ENTREGA FINAL DEL PROYECTO

ALEJANDRO ARIAS ORTIZ
LUIS MATEO OCHOA AGUDELO
DUVAN ESNEIDER GALLEGO JIMENEZ

PROFESOR:
Raúl Ramos Pollan

MODELOS DE SISTEMAS 1 / INTRODUCCIÓN A LA INTELIGENCIA ARTIFICIAL UNIVERSIDAD DE ANTIOQUIA
FACULTAD DE INGENIERÍA MEDELLÍN 2023

Contenido

Introducción.

Los accidentes de tránsito son un serio problema que afecta a muchas personas cada año. Estos accidentes, a menudo causados por imprudencias o condiciones de los conductores, se han convertido en una preocupación importante para los países. Además de enfrentar el reto de controlar el tráfico, también hay que lidiar con el impacto que estos accidentes tienen en la salud de los involucrados.

La gestión adecuada de los datos es clave en este escenario, sobre todo cuando se trata de riesgos relacionados con la seguridad vial, como los accidentes de coches. Con un conjunto de datos exhaustivo, se busca pronosticar la probabilidad de accidentes considerando sus causas y efectos, lo que permitirá enfocar los esfuerzos en la prevención y reducción de accidentes, así como proveer información vital a los interesados, como hospitales, para que puedan anticiparse a un posible incremento en la llegada de pacientes debido a accidentes de tráfico.

La inteligencia artificial (IA) surge como una solución prometedora en este ámbito, ya que puede procesar y analizar grandes cantidades de datos para extraer información valiosa. Con la IA, se busca obtener conocimientos que permitan acciones más efectivas contra las causas y efectos de los accidentes viales, y se espera que también facilite la toma de decisiones en la planificación y respuesta médica. En última instancia, la combinación de IA y análisis de datos tiene un gran potencial para mejorar la prevención de accidentes de tránsito y reducir su impacto negativo.

Exploración descriptiva del Dataset

El conjunto de datos de accidentes de tráfico de EE. UU. contiene información detallada sobre 259.077 accidentes automovilísticos que ocurrieron en todos los estados de EE. UU. (excepto en áreas remotas de Alaska y Hawái) entre 2016 y 2020. Los conjuntos de datos tienen información muy detallada sobre cada accidente, incluida la ubicación, la hora, el clima, los factores contribuyentes y la información demográfica de los automovilistas involucrados en el accidente. El programa de muestreo se vio profundamente afectado por la pandemia COVID19, por lo que no hay nada más allá del año 2020 debido a una revisión importante del programa para acomodar el trabajo remoto. Los datos de 2020 pueden ser muy escasos para los meses de marzo a mayo, dado el hecho de que estos informes provienen de los departamentos de policía locales y muchos de ellos se vieron afectados en la primavera de 2020.

La base de datos proporciona una gran cantidad de información relevante sobre cada accidente registrado. Algunos de los datos incluidos son:

1. Fecha y hora: Información sobre cuándo ocurrió el accidente.
2. Tipo: Clasificación del tipo de accidente, como colisión de vehículos, atropello, choque con un objeto fijo, etc.
3. Condiciones: Descripción de las condiciones climáticas en el momento del accidente.
4. Estado: Información sobre el estado de la carretera, como seca, mojada, helada, etc.
5. Contribuciones: Factores que se consideran que contribuyen al accidente, como la velocidad, el consumo de alcohol, el uso del cinturón de seguridad, etc.
6. Gravedad: Indicación de la gravedad del accidente en términos de personas fallecidas, heridas graves o heridas leves.

El análisis de estos datos proporcionará información valiosa sobre las causas y consecuencias de los accidentes de tráfico. Esto a su vez facilitará el desarrollo de estrategias de prevención y la toma de decisiones informadas para mejorar la seguridad vial y reducir la tasa de accidentes en el futuro.

1. STRATUM: Podría ser una clasificación o estratificación de los datos.
2. REGION: La región geográfica donde ocurrió el accidente.
3. URBANICITYNAME: Indica si el área es urbana o rural.
4. VE_TOTAL: Número total de vehículos involucrados en el accidente.
5. NUM_INJNAME: Número de personas heridas en el accidente.
6. MONTHNAME: El mes en que ocurrió el accidente.
7. YEARNAME: El año del accidente.
8. DAY_WEEKNAME: Día de la semana cuando ocurrió el accidente.
9. HOURNAME: Hora del día cuando ocurrió el accidente.
10. MINUTENAME: Minuto exacto cuando ocurrió el accidente.
11. HARM_EV: Código o identificador del evento dañino principal en el accidente.
12. HARM_EVNAME: Descripción del evento dañino principal.
13. ALCOHOL: Indicador de si el alcohol estuvo involucrado en el accidente.
14. ALCOHOLNAME: Descripción de la implicación del alcohol.
15. MAX_SEV: Gravedad máxima del accidente.
16. MAX_SEVNAME: Descripción de la gravedad máxima del accidente.
17. MAN_COLL: Tipo de colisión (por ejemplo, frontal, trasera, lateral).

18. MAN_COLLNAME: Descripción del tipo de colisión.
19. RELJCT1NAME: Información sobre si el accidente ocurrió en una intersección.
20. RELJCT2NAME: Información más específica sobre la relación con la intersección.
21. TYP_INT: Tipo de intersección donde ocurrió el accidente.
22. WRK_ZONENAME: Indica si el accidente ocurrió en una zona de trabajo.
23. REL_ROADNAME: Relación del accidente con la carretera (en la carretera, adyacente, etc.).
24. LGT_CONDNAME: Condiciones de luz en el momento del accidente.
25. SCH_BUSNAME: Indica si un autobús escolar estuvo involucrado en el accidente.
26. INT_HWYNAME: Indica si el accidente ocurrió en una carretera interestatal.
27. WEATHERNAME: Condiciones meteorológicas en el momento del accidente.
28. WKDY_IMNAME: Posiblemente una codificación del día de la semana.
29. HOUR_IMNAME: Posiblemente una codificación de la hora del accidente.
30. MINUTE_IMNAME: Posiblemente una codificación del minuto del accidente.
31. EVENT1_IMNAME: Primer evento importante registrado en el accidente.
32. MANCOL_IMNAME: Codificación del tipo de colisión.
33. RELJCT1_IMNAME: Codificación de la relación del accidente con la intersección.
34. RELJCT2_IMNAME: Codificación de información más específica sobre la intersección.
35. LGTCON_IMNAME: Codificación de las condiciones de luz.
36. WEATHR_IMNAME: Codificación de las condiciones meteorológicas.
37. MAXSEV_IMNAME: Codificación de la gravedad máxima del accidente.
38. NO_INJ_IMNAME: Número de personas no heridas o codificación de este dato.
39. ALCHL_IMNAME: Codificación de la implicación del alcohol en el accidente.
40. WEIGHT: Peso o factor de ponderación para estadísticas o muestras.

