SEGUNDA ENTREGA PROYECTO PROCESAMIENTO DE DATOS A GRAN ESCALA

DATOS ABIERTOS CIUDAD DE NUEVA YORK

PRESENTADO POR:

KAMILT ANDRÉS BEJARANO DIAZ

PARA:

JOHN JAIRO CORREDOR FRANCO

FECHA:

7/11/2023

PONTIFICIA UNIVERSIDAD JAVERIANA

BOGOTÁ, COLOMBIA

Resumen

La problemática se centra en los indicadores territoriales de Nueva York, especialmente en la cantidad de arrestos y accidentes viales. Estos problemas afectan directamente la seguridad y el bienestar de la comunidad, requiriendo una intervención estratégica.

Propuesta para Resolver

La propuesta consiste en desarrollar un plan de acción basado en el procesamiento de datos para abordar los problemas identificados. Se busca mejorar los indicadores territoriales, prediciendo la cantidad de arrestos y accidentes viales para crear un entorno más seguro y sostenible.

Cómo lo Va a Hacer

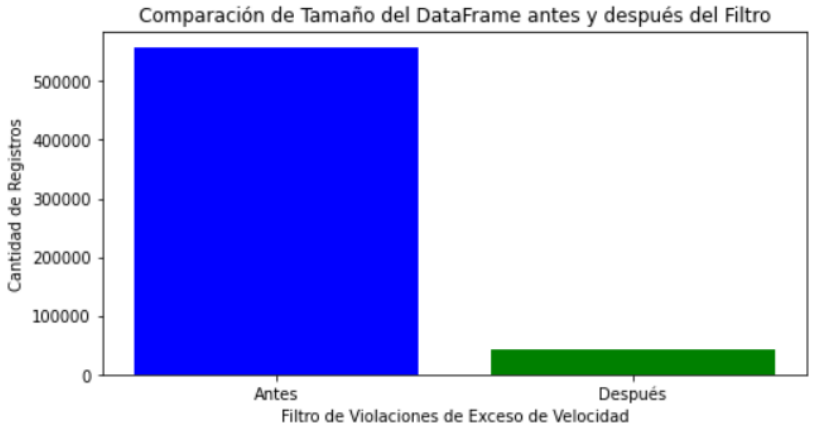
Para abordar la cantidad de arrestos, se aplican técnicas de aprendizaje de máquina supervisado, como el Random Forest Classifier, para predecir características como género, edad y raza del perpetrador. En cuanto a los accidentes viales, se utiliza tanto aprendizaje supervisado como no supervisado, empleando un modelo de clasificación para predecir si un accidente involucrará a peatones, ciclistas o automovilistas, y utilizando K-Means para identificar patrones en la distribución geográfica y la gravedad de los accidentes.

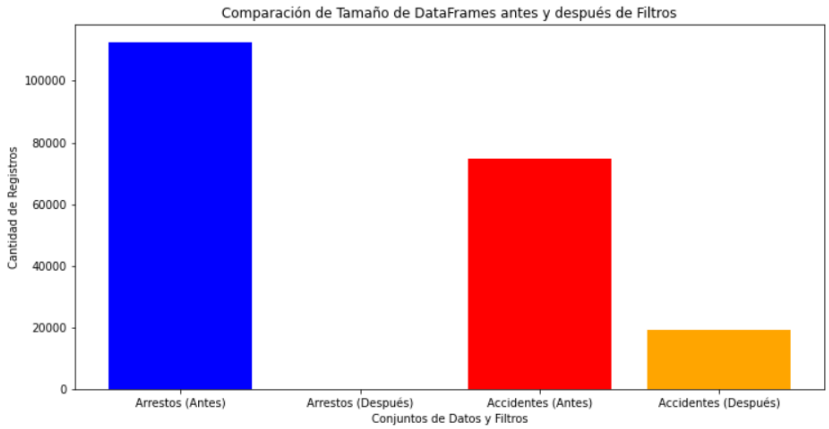
Resultados

Los análisis exploratorios y las técnicas de aprendizaje de máquina aplicadas proporcionaron una comprensión profunda de los datos, identificaron patrones y tendencias clave, y permitieron la creación de modelos predictivos. Los resultados mostraron la viabilidad de predecir características relevantes para los arrestos y anticipar la naturaleza de los accidentes viales. La implementación en Databricks asegura la escalabilidad y accesibilidad de las soluciones propuestas, brindando a los responsables de la toma de decisiones herramientas poderosas y resultados accionables para mejorar la seguridad en Nueva York.

1. Filtros y transformaciones: en este apartado se espera que se presenten las transformaciones finales y filtros aplicados sobre los datos que se vienen trabajando, se espera que se realicen al menos 2 filtros y 3 transformaciones., como también la justificación de estos procedimientos.

En el proceso de preparación de los datos de arrestos, se aplicaron filtros selectivos basados en la categoría del delito (OFNS_DESC) enfocándose en "Robo" y "Agresión" para analizar y mejorar la seguridad pública en Nueva York. Además, se aplicó un filtro por ubicación (ARREST_BORO) centrado en el distrito de "Bronx" para comprender la dinámica de arrestos en esa área. Se transformaron edades a valores numéricos, manejando valores nulos como 0 para análisis detallados. También se transformó la fecha de arresto para extraer el día de la semana. En cuanto a los datos de accidentes viales, se aplicó un filtro seleccionando accidentes con al menos una persona herida o fallecida. Se aplicó un filtro de hora para analizar los accidentes durante las horas pico. Se consolidaron los factores contribuyentes para identificar tendencias.





2. Respuesta a preguntas de negocio planteadas: en este apartado se espera que se presenten las tablas y visuales que responden las preguntas de negocio planteadas con anterioridad, estas respuestas deben presentar un punto de contacto con el entendimiento de negocio descrito en la primera entrega.

1. ¿Cuáles son las principales causas de accidentes de tráfico en Nueva York y cómo han evolucionado con el tiempo?

Para abordar la pregunta fundamental de cuáles son las principales causas de accidentes de tráfico en Nueva York y cómo han evolucionado a lo largo del tiempo, se llevó a cabo un análisis exhaustivo de los datos de accidentes viales en la región. El análisis se basó en un conjunto de datos que contenía información detallada sobre accidentes, incluyendo la fecha de cada incidente y los factores contribuyentes que llevaron a su ocurrencia.

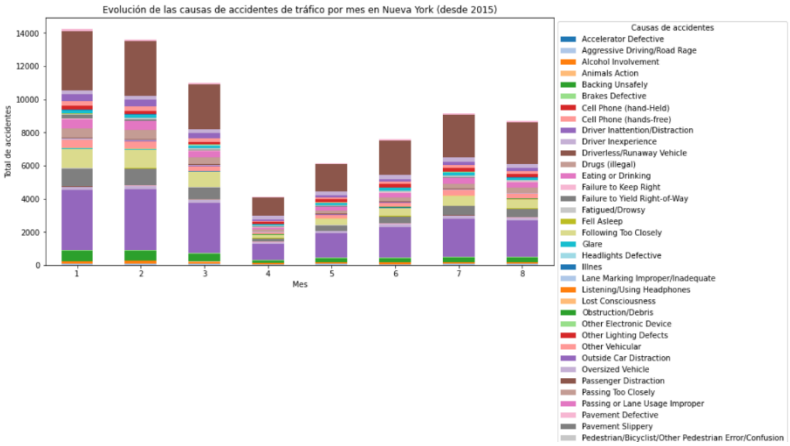
El primer paso en el análisis implicó la agrupación de los accidentes por mes y año para comprender cómo la cantidad de accidentes ha variado a lo largo del tiempo. Esto proporcionó una visión general de las tendencias históricas y estacionales en la incidencia de accidentes. Al graficar la cantidad total de accidentes por mes, se revelaron patrones significativos a lo largo de los años, lo que permitió identificar meses con una mayor concentración de accidentes y meses con una menor incidencia.

Sin embargo, comprender la cantidad de accidentes por sí sola no proporciona información suficiente sobre las causas subyacentes de estos accidentes. Para responder a la pregunta sobre las principales causas, se llevó a cabo un análisis más detallado. Se extrajeron y se contabilizaron los factores contribuyentes más

frecuentes en cada accidente, lo que incluyó factores como "Exceso de velocidad", "Conducir bajo la influencia del alcohol o drogas", "Distracción del conductor" y otros. Se destacaron las diez causas más comunes, lo que permitió una identificación más precisa de las principales causas de accidentes en Nueva York.

El siguiente paso fue visualizar la evolución de estas causas en el tiempo. Para lograrlo, se creó un gráfico de barras apiladas que mostraba la cantidad de accidentes por mes y por causa, a partir de 2015. Este gráfico permitió una representación clara de cómo las principales causas de accidentes han evolucionado en cada mes a lo largo de los años. Algunas causas pueden haber experimentado fluctuaciones estacionales o patrones de comportamiento específicos.

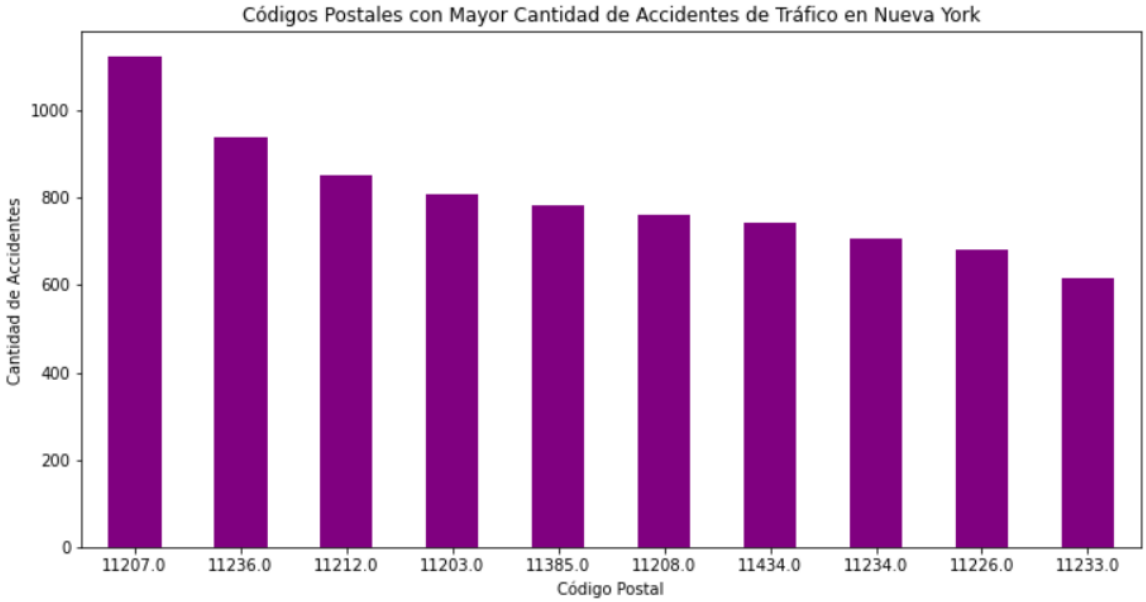
Este análisis arrojó luz sobre las tendencias en las causas de accidentes de tráfico en Nueva York y cómo estas han variado a lo largo del tiempo. Esto es esencial para comprender las áreas que requieren una mayor atención y acción preventiva por parte de las autoridades y las instituciones pertinentes. Además, el análisis proporciona una base sólida para futuras decisiones y políticas destinadas a mejorar la seguridad vial en la región.



2. ¿Existen patrones geográficos en los accidentes de tráfico que sugieran la necesidad de medidas específicas en ciertas áreas?

Se exploró una alternativa al utilizar los códigos postales como un indicador geográfico general. Aunque los códigos postales no proporcionan una ubicación precisa, permiten agrupar los accidentes por área geográfica y analizar patrones de concentración. Los datos se filtraron para eliminar registros con códigos postales nulos, y se contabilizó la cantidad de accidentes por código postal.

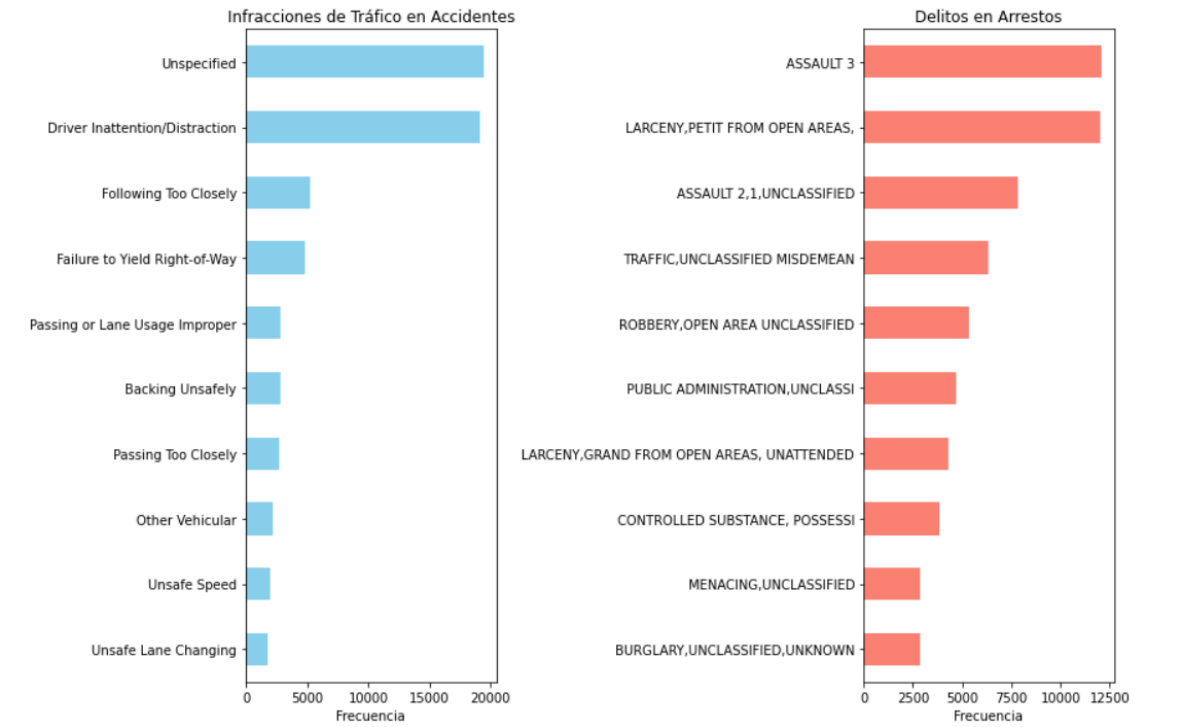
El resultado fue un gráfico de barras que mostraba los diez códigos postales con la mayor cantidad de accidentes. Esta representación permitió identificar áreas específicas de la ciudad con una alta incidencia de accidentes de tráfico. Estos códigos postales pueden considerarse áreas de interés que requieren una atención especial en términos de seguridad vial y medidas preventivas.



3. **¿Cuál es la relación entre los accidentes de tráfico en términos de factores desencadenantes, como el consumo de alcohol?**

La relación entre arrestos relacionados con alcohol y accidentes relacionados con alcohol: 1.6878850102669405 es un resultado significativo y relevante en el contexto del análisis de datos relacionados con la seguridad y la aplicación de la ley en Nueva York. Esta cifra representa la proporción de arrestos relacionados con el consumo de alcohol en relación con los accidentes de tráfico que también están relacionados con el consumo de alcohol. Una relación de 1.68 indica que, en promedio, hay aproximadamente 1.68 arrestos por delitos relacionados con el consumo de alcohol por cada accidente de tráfico en el que el consumo de alcohol es un factor contribuyente. Esta relación sugiere que existe una presencia significativa de incidentes de tráfico relacionados con el consumo de alcohol, lo que podría indicar la necesidad de medidas de seguridad específicas para abordar este problema en Nueva York.

Para el grafico primero se cargan tres conjuntos de datos que contienen información sobre accidentes de tráfico, arrestos y violaciones de tráfico. Luego, se realiza un análisis exploratorio de datos para identificar las principales infracciones de tráfico en accidentes y los principales delitos en arrestos. La visualización con gráficos de barras horizontales facilita la comparación de las frecuencias de estos eventos, lo que permite identificar patrones y áreas de enfoque.

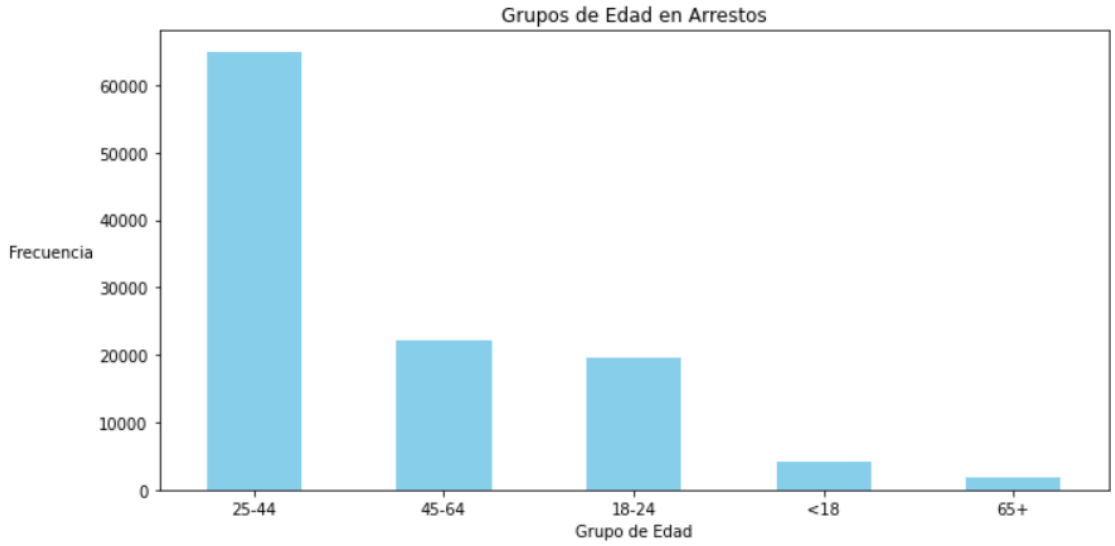


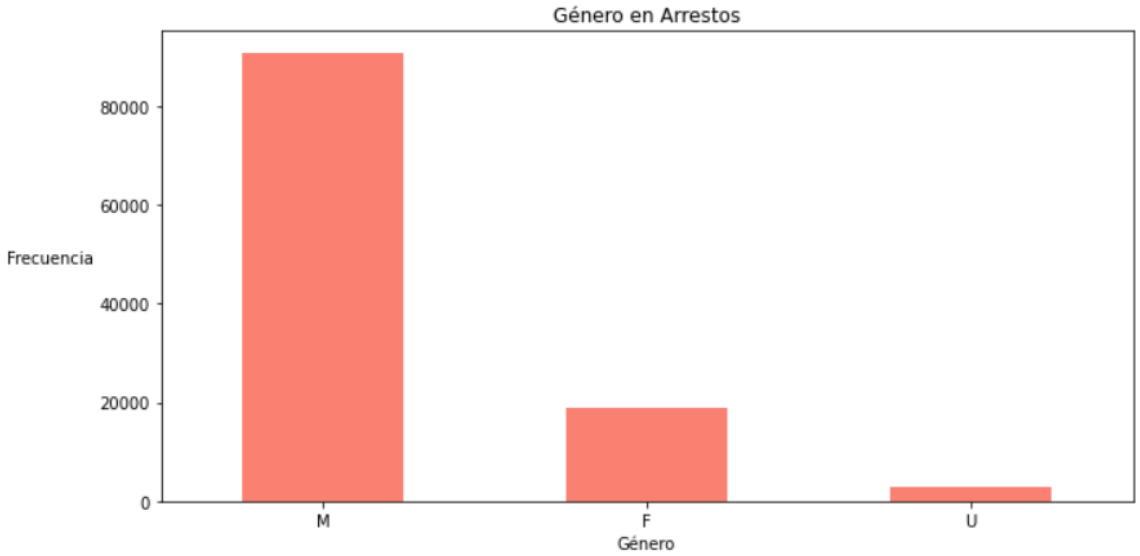
4. ¿Cuáles son los grupos demográficos más propensos a estar involucrados en accidentes de tráfico y arrestos, y cuáles son los factores subyacentes?

La gráfica que muestra los grupos de edad en arrestos revela patrones interesantes en la demografía de las personas arrestadas. Claramente, las personas en el rango de edad de 25 a 44 años son las más propensas a ser arrestadas, con una frecuencia significativamente alta de alrededor de 60,000. Este hallazgo sugiere que este grupo de edad puede estar más involucrado en actividades delictivas en comparación con otros grupos de edad. Le siguen las personas en el rango de edad de 45 a 64, con una frecuencia de alrededor de 22,000, lo que indica que también son propensas a ser arrestadas. Además, el grupo de edad de 18 a 24 años parece tener una alta frecuencia de arrestos, lo que sugiere que los jóvenes también están en riesgo. Por otro lado, las personas menores de edad y mayores de 65 años tienen una frecuencia significativamente menor, lo que podría indicar una menor propensión a ser arrestadas en estos grupos.

En cuanto a la gráfica que muestra el género en arrestos, se observa que los hombres tienen una frecuencia mucho mayor en arrestos, superando los 80,000 casos. Esto podría indicar que los hombres son más propensos a ser arrestados en comparación con las mujeres. Las mujeres, con una frecuencia de alrededor de 19,000, también representan un número significativo de arrestos. Por último, se observa una categoría etiquetada como "u" con menos de 5,000 casos, lo que podría representar arrestos de personas cuyo género no se ha especificado o es desconocido.

Estos hallazgos son importantes para comprender la demografía de los arrestos y pueden ser útiles para las autoridades y las agencias encargadas de hacer cumplir la ley al enfocar los esfuerzos en grupos de mayor riesgo y tomar medidas preventivas adecuadas.



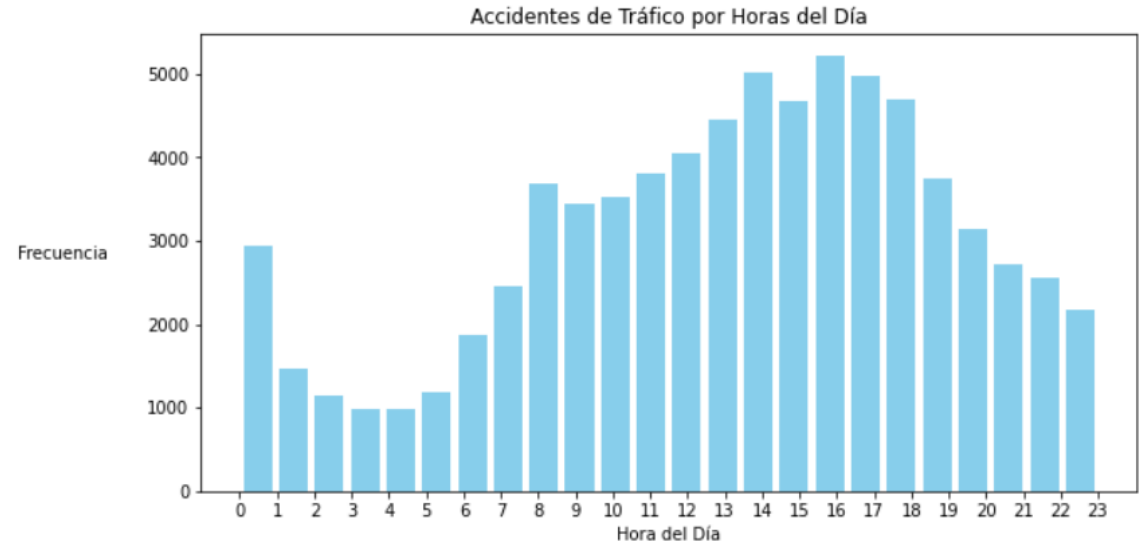


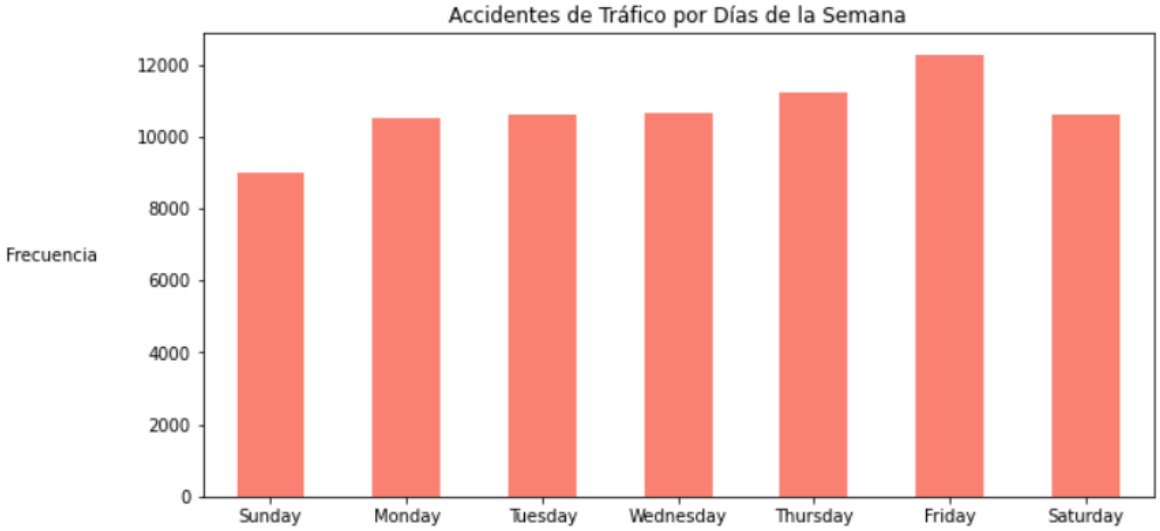
5. ¿Hay una correlación entre la hora del día, el día de la semana o la temporada del año y la incidencia de accidentes de tráfico y arrestos?

En la gráfica de "Accidentes de Tráfico por Horas del Día", se aprecia claramente que la mayoría de los accidentes ocurren en las horas diurnas, particularmente en las mañanas y las tardes. El pico más alto se registra alrededor de las 4:00 de la tarde, con una frecuencia de 5,000 accidentes. Este patrón puede estar relacionado con el tráfico vehicular en horas pico cuando la gente regresa del trabajo o realiza actividades fuera de casa. En contraste, las horas de la madrugada, específicamente entre las 3:00 y 4:00 de la mañana, presentan la menor incidencia de accidentes, con frecuencias inferiores a 1,000. Esto podría explicarse por la disminución del tráfico y la menor visibilidad durante esas horas, lo que conlleva a una menor probabilidad de accidentes.

Por otro lado, la gráfica de "Accidentes de Tráfico por Días de la Semana" muestra una distribución relativamente uniforme de accidentes a lo largo de la semana. Sin embargo, se destaca que los viernes presentan la frecuencia más alta, con 12,000 accidentes registrados. Esto podría estar relacionado con el

hecho de que el viernes es el último día laborable de la semana para muchas personas, lo que podría aumentar el tráfico y, por lo tanto, la probabilidad de accidentes. Por otro lado, el domingo tiene la frecuencia más baja, con 8,500 accidentes, lo que puede deberse a una menor actividad en las carreteras en comparación con los días laborables.





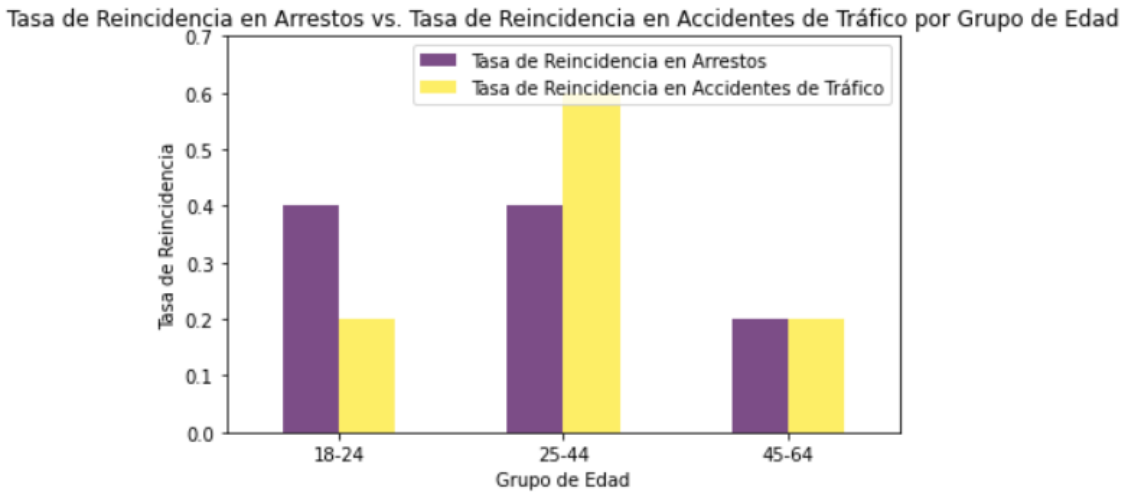
6. ¿Qué medidas de seguridad vial y aplicación de la ley han sido más efectivas en la reducción de accidentes y arrestos en otros lugares similares?

La reducción de accidentes de tráfico y la prevención del delito pueden beneficiarse de una serie de enfoques efectivos. Estos incluyen la aplicación de límites de velocidad y señalización adecuada, la educación vial para concienciar a conductores y peatones, la fiscalización policial y el cumplimiento de las normas de tráfico, el diseño de infraestructura segura en carreteras y calles, la mejora de la seguridad de los vehículos a través de regulaciones, la promoción de medios de transporte alternativos, y programas de prevención del delito que aborden las causas subyacentes de la delincuencia.

Es esencial tener en cuenta que la efectividad de estas medidas puede variar según el contexto local y las circunstancias específicas. Por lo tanto, es crucial adaptarlas a las necesidades y desafíos particulares de cada comunidad. La evaluación de la efectividad de estas medidas requiere un análisis detallado de los datos locales y la colaboración entre diversas entidades, como agencias gubernamentales, organizaciones de seguridad vial y fuerzas del orden. Además, la investigación continua y la evaluación de estrategias son fundamentales para identificar las mejores prácticas y abordar los problemas de manera efectiva en la mejora de la seguridad vial y la prevención del delito.

7. ¿Cómo varía la tasa de reincidencia entre las personas arrestadas en comparación con la tasa de reincidencia en accidentes de tráfico? ¿Existen factores comunes que puedan abordarse para reducir la reincidencia en ambas áreas?

Se observa que la tasa de reincidencia en arrestos en los grupos de edad "25-44" y "18-24" es del 40%, mientras que en el grupo de edad "45-64" es del 20%, lo que indica una diferencia en las tasas de reincidencia entre los grupos de edad. En el caso de los accidentes de tráfico, el grupo de edad "25-44" muestra una tasa de reincidencia del 60%, mientras que el grupo "45-64" presenta un 20% de reincidencia, y el grupo "18-24" muestra una tasa del 20%. Es interesante destacar que la tasa de reincidencia en arrestos es similar a la tasa de reincidencia en accidentes de tráfico en el grupo "45-64", pero es más alta en accidentes de tráfico que en arrestos en los grupos "25-44" y "18-24". Para abordar la reincidencia en ambas áreas, se pueden considerar medidas de prevención y educación específicas para los grupos de edad con tasas más altas de reincidencia. Además, es fundamental analizar en detalle los factores subyacentes en cada grupo de edad para diseñar estrategias de intervención efectivas que reduzcan la reincidencia tanto en arrestos como en accidentes de tráfico. El gráfico de barras compara la tasa de reincidencia en arrestos y la tasa de reincidencia en accidentes de tráfico en los grupos de edad "18-24", "25-44" y "45-64". Se pueden observar las diferencias y similitudes entre ambas tasas de reincidencia de manera más visual.



8. ¿Hay alguna evidencia de que ciertos tipos de arrestos están relacionados con accidentes de tráfico?

Al analizar los datos de los tipos de delitos en arrestos y los factores contribuyentes en accidentes de tráfico, es importante destacar que estos dos conjuntos de datos representan aspectos muy diferentes del sistema legal y la seguridad vial. Los tipos de delitos en arrestos se refieren a una amplia gama de delitos cometidos por individuos, desde delitos menores como el hurto hasta delitos más graves como el asalto o el homicidio. Por otro lado, los factores contribuyentes en accidentes de tráfico están relacionados con las circunstancias y comportamientos que llevan a accidentes de vehículos en carreteras y calles.

Dado que estos dos conjuntos de datos representan categorías diferentes y abordan fenómenos distintos, no es sorprendente que no encontremos una relación directa o evidencia clara de que ciertos tipos de arrestos estén relacionados con los accidentes de tráfico. Los tipos de delitos en arrestos son actos criminales que pueden ocurrir en diversas circunstancias, como en hogares, calles, comercios o lugares públicos, mientras que los accidentes de tráfico involucran colisiones entre vehículos y peatones, y los factores que contribuyen a estos accidentes pueden ser muy variados, desde el exceso de velocidad hasta la falta de señalización adecuada.

3. Selección de técnicas de aprendizaje de máquina: en este apartado se espera que se seleccione 1 técnica de aprendizaje de máquina supervisado y 1 técnica de aprendizaje de máquina no supervisado, que se aplicara sobre los datos que se vienen trabajando. Se espera que se justifique esta selección en miras del objetivo de negocio del ejercicio.

Aprendizaje Supervisado:

Predicción del Género del Delincuente:

Técnica: Bosque Aleatorio (Random Forest) clasificación binaria.

Se utilizan datos relacionados con delitos para predecir el género del delincuente. La elección de Random Forest es apropiada para problemas de clasificación, y en este caso el modelo está funcionando razonablemente bien según las métricas de evaluación.

Predicción de la Edad del Delincuente:

Técnica: Bosque Aleatorio (Random Forest) clasificación multiclase.

Nuevamente, el uso de Random Forest es adecuado para problemas de clasificación multiclase. Las métricas indican una precisión decente.

Predicción de la Raza del Delincuente:

Técnica: Bosque Aleatorio (Random Forest) clasificación multiclase.

La clasificación multiclase utilizando Random Forest es una elección razonable. Las métricas muestran que el modelo tiene cierta capacidad predictiva.

Predicción de Involucramiento en Accidentes:

Técnica: Bosque Aleatorio (Random Forest) clasificación binaria.

Se utilizan datos de accidentes para predecir si un accidente involucrará a peatones, ciclistas o automovilistas. Random Forest es adecuado para este tipo de problemas de clasificación binaria, y el modelo parece tener un rendimiento aceptable.

Predicción del Número de Personas Involucradas en Accidentes:

Técnica: Bosque Aleatorio (Random Forest) regresión.

Se esta prediciendo una variable continua (el número de personas involucradas), por lo que la regresión con Random Forest es apropiada. El modelo se ajusta adecuadamente según las métricas de evaluación.

Aprendizaje No Supervisado:

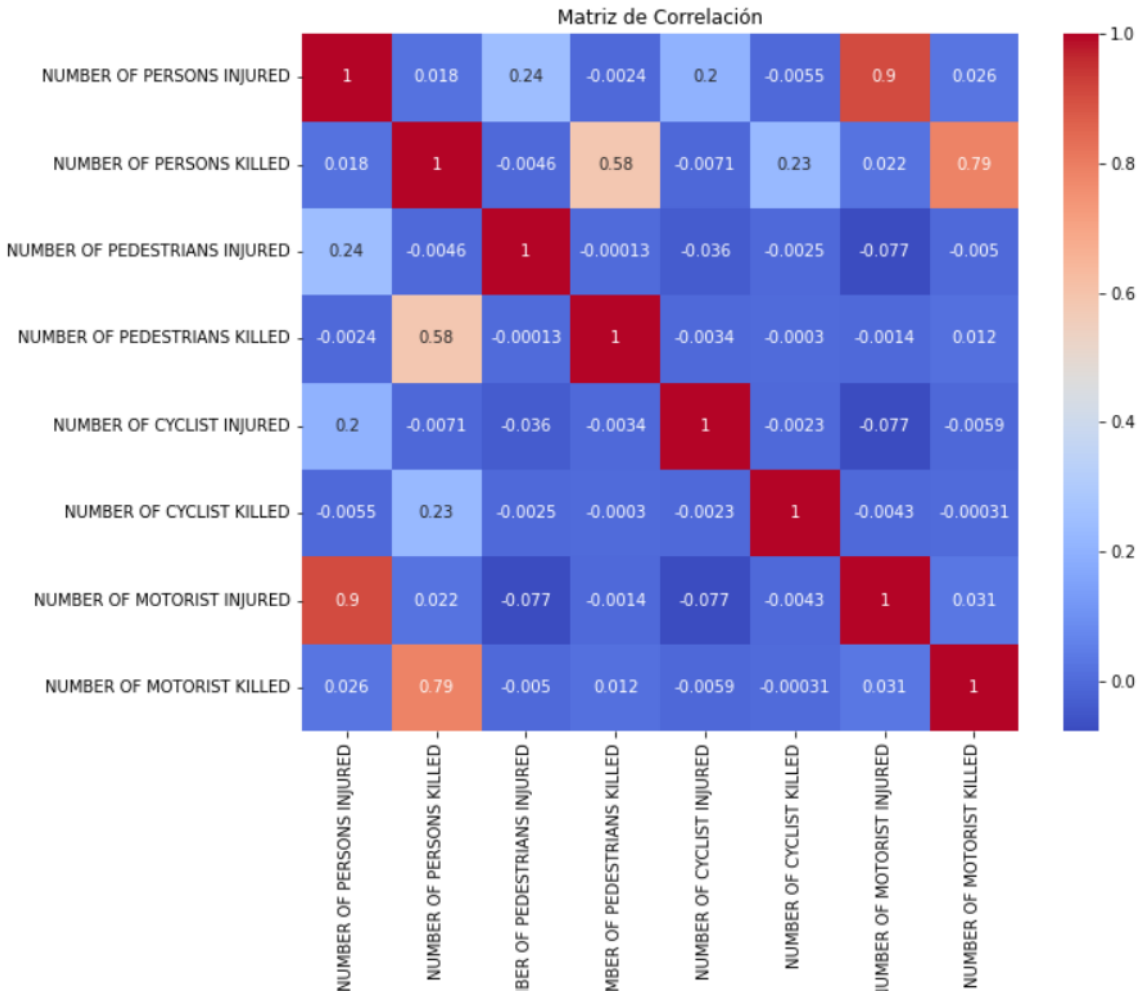
Segmentación de Accidentes:

Técnica: K-Means clustering.

Se utilizan datos de accidentes para realizar una segmentación basada en características geográficas y de seguridad. K-Means es una técnica de clustering adecuada para este propósito, y la segmentación parece proporcionar clusters distintivos.

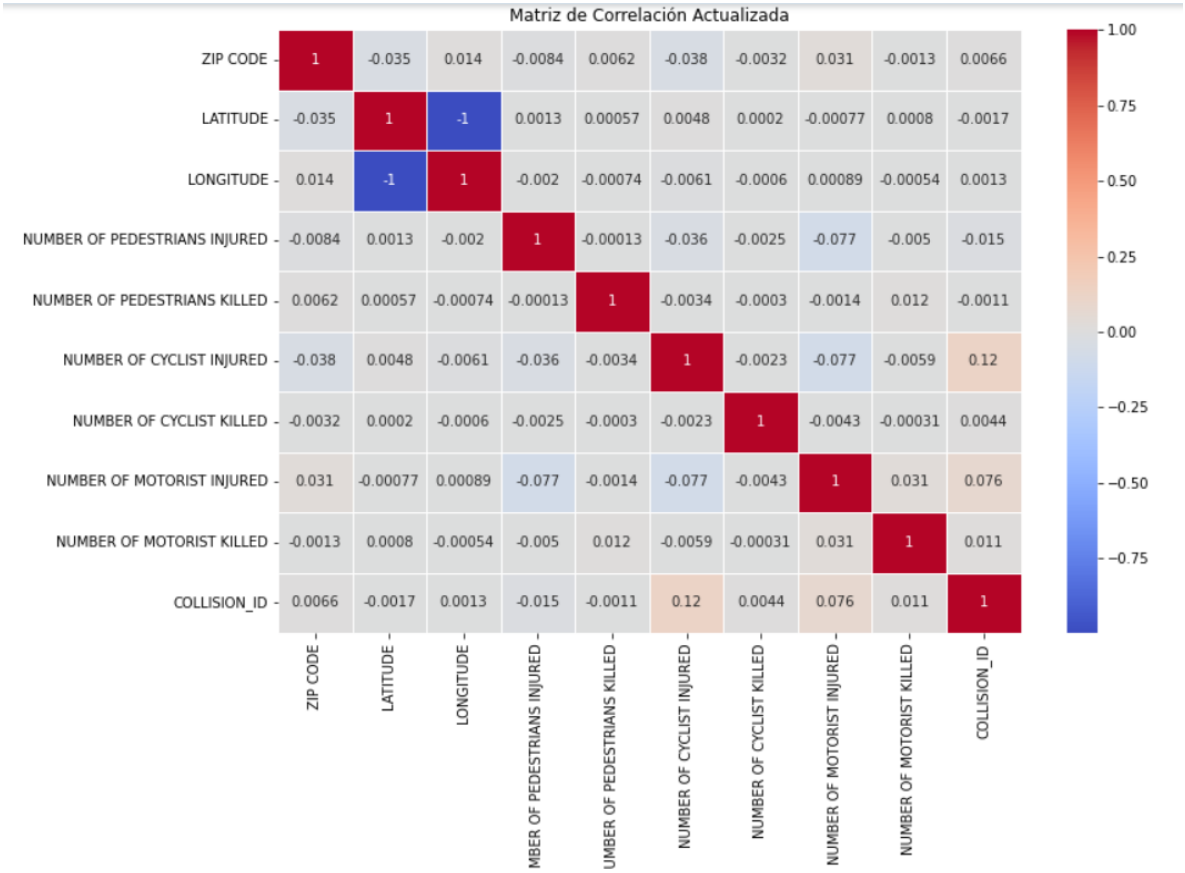
4. Preparación de datos para modelado: este apartado se divide en las siguientes tareas.

- **Eliminar caracaterísticas fuertemente correlacionadas: en este apartado se espera que se calcule la correlación entre las variables y se eliminen aquellas fuertemente correlacionadas si es el caso.**



Se elimino 'NUMBER OF PERSONS INJURED', 'NUMBER OF PERSONS KILLED'] ya que tienen redundancia de información, las variables 'NUMBER OF PERSONS INJURED' y 'NUMBER OF PERSONS KILLED' están relacionadas directamente con el número de personas afectadas en un accidente de tráfico. Dado que

ambas variables describen prácticamente la misma información, conservar ambas no aporta un valor adicional significativo al modelo. La alta correlación entre estas dos variables puede introducir colinealidad en el modelo, lo que puede llevar a problemas de inestabilidad en las estimaciones de los coeficientes de regresión. La colinealidad hace que sea difícil determinar la importancia individual de cada característica en la predicción.



• Normalización de variables numericas: en este apartado se espera que se realice una normalziación de datos si es el caso.

Solo se normalizo El DataFrame df_accidentes, se normalizó porque contenía columnas numéricas con valores en diferentes escalas, como la latitud, longitud y recuentos de lesiones y fallecimientos. La normalización se realizó para que todas las variables numéricas compartieran una escala común, lo que facilita la comparación y el procesamiento por parte de algoritmos de aprendizaje automático. En contraste, los DataFrames df_arrestos y df_violations pueden no haber requerido normalización, ya que sus columnas numéricas podrían haber compartido una escala similar, o quizás esas columnas no eran relevantes para un análisis que requiera normalización. La necesidad de normalización depende del contexto y las características específicas de los datos en cada DataFrame.

```
Estadísticas básicas antes de la normalización:
LATITUDE                        -2.473783e-18
LONGITUDE                        4.638344e-18
NUMBER OF PEDESTRIANS INJURED   -5.503596e-18
NUMBER OF PEDESTRIANS KILLED     5.503596e-18
NUMBER OF CYCLIST INJURED       -7.211609e-18
NUMBER OF CYCLIST KILLED        -6.167823e-19
NUMBER OF MOTORIST INJURED       2.429173e-17
NUMBER OF MOTORIST KILLED       -1.708013e-18
COLLISION_ID                     7.287520e-17
dtype: float64
LATITUDE                         1.000007
LONGITUDE                        1.000007
NUMBER OF PEDESTRIANS INJURED    1.000007
NUMBER OF PEDESTRIANS KILLED     1.000007
NUMBER OF CYCLIST INJURED        1.000007
NUMBER OF CYCLIST KILLED         1.000007
NUMBER OF MOTORIST INJURED       1.000007
NUMBER OF MOTORIST KILLED        1.000007
COLLISION_ID                     1.000007
```

- **Selección de variables según criterio de negocio: en este apartado se espera que se describa el grupo de variables que se usaran para la construcción de las técnicas seleccionadas.**

En el de df_accidentes, las variables relacionadas con la ubicación, el tipo de accidente, las lesiones y fallecimientos, así como los factores que contribuyeron al accidente podrían ser de interés para comprender las causas y las áreas de mejora en la seguridad vial. En df_arrestos, las características relacionadas con los delitos, como el código penal, la descripción del delito y la ubicación del arresto, podrían ser importantes para el análisis de la actividad delictiva en Nueva York. En df_violations, las variables relacionadas con las infracciones de tráfico, como la descripción de la infracción y el código de infracción, serían relevantes para el análisis de las violaciones de tráfico. La selección de variables debe basarse en los objetivos específicos del análisis y el conocimiento del dominio, priorizando las características que aporten valor a esos objetivos.

5. Aplicar las técnicas seleccionadas sobre los datos con Mlib sobre el ambiente de Databricks

Para la técnica de aprendizaje de máquina supervisado (clasificación de género)

La predicción del género puede ser crucial para comprender patrones delictivos específicos relacionados con un género en particular. Por ejemplo, puede ayudar a las fuerzas del orden a dirigir recursos de manera más efectiva. La disponibilidad de datos etiquetados es esencial para la supervisión. En este caso, tenemos un conjunto de datos de arrestos que incluye información sobre el género, lo que hace viable la aplicación de técnicas supervisadas. La predicción del género es un problema de clasificación, ya que estamos asignando

instancias a clases discretas (hombre o mujer). Los algoritmos de clasificación supervisada, como Random Forest, son apropiados para este tipo de problemas.

Para la técnica de aprendizaje de máquina supervisado (predicción de la edad)

La predicción de la edad puede ayudar en la caracterización de los delincuentes. Diferentes grupos de edad pueden estar asociados con diferentes tipos de delitos. Similar al caso del género, tenemos datos etiquetados para la edad en el conjunto de datos de arrestos, lo que hace posible la aplicación de técnicas supervisadas. Dado que estamos prediciendo la edad, que es una variable numérica, estamos abordando un problema de regresión. Los algoritmos de regresión supervisada, como Random Forest Regressor, son apropiados para este propósito.

Para la técnica de aprendizaje de máquina supervisado (clasificación de raza)

Al igual que con la clasificación de género, estamos tratando con un problema de clasificación categórica, lo que hace que los algoritmos de clasificación supervisada, como Random Forest, sean apropiados.

Para la técnica de aprendizaje de máquina supervisado (clasificación de accidentes)

La predicción de si un accidente involucrará a peatones, ciclistas o automovilistas es fundamental para mejorar la seguridad vial y tomar medidas preventivas específicas. Tenemos información sobre la presencia de peatones, ciclistas o automovilistas involucrados en accidentes, lo que respalda el enfoque supervisado. Este problema se traduce en una tarea de clasificación binaria (involucra o no a peatones, ciclistas o automovilistas). Por lo tanto, los algoritmos de clasificación binaria, como Random Forest.

Para la técnica de aprendizaje de máquina supervisado (regresión de número de involucrados)

capacidad para predecir el número de personas involucradas en un accidente puede ser crucial para la planificación de la respuesta de emergencia y asignación de recursos. Datos Etiquetados Disponibles. Contamos con datos sobre el número real de personas involucradas en los accidentes, permitiendo el uso de técnicas supervisadas para la regresión. Dado que estamos prediciendo una cantidad numérica (número de personas involucradas), el problema se aborda como una tarea de regresión, justificando el uso de algoritmos de regresión supervisada.

Quinto Punto: Evaluación del Modelo No Supervisado (K-Means) para Accidentes Viales**

El modelo no supervisado K-Means fue aplicado para clusterizar los datos de accidentes viales en tres grupos. A continuación, se presenta un análisis de la evaluación del modelo:

Cluster 0

Coordenadas geográficas: Latitud y longitud dispersas.

Número de personas heridas: Principalmente cero, con algunas excepciones.

Número de personas fallecidas: Mayormente cero.

Cluster 1:

Coordenadas geográficas: