo utilizando la moda, que representa el valor más frecuente en los datos existentes.

En el caso de la columna Road_Type, en la cual se encuentran datos faltantes, se sugiere agrupar la totalidad de los valores ausentes bajo la categoría 'Unknown'. Esta acción proporcionará una identificación clara de los casos en los cuales no se dispone de información precisa sobre el tipo de vía en el momento del accidente.

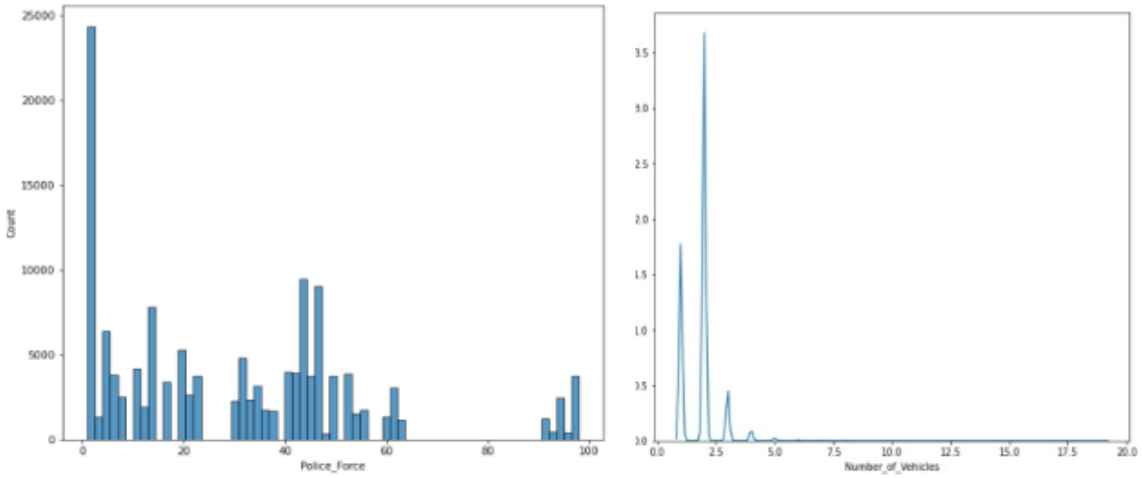


Figura 10. numero de policias(derecha) y numero de vehiculos usados(izq)

7.1. ELIMINACION DE DATOS

Se recomienda la eliminación de ciertas variables del dataset debido a la falta de aportes sustanciales o la presencia de duplicación de información:

Location_Easting_OSGR y Location_Northing_OSGR: Estos campos contienen información duplicada de las columnas Longitude y Latitude, respectivamente. Su eliminación se justifica para evitar redundancia en el conjunto de datos.

Year: Dado que todos los registros en el dataset corresponden al año 2014, la variable "Year" carece de variabilidad y puede ser eliminada sin afectar el análisis.

Number_of_Casualties_log: Esta columna, creada con fines de visualización de datos, no es esencial para el análisis y modelado. Su eliminación no comprometerá la información principal del conjunto de datos.

Unnamed: 0 y Accident_Index: Aunque estos campos sirven como identificadores únicos para cada accidente, no aportan información relevante al análisis. Por lo tanto, se sugiere su eliminación para mantener un dataset más conciso sin perder información valiosa.

Carriageway_Hazards y Special_Conditions_at_Site: Dado que estas variables contienen principalmente datos nulos, indicando información limitada, su eliminación contribuirá a mantener un dataset más limpio y completo.

Pedestrian_Crossing-Physical_Facilities y Pedestrian_Crossing-Human_Control: La mayoría de los accidentes no involucran facilidades de movilidad o control peatonal en un radio de 50 metros. Por ende, estas variables no ofrecen información significativa y se

sugiere su eliminación. Este paso simplificará el dataset, focalizándose en variables más relevantes para el análisis de accidentes de tránsito en el Reino Unido durante el año 2014.

7.2. GENERACION DE DATOS

Se han introducido variables adicionales para mejorar la descripción de la situación durante los accidentes. Estas nuevas variables incluyen:

I. Variable de día o noche: Indica la condición de iluminación en el momento del accidente, diferenciando entre el periodo diurno y nocturno.
II. Variable de estación del año: Proporciona información sobre la estación meteorológica durante el accidente, clasificando en primavera, verano, otoño e invierno.
III. Variable binaria de iluminación: Esta variable señala si la zona estaba bien iluminada en el momento del accidente, ofreciendo una clasificación binaria.

Adicionalmente, se ha creado una variable categórica para clasificar la gravedad de las víctimas en los accidentes:

- Accidentes leves: Menos de 5 víctimas.
- Accidentes moderados: Menos de 10 víctimas.
- Accidentes graves: Más de 10 víctimas

8. METODOS SUPERVISADOS

En la fase de métodos no supervisados, se evaluaron tres modelos: Random Forest Classifier, Decision Tree Classifier y Support Vector Classifier (SVC). Se empleó el proceso de validación cruzada y se adaptó el código proporcionado como ejemplo para seleccionar el modelo más adecuado para los datos. Es importante señalar que las diferencias en los errores de Root Mean Square Error (RSME) entre los tres modelos son mínimas, y cada uno podría generar modelos efectivos. No obstante, la decisión final se inclinó hacia trabajar exclusivamente con el clasificador "Decision Tree". En la figura se presentan los errores obtenidos para cada modelo: Random Forest Classifier, Decision Tree Classifier y SVC, respectivamente.

```
---------
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)
```

A continuación, se lleva a cabo la búsqueda de los mejores hiperparámetros para el modelo utilizando GridSearchCV, una herramienta de Scikit Learn que realiza validación cruzada utilizando diversos parámetros especificados antes de la ejecución del código. Los resultados obtenidos son los siguientes:

```
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. METODOS NO SUPERVISADOS

Para los métodos no supervisados, se llevó a cabo un Análisis de Componentes Principales (PCA), una función que extrae las características más representativas del dataset para realizar una transformación y mejorar los resultados con el modelo del Decision Tree. El análisis se realizó mediante el 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')
```

10. CURVAS DE APRENDIZAJE

Las curvas de aprendizaje proporcionan una representación visual de cómo se comporta el modelo a medida que se agregan más datos a lo largo del tiempo. En ambas gráficas, se observa una tendencia hacia el sesgo, indicando que los modelos pueden ser demasiado simples o que hay datos mezclados, lo que sugiere la necesidad de incluir más columnas en el dataset. Es importante destacar que en el modelo específico con PCA, se observa un comportamiento errático de los datos de entrenamiento, además de no presentar una mejora significativa en la métrica en comparación con el modelo supervisado. Esto podría deberse al uso de una reducción sustancial de columnas con PCA.

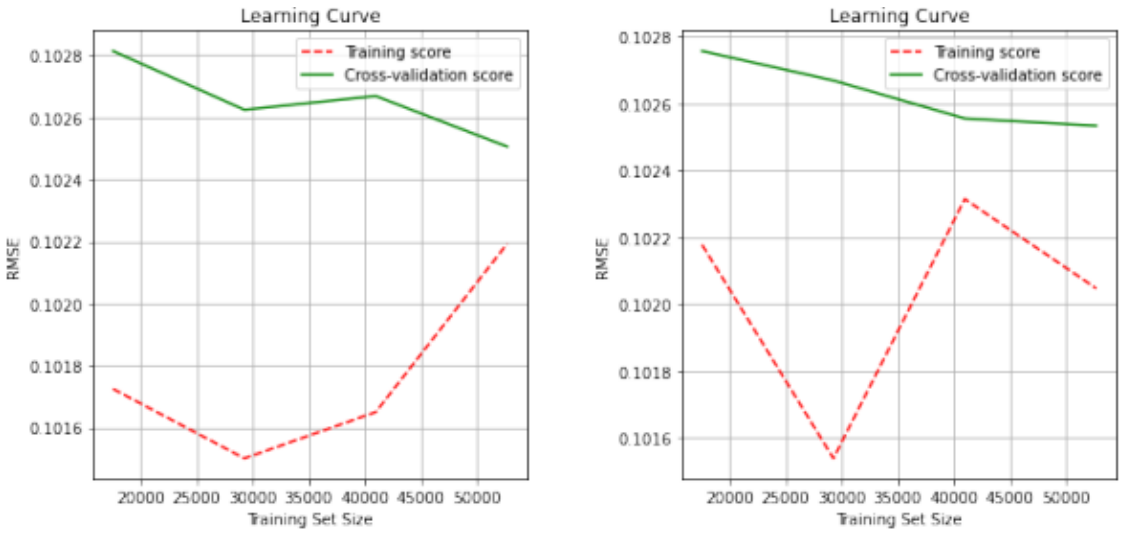


Figura 11. Curvas de aprendizaje, (De derecha a izquierda: Decision tree, Decision tree + PCA)

11. RETOS EN EL DESPLIEGUE DEL MODELO

El desarrollo del modelo de predicción de víctimas en accidentes automovilísticos enfrenta desafíos sustanciales, especialmente en la obtención de datos significativos. A pesar de la abundancia de información en el conjunto de datos actual sobre accidentes, no se logra alcanzar un rendimiento óptimo en términos de predicción. Esta circunstancia motiva la consideración de la inclusión de nuevas variables en el conjunto de datos, lo cual puede conllevar costos adicionales.

Para lograr la implementación exitosa de un modelo en entornos de producción, es imperativo abordar los siguientes desafíos:

I. Recolección de datos adicionales: Se hace necesario recopilar más información relativa a los accidentes para enriquecer el conjunto de datos existente. Variables como condiciones climáticas, estado de la vía y presencia de señales de tránsito podrían ser fundamentales. No obstante, llevar a cabo esta tarea implica consideraciones logísticas y puede generar costos asociados.

II. Evaluación con profesionales de la salud y servicios de emergencia: Se plantea la necesidad de establecer una colaboración activa con centros de salud, servicios de ambulancias y profesionales paramédicos. Esta colaboración permitiría evaluar la viabilidad del modelo, determinando si su implementación conlleva una mejora sustancial en la eficiencia de la atención a los accidentes automovilísticos.

12. CONCLUSIONES

Es altamente recomendable realizar esfuerzos para obtener datos adicionales que enriquezcan el conjunto de datos actual. Esta acción no solo mejorará el rendimiento del modelo, sino que también contribuirá a reducir el sesgo inherente presente en los datos actuales. La obtención de datos adicionales ofrecerá una muestra más completa y diversa, permitiendo a los modelos capturar de manera más efectiva la variabilidad de los accidentes

de tránsito. Un conjunto de datos más extenso proporcionará más ejemplos para el entrenamiento de modelos, aumentando así su capacidad de generalización.

En cuanto a la elección del modelo, es crucial considerar que los resultados pueden ser comparables entre los tres modelos evaluados inicialmente. No obstante, se recomienda explorar y analizar otros modelos disponibles en el ámbito de la predicción de accidentes de tránsito. Cada modelo posee sus propias fortalezas y debilidades, y la exploración de diferentes enfoques brindará una visión más completa y robusta de las posibles soluciones.

El sesgo presente en los datos puede ser influenciado por la propia naturaleza de los accidentes de tránsito. Si existe una concentración significativa de valores cercanos a 1 en los datos, esto puede plantear desafíos para que los modelos "aprendan" de manera efectiva. Los modelos podrían enfrentar dificultades para identificar y generalizar patrones en los datos debido a esta falta de variabilidad. Por lo tanto, es esencial considerar técnicas de preprocesamiento y manejo de datos que aborden este sesgo, como la transformación logarítmica mencionada anteriormente. Estas técnicas mejorarán la capacidad de los modelos para aprender y realizar predicciones más precisas.