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_i corresponde a los valores observados en la serie y $\hat{y}_i$ 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.

Iteraciones de desarrollo.

Preprocesado de datos

Análisis de la variable objetivo

Es crucial examinar y entender el comportamiento de la variable clave en este proyecto, "NUM_INJNAME" (Número de Víctimas). Al analizarla, se nota una tendencia desigual hacia cifras cercanas a 1. Considerando que representa valores enteros, es esencial evaluar su distribución para confirmar que no todos los casos reporten un valor de 1.

Por lo tanto, se realiza un análisis de los distintos valores que toma esta variable. Si se descubre una gama variada de números diferentes a 1, se considerará la posibilidad de aplicar una transformación logarítmica. Esta técnica mejora la visualización de los datos y disminuye la asimetría, facilitando así su interpretación y análisis.

La aplicación de la transformación logarítmica a "NUM_INJNAME" proporciona una visión más clara de su distribución y patrones. También atenúa el impacto de los valores extremos y ofrece una representación más equilibrada y propicia para el análisis.

Al comprender mejor esta variable y emplear las transformaciones pertinentes, se pueden tomar decisiones más informadas y crear modelos predictivos más exactos y fiables. Esto contribuye a abordar eficazmente los problemas relacionados con los accidentes de tráfico y a implementar estrategias preventivas apropiadas en EE. UU.

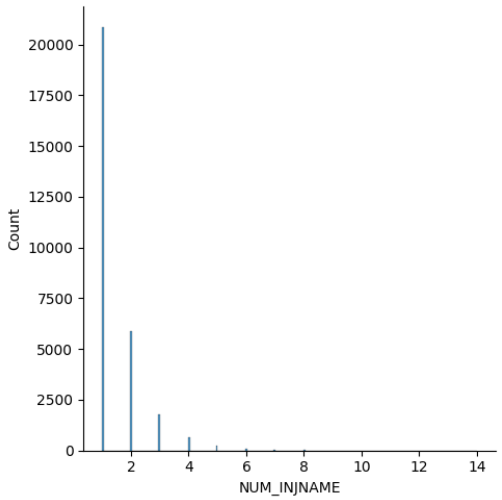


Ilustración 1: Distribución de la variable objetivo.

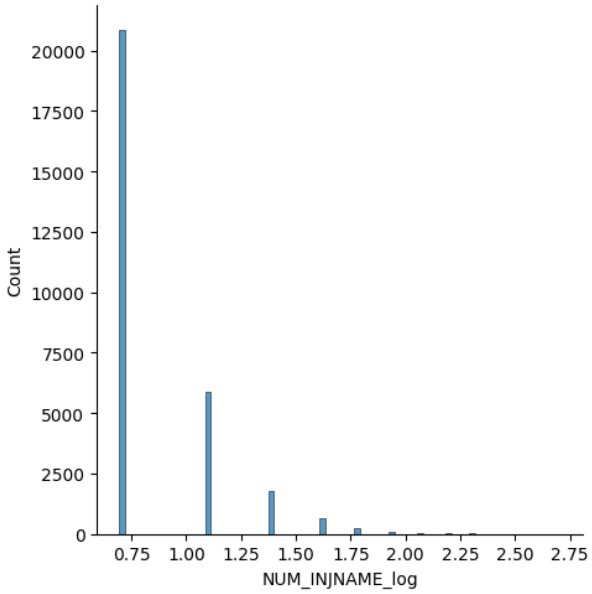


Ilustración 2: Transformación logarítmica.

En la Figura 2, se evidencia que la distribución de "NUM_INJNAME", tras la transformación logarítmica, muestra una mejora significativa y resulta más apta para su análisis. Este cambio ha permitido el aprovechamiento de una mayor cantidad de datos que anteriormente se veían afectados por el sesgo en algunos rangos de la gráfica original en la Figura 1.

La alteración hecha a la variable clave mediante la transformación logarítmica ha optimizado notablemente su distribución. Esto se traduce en un incremento de los datos útiles para el análisis, y para la realización de pruebas de programación y procesos algorítmicos. Al usar la variable modificada, se facilita un análisis más detallado y efectivo, maximizando el uso de todos los datos disponibles y minimizando el sesgo previo. Esto conducirá a resultados más confiables y a decisiones más acertadas.

Exploración de variables

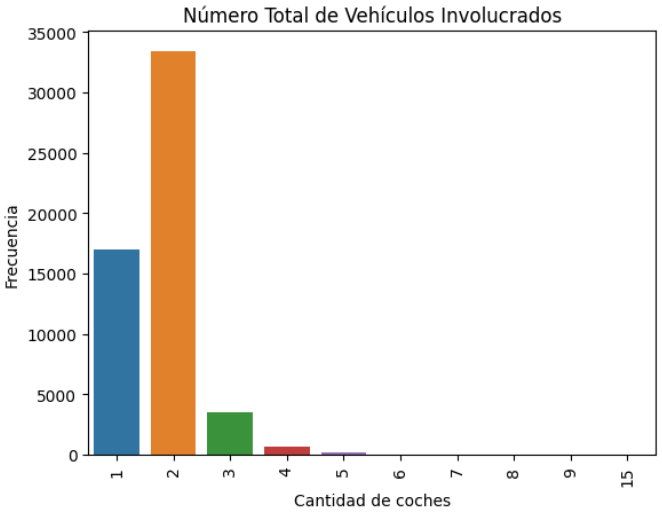
La investigación de las variables es un paso esencial para la elaboración del modelo, ya que ofrece una perspectiva sobre su conexión con la variable principal. Este proceso implica tener un conjunto predefinido de variables para su análisis. Dichas variables se importan y se emplean para calcular estadísticas y elaborar histogramas, los cuales son clave para entender y describir la situación de los accidentes en EE. UU. y sus consecuencias.

El estudio de estas variables nos ayuda a analizar cómo se relacionan con la variable principal. Utilizando medidas estadísticas como el promedio, la desviación estándar y la correlación, adquirimos conocimientos importantes acerca de la influencia de cada variable en los accidentes y sus resultados. Además, al crear histogramas y otras visualizaciones, podemos detectar patrones, tendencias y posibles factores de riesgo ligados a los accidentes de tráfico en EE. UU.

La lista de variables importadas y examinadas en esta fase nos brinda una base firme para comprender a fondo el problema de los accidentes y sus implicaciones asociadas. Al evaluar estas variables de manera integral, estaremos en posición de tomar decisiones más informadas y desarrollar estrategias efectivas para abordar la problemática de los accidentes en

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.



La conclusión principal que se desprende del histograma es que los eventos que involucran dos vehículos son los más frecuentes, mientras que los incidentes que involucran a tres o más vehículos son mucho menos comunes. La prevalencia de eventos con dos vehículos podría estar influenciada por factores específicos del entorno o las circunstancias que fueron objeto de estudio, mientras que los incidentes con muchos vehículos son eventos atípicos y raros en este contexto particular.

Ilustración 3: Automóviles Involucrados

Este histograma titulado "Distribución de los Mes del Accidente" muestra la frecuencia de accidentes por mes. Se puede observar que los meses de verano y finales de otoño (julio, octubre y noviembre) tienen las frecuencias más altas, lo que sugiere un aumento en los accidentes durante estos períodos. Por otro lado, los meses de primavera como marzo y mayo muestran las frecuencias más bajas. Esta tendencia puede estar relacionada con factores estacionales como las condiciones del clima, los patrones de viaje o las vacaciones, que podrían influir en la cantidad de accidentes que ocurren cada mes.

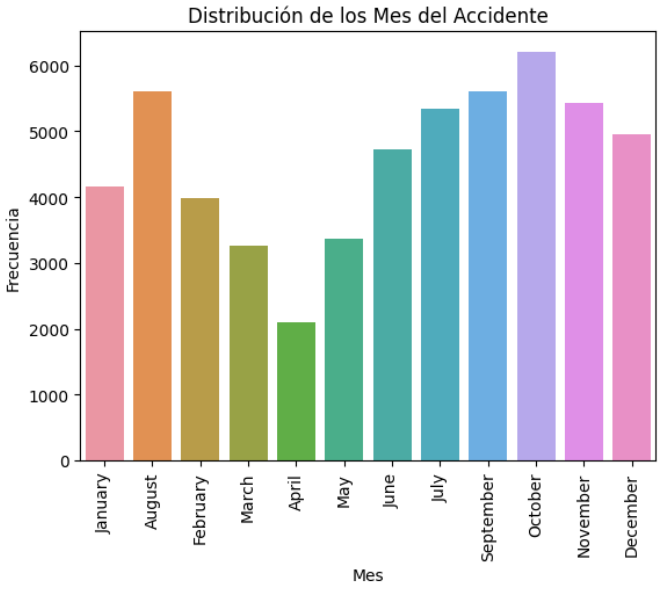
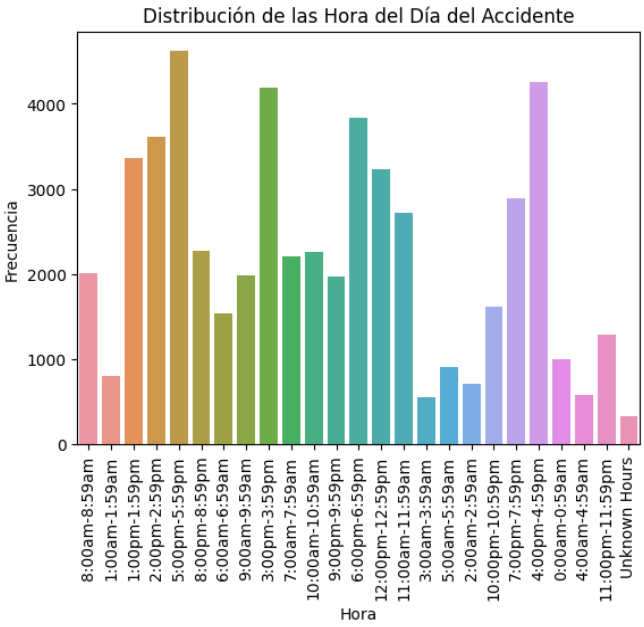
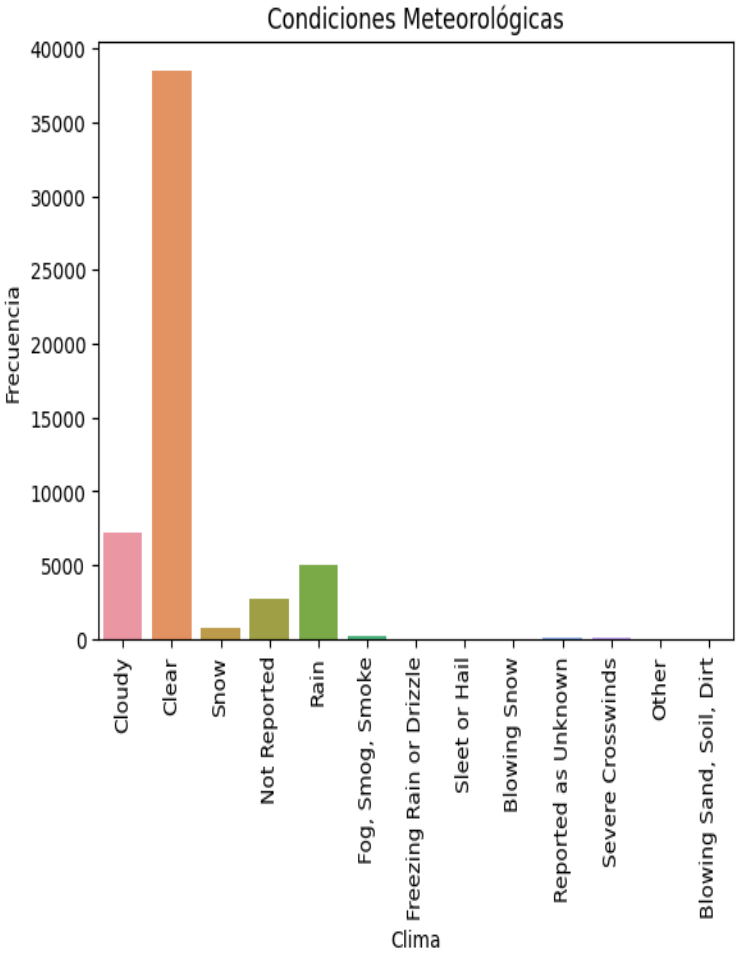


Ilustración 4: Distribución de los Mes del Accidente



Se muestra la frecuencia de accidentes distribuida a lo largo de diferentes horas del día. La barra más alta ocurre en el intervalo de tiempo de "4:00 pm-4:59 pm", lo que indica que es la hora del día con más accidentes registrados, posiblemente debido al tráfico de la hora punta de la tarde cuando la gente regresa a casa del trabajo o la escuela. También hay un pico notable durante la hora de "8:00 am-8:59 am", otra hora punta común. Las horas entre medianoche y la mañana temprano tienen las frecuencias más bajas, lo que puede reflejar menos tráfico en las carreteras durante estas horas. Curiosamente, hay una categoría para "Unknown hours", lo que sugiere que para algunos accidentes, la hora exacta no fue registrada o determinada. Estos patrones pueden ser importantes para las iniciativas de seguridad vial y planificación urbana.

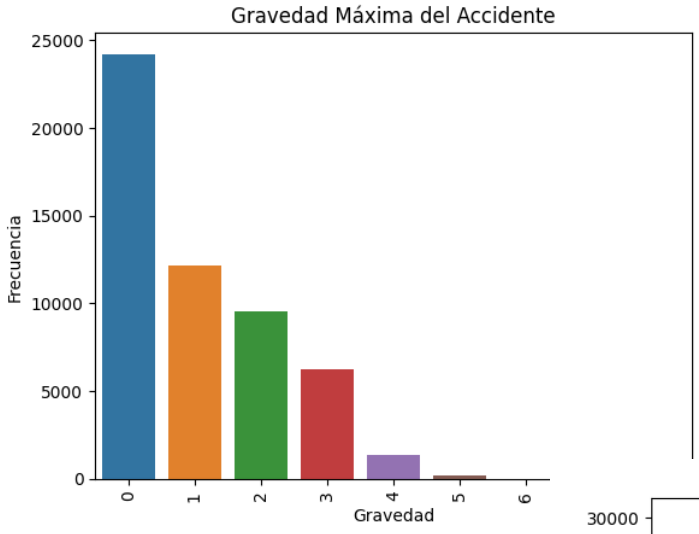
Ilustración 5: Distribución de las Hora del Día del Accidente



El histograma muestra que la mayoría de los eventos registrados ocurren en condiciones climáticas despejadas ("Clear"), sugiriendo que los eventos no están necesariamente relacionados con el mal tiempo. Las condiciones de "Cloudy" y "Rain" también son comunes pero mucho menos frecuentes. Eventos durante condiciones adversas como "Snow" y "Fog/Smog/Smoke" son aún menos comunes, indicando que los días con mal tiempo podrían no ser los más propensos para eventos o que las personas toman más precauciones durante estos tiempos. Esto subraya la importancia de considerar la influencia del clima en la planificación de la seguridad vial. Además, la categoría "Other" tiene una frecuencia muy baja, lo que puede incluir condiciones climáticas menos comunes o más específicas no listadas en el histograma. Estos datos pueden ser cruciales para comprender la correlación entre las condiciones climáticas y la frecuencia de eventos para mejorar la seguridad y la planificación.

Ilustración 6: Condiciones Meteorológicas

El histograma titulado "Gravedad Máxima del Accidente" presenta una distribución de la frecuencia de accidentes por niveles de gravedad, que parecen estar codificados numéricamente en el eje horizontal. La barra más alta corresponde a la categoría de gravedad "0", lo que sugiere que la mayoría de los accidentes registrados se clasifican en la categoría de gravedad más baja, que podría representar accidentes sin lesiones o con daños mínimos. Las categorías con mayor gravedad tienen una frecuencia progresivamente menor, lo que indica que los accidentes graves son



menos frecuentes. Esto podría reflejar que, aunque ocurren accidentes, la mayoría no resultan en lesiones graves o daños significativos.

Ilustración 7: Gravedad de accidentes

El histograma "Indicador de Alcohol Involucrado" muestra la frecuencia de accidentes clasificados según la presencia de alcohol. La categoría "No Alcohol Involved" tiene la frecuencia más alta, lo que sugiere que la mayoría de los accidentes registrados no estuvieron relacionados con el alcohol. La categoría "Reported as Unknown" también es significativa, indicando que en muchos casos no se conoce o no se informa si el alcohol estuvo involucrado. Los accidentes donde sí se confirma la implicación de alcohol ("Alcohol Involved") son considerablemente menos frecuentes. Esto puede indicar que, aunque el alcohol es un factor en algunos accidentes, no es el predominante en la mayoría de los casos según este conjunto de datos. La presencia de la categoría "No applicable person" podría referirse a accidentes donde no había una persona relevante para atribuir la influencia del alcohol, como podría ser en accidentes con vehículos no tripulados o en propiedad privada sin conductores implicados.

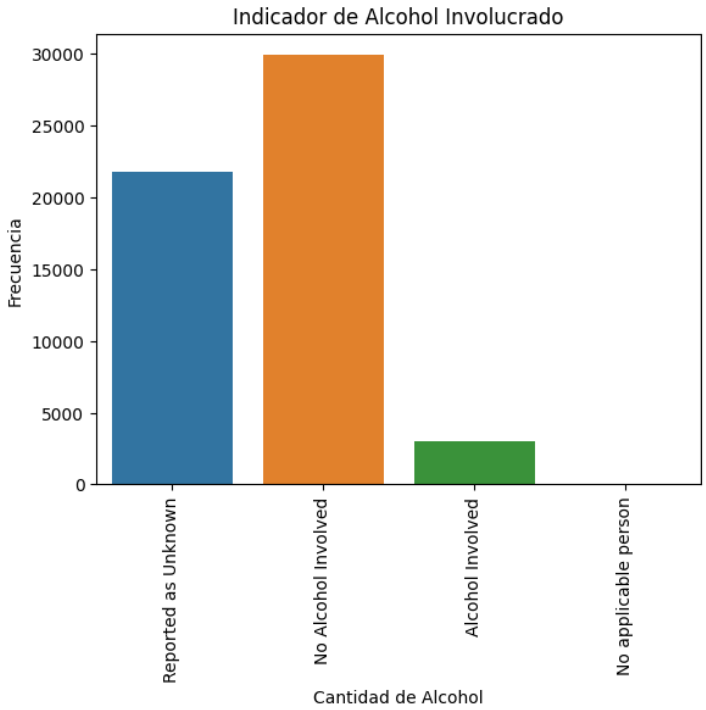


Ilustración 8: Alcohol Involucrado

Accidentes por hora y mes.

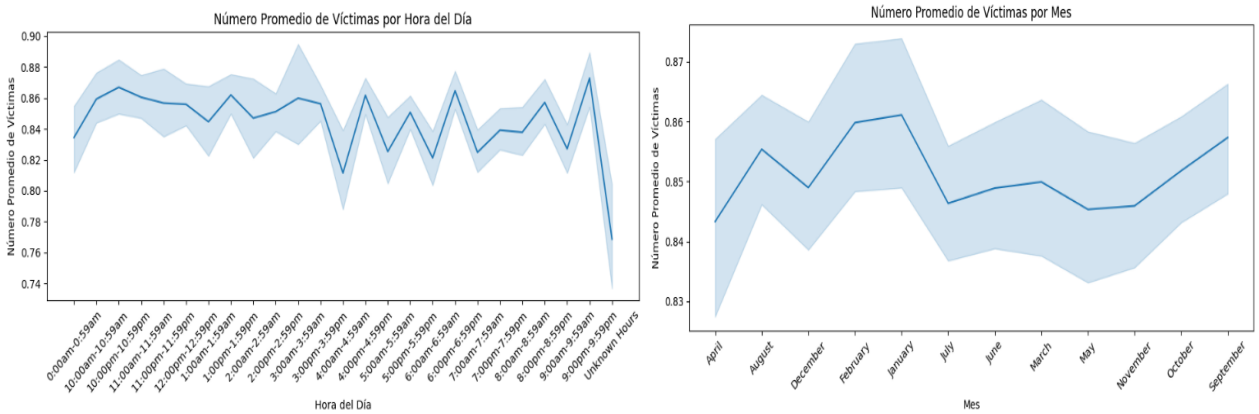


Ilustración 9: Correlación de los accidentes por hora y meses.

Al analizar las gráficas proporcionadas, las cuales representan el número promedio de víctimas por hora del día y por mes, se pueden realizar las siguientes observaciones y conclusiones:

Número Promedio de Víctimas por Hora del Día:

La gráfica muestra fluctuaciones en el número promedio de víctimas a lo largo del día. Hay una disminución notable en las primeras horas de la mañana, seguida de un incremento a medida que avanza el día, con picos y valles que pueden coincidir con las horas pico de tráfico.
Se observan bajos en horarios que pueden corresponder a periodos de menor actividad vehicular, como las horas tempranas de la mañana.
Por la noche, el número promedio de víctimas tiende a disminuir, lo que podría sugerir menos tráfico o que hay menos posibilidad de accidentes graves durante este tiempo.

Número Promedio de Víctimas por Mes:

La gráfica mensual muestra variaciones a lo largo del año, con algunos meses presentando mayores promedios de víctimas que otros.
Podría haber un aumento en el número promedio de víctimas en meses específicos, lo que podría estar influenciado por condiciones estacionales, como condiciones climáticas adversas o aumento del tráfico debido a las vacaciones.
En resumen, las tendencias en estas gráficas pueden indicar cómo factores como la hora del día y la estacionalidad afectan la incidencia y gravedad de los accidentes. Estos datos son cruciales para desarrollar estrategias de prevención y para implementar medidas de seguridad vial que puedan adaptarse a los patrones observados, con el objetivo de reducir la cantidad de víctimas en accidentes de tráfico.

Identificación Área Urbana/ Rural.

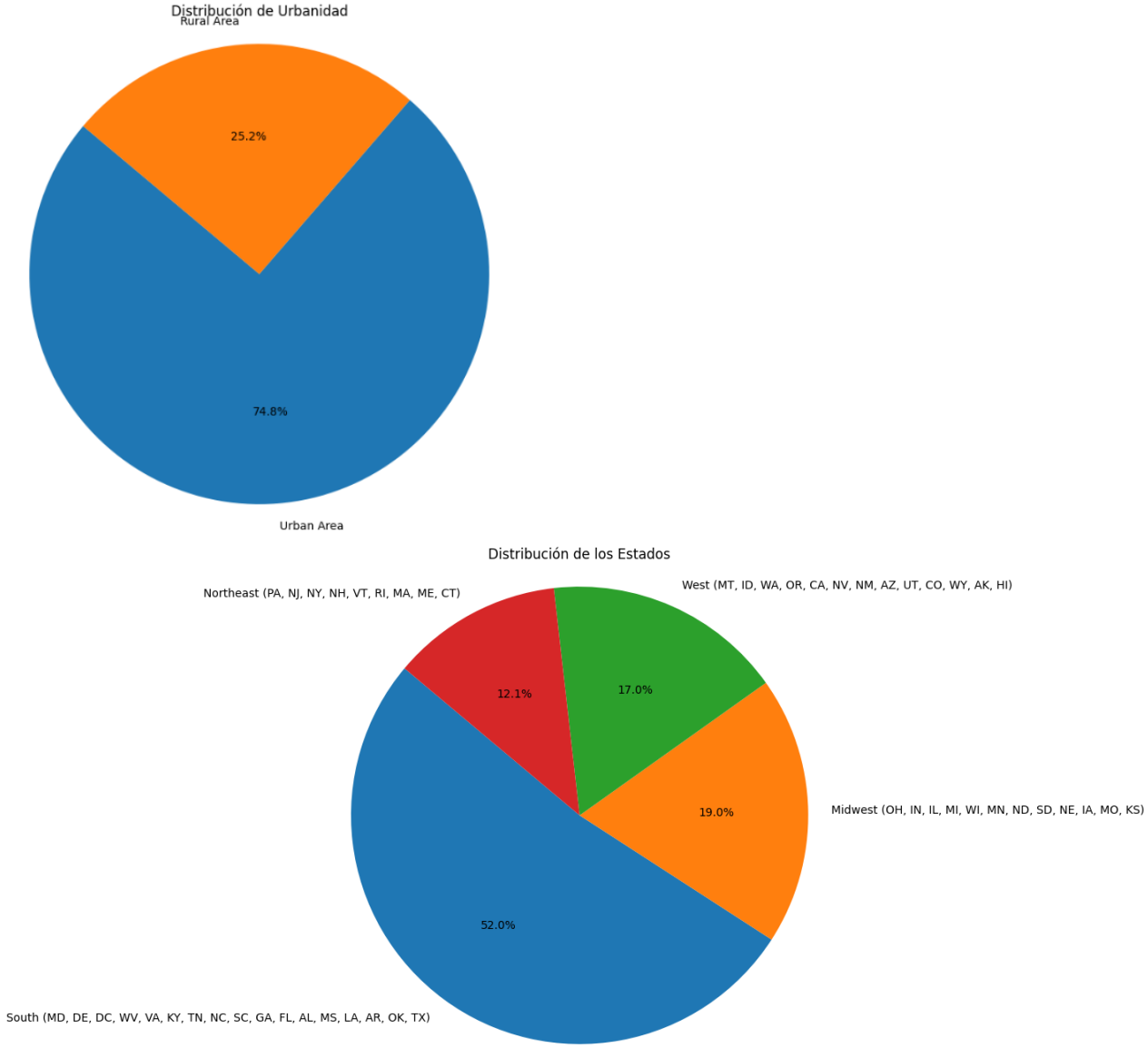


Ilustración 10. Identificación de las zonas geograficas

La primera gráfica, "Distribución de los Estados", muestra la proporción de eventos, que podrían ser accidentes de tráfico, divididos por regiones geográficas en Estados Unidos. La región Sur tiene la mayor proporción con un 52%, seguida por el Medio Oeste con un 19%, el Oeste con un 17% y el Noreste con el 12.1%. Esta distribución sugiere que la región Sur tiene una cantidad significativamente mayor de eventos en comparación con otras regiones, lo que podría deberse a una serie de factores como la densidad de población, la extensión geográfica, o incluso el clima y las políticas de tráfico regionales.

La segunda gráfica, "Distribución de Urbanidad", contrasta eventos en áreas urbanas y rurales, con un predominio en las áreas urbanas (74.8%) sobre las rurales (25.2%). Esta distribución puede reflejar la mayor densidad de tráfico en áreas urbanas y, por lo tanto, una mayor incidencia de eventos.

Al integrar ambos gráficos, se podría inferir que dentro de las regiones geográficas, las áreas urbanas son más propensas a eventos. Por ejemplo, aunque la región Sur lidera en la cantidad de eventos, sería interesante analizar cómo esta cifra se correlaciona con la urbanización de dicha región. Los datos sugieren que las

políticas de prevención y seguridad vial podrían necesitar un enfoque más concentrado en áreas urbanas, especialmente en la región Sur, para abordar eficientemente la incidencia de eventos. Además, se podría considerar que las diferencias regionales en las estadísticas de eventos pueden estar influenciadas por la urbanización y factores sociodemográficos específicos de cada región.

Correlación entre parámetros y variable objetivo

	NUM_INJNAME
NUM_INJNAME_log	1.000000
NUM_INJNAME	1.000000
NUM_INJ	0.977060
NO_INJ_IM	0.956527
PERMVIT	0.535134
VE_FORMS	0.349644
VE_TOTAL	0.336984
MANCOL_IM	0.228385
MAN_COLL	0.147616
STRATUM	0.139289
MAX_SEV	0.133855
MAXSEV_IM	0.132575
REGION	0.048952
PSUSTRAT	0.043562
HOUR_IM	0.030157
INT_HWY	0.027698
PJ	0.024289
DAY_WEEK	0.012740
WKDY_IM	0.012740
WEATHR_IM	0.012654
WEATHER	0.011620
RELJCT1_IM	0.011411
HOUR	0.009124
SCH_BUS	0.007771
MINUTE_IMNAME	0.006742
MINUTE_IM	0.006742
RELJCT2_IM	0.006055
RELJCT2	0.005793
URBANICITY	0.004302
RELJCT1	0.003086
MINUTE	0.002658
PVH_INVL	0.002455
WRK_ZONE	0.002426
TYP_INT	-0.001820
MONTH	-0.006329
LGTCON_IM	-0.007819
PSU	-0.014201
LGT_COND	-0.014316
CASENUM	-0.015080
PSU_VAR	-0.016018
ALCHL_IM	-0.023087
ALCOHOL	-0.055863
EVENT1_IM	-0.057278
HARM_EV	-0.059141
REL_ROAD	-0.077968
PERNOTMVIT	-0.167438
PEDS	-0.180857
YEAR	NaN
YEARNAME	NaN

Variables Positivas: Los coeficientes positivos sugieren que hay una relación directa entre la variable y el resultado; es decir, a medida que la variable aumenta, también lo hace el resultado. Por ejemplo, una variable con un coeficiente de 0.009124 aumenta ligeramente el resultado cuando aumenta.

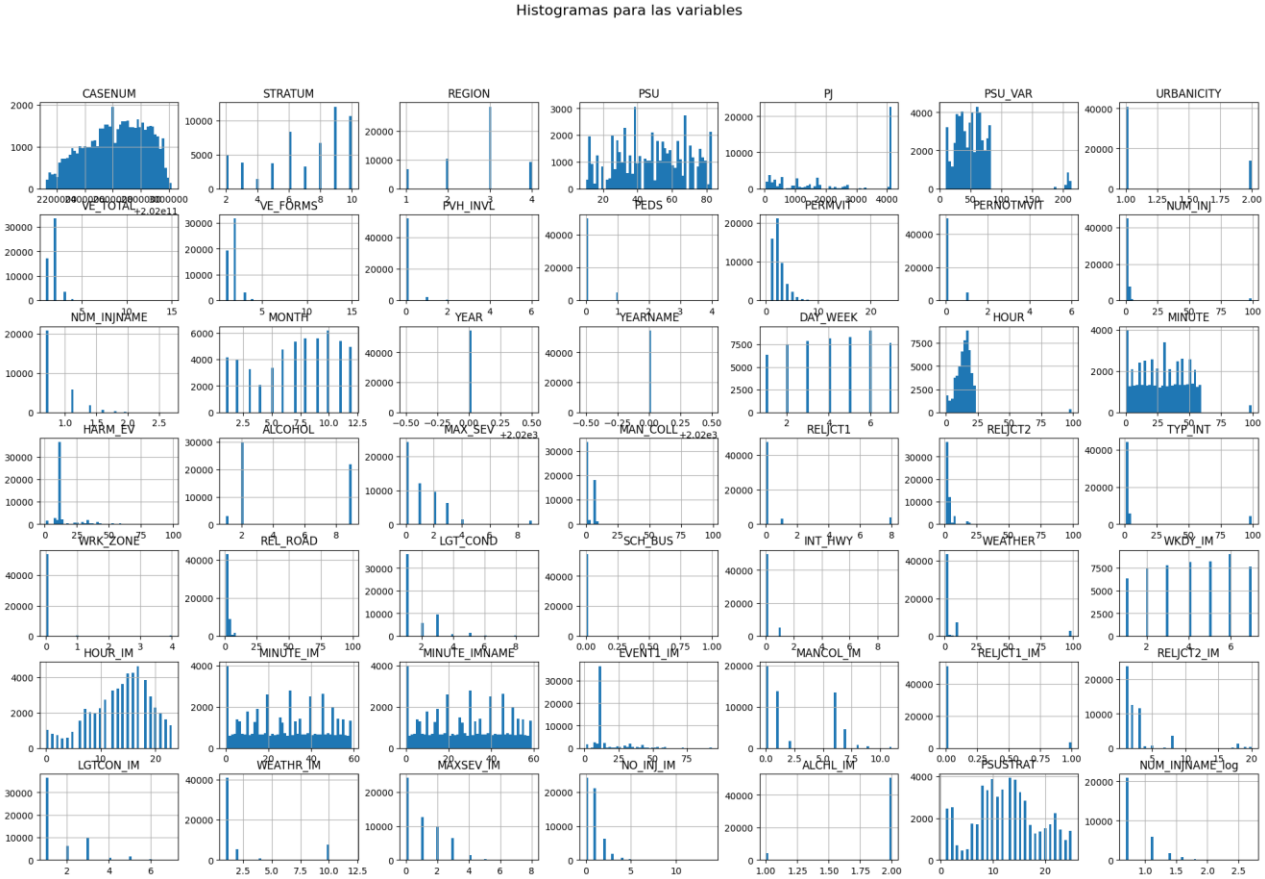
Variables Negativas: Los coeficientes negativos indican una relación inversa; a medida que la variable aumenta, el resultado disminuye. Por ejemplo, una variable con un coeficiente de -0.157438 disminuiría significativamente el resultado a medida que aumenta.

Variables Significativas: Variables con coeficientes más grandes en valor absoluto son generalmente consideradas más significativas en su impacto en el resultado. La significancia también puede ser evaluada en base a pruebas estadísticas, que no se muestran aquí.

Interpretación de Variables Específicas:

ALCOHOL y ALCHL_IM con coeficientes negativos sugieren que la presencia de alcohol está inversamente relacionada con el resultado. Esto puede parecer contra-intuitivo y requeriría una investigación adicional para entender el contexto.
NUM_INJ o NO_INJ_IM están muy correlacionados con el resultado, lo que indica que el número de lesiones es un predictor fuerte de este.
Variables como WEATHER y WEATHR_IM tienen coeficientes bajos, lo que sugiere un impacto menor en el resultado.

Distribución de las variables numéricas.



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 NUM_INJUNAME. Con el fin de simular la falta de datos en dos columnas adicionales, se han seleccionado STRATUM y REGION. De esta manera, los datos faltantes se distribuyen de la siguiente manera:

	Total	Percent
NUM_INJNAME	25241	46.106494
STRATUM	2737	4.999543
REGION	2737	4.999543

Tratamiento de datos.

Durante la fase de limpieza de datos, identificamos que aproximadamente el 46% de los valores en la columna NUM_INJ_INNAME estaban ausentes. Dado que estos datos faltantes corresponden a una nomenclatura cuya completitud no resulta crítica para nuestro análisis y cuyo proceso de imputación sería poco práctico, optamos por excluir esta variable del conjunto de datos. Esta acción se realizó mediante la función drop de la biblioteca pandas, asegurando así la integridad y la calidad de nuestro análisis de datos.

Posteriormente, nos enfocamos en la visualización y el análisis de la distribución de las variables categóricas STRATUM y REGION. Utilizamos un histograma para representar la distribución de STRATUM, donde observamos una prevalencia notable de ciertos estratos sobre otros, destacando en particular el estrato 10 con la mayor frecuencia. Esta observación nos indica una concentración de los datos en categorías específicas de

estrato que podrían ser de interés para análisis posteriores.

Para la variable REGION, aplicamos un gráfico de densidad kernel para visualizar la probabilidad de la distribución de las observaciones a lo largo de las categorías regionales. Identificamos cuatro picos distintos que reflejan la existencia de cuatro regiones predominantes, siendo la región 3 la más destacada entre ellas.

Las estadísticas descriptivas para estas variables revelaron que la mediana y la moda para STRATUM son 8 y 9 respectivamente, mientras que para REGION, ambas medidas son 3, corroborando la prominencia de esta última en nuestro conjunto de datos.

La interpretación de estas visualizaciones nos proporciona una comprensión detallada de la estructura de nuestro conjunto de datos, lo cual es fundamental para dirigir los esfuerzos analíticos futuros y para el desarrollo de modelos estadísticos robustos.

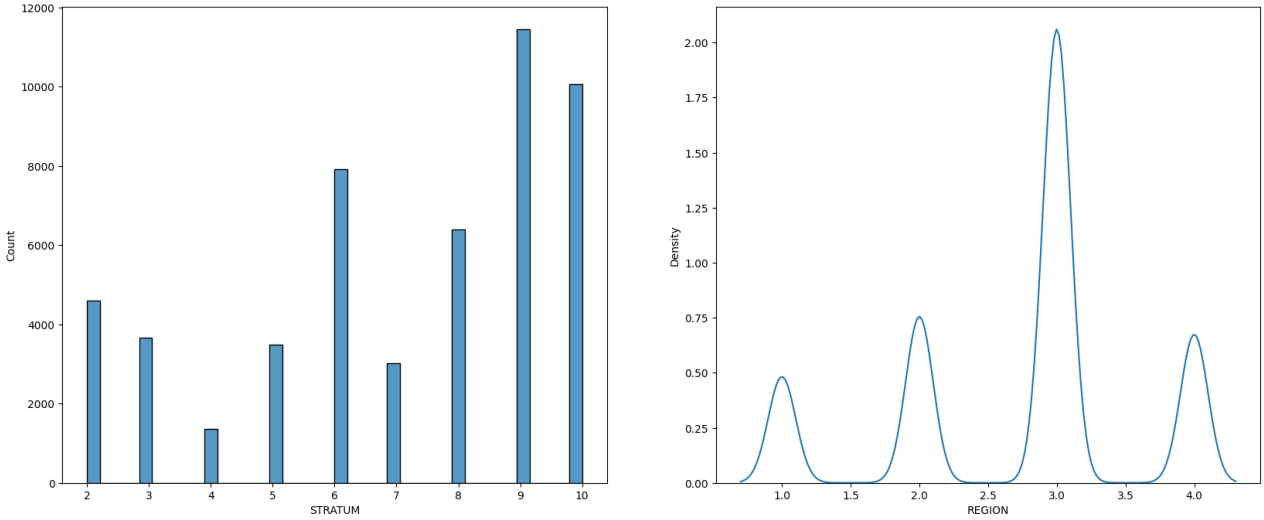


Ilustración 11: Graficas de STRATUM y REGION

Modelos 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:  0.02063 (± 0.00258031 )
RMSE  Train:  0.01923 (± 0.00370340 )
---------
RMSE Test:  0.01807 (± 0.00289039 )
RMSE  Train:  0.01769 (± 0.00151044 )
---------
RMSE Test:  0.02212 (± 0.00315717 )
RMSE  Train:  0.01788 (± 0.00217220 )
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': 5}
```

```
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: 0.024675191498735514
 En train: 0.019113736784006 46
```

Modelos no supervisados

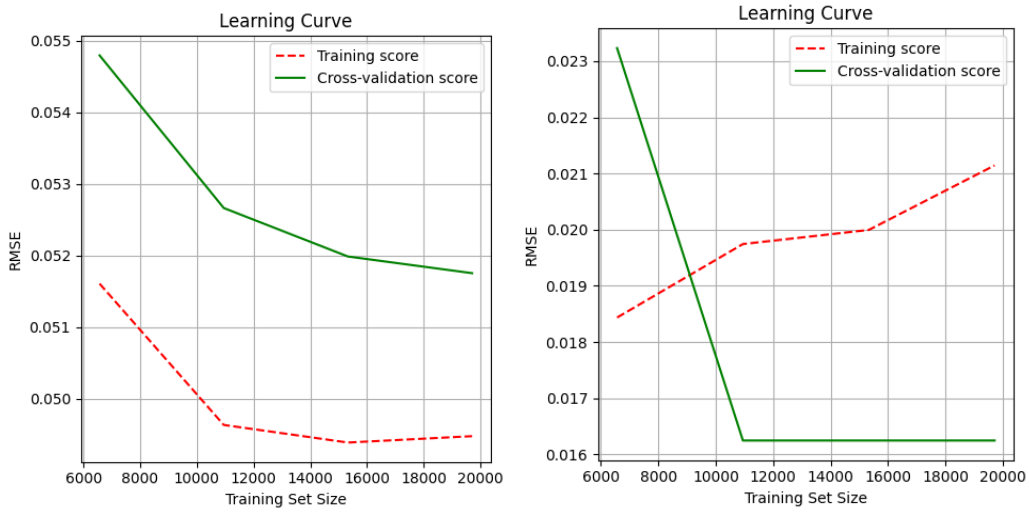
Para los métodos no supervisados se procedió a realizar un PCA, el cual es una función que permite obtener los datos más representativos del dataset con el fin de realizarles una transformación y obtener mejores resultados con el modelo del decision tree. El análisis se realizó a través del siguiente código:

▾ Principal Component Analysis

```
X = Data.drop(columns = ['Gravedad_Accidente', 'NUM_INJ'])
y = Data['Gravedad_Accidente'].values
y = 1.*(y == 'Accidente Leve') + 2.*(y == 'Accidente Moderado') +  3.*(y == 'Accidente Grave')

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')
```

```
(38321, 1) (16424, 1)
RMSE del modelo con  1 elementos:  0.05117
------------
(38321, 3) (16424, 3)
RMSE del modelo con  3 elementos:  0.03490
------------
(38321, 5) (16424, 5)
RMSE del modelo con  5 elementos:  0.02468
------------
Mejor RMSE:  0.02468   ; obtenido con  5  componentes para PCA
```

Resultados, métricas y curvas de aprendizaje.



Las curvas de aprendizaje proporcionadas muestran el desempeño del modelo en términos de RMSE (Root Mean Square Error) tanto para el conjunto de entrenamiento como para la validación cruzada en función del tamaño del conjunto de entrenamiento.

Para la primera gráfica, observamos que el RMSE de entrenamiento aumenta ligeramente con el tamaño del conjunto de entrenamiento, lo que indica que el modelo se ajusta bien a los datos de entrenamiento con conjuntos de menor tamaño y experimenta un leve sobreajuste a medida que aumenta el tamaño del conjunto. Sin embargo, el RMSE de la validación cruzada disminuye y se estabiliza, lo que sugiere que el modelo generaliza bien y mejora su capacidad de hacer predicciones en datos no vistos a medida que dispone de más datos de entrenamiento.

En la segunda gráfica, tanto el RMSE de entrenamiento como el de validación cruzada disminuyen a medida que aumenta el tamaño del conjunto de entrenamiento, lo que indica que el modelo se beneficia de más datos y mejora su rendimiento. El RMSE de validación cruzada es consistentemente mayor que el de entrenamiento, lo que es típico y muestra que el modelo no está sobreajustado.

En resumen, las curvas sugieren que el modelo tiene un buen rendimiento y que proporcionarle más datos de entrenamiento mejora su capacidad de hacer predicciones precisas en ambos conjuntos de datos.

Retos y consideraciones de despliegue

El modelo de predicción de víctimas en accidentes automovilísticos tiene dificultad para predecir con precisión el número de víctimas. Esto se debe a que el dataset actual no incluye toda la información necesaria, como las condiciones climáticas, el estado de la vía o la presencia de señales de tránsito. Para mejorar el modelo, es necesario recopilar más datos y evaluarlo con profesionales de la salud y servicios de emergencia.

El modelo actual no incluye información sobre factores que pueden afectar el número de víctimas, como las condiciones climáticas o el estado de la vía. Para mejorar el modelo, es necesario recopilar datos sobre estos factores. Esto podría implicar costos adicionales, ya que se necesitaría realizar encuestas o recopilar información de fuentes secundarias.

Evaluación con los profesionales de la salud y servicios de emergencia: Es importante evaluar el modelo con profesionales de la salud y servicios de emergencia para determinar si es preciso y útil. Esta evaluación podría realizarse mediante simulaciones o pruebas piloto.

Conclusiones

Para mejorar el rendimiento y reducir el sesgo del modelo de predicción de accidentes de tránsito, es necesario obtener más datos representativos. Estos datos deben ser más completos y diversos para que los modelos puedan capturar mejor la variabilidad de los accidentes. También es importante considerar otros modelos disponibles en el campo de la predicción de accidentes de tránsito, ya que cada modelo tiene sus propias fortalezas y debilidades. Por último, es posible que sea necesario aplicar técnicas de preprocesamiento y manejo de datos para abordar el sesgo presente en los datos.

1. Obtención de más datos representativos: El dataset actual no es representativo de la variabilidad de los accidentes de tránsito. Esto se debe a que los datos actuales están sesgados hacia accidentes con pocas víctimas. Para mejorar el rendimiento del modelo, es necesario obtener más datos que reflejen la realidad de los accidentes de tránsito. Estos datos deben incluir accidentes con muchas víctimas, así como accidentes con pocas víctimas.

2. Selección del modelo: Los tres modelos evaluados inicialmente tienen resultados similares. Sin embargo, es importante considerar 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.

3. Tratamiento del sesgo: El sesgo en los datos 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.

La mejora del modelo de predicción de accidentes de tránsito requiere superar los desafíos de la obtención de datos representativos, la selección del modelo y el tratamiento del sesgo. Sin embargo, un modelo preciso podría ayudar a mejorar la seguridad vial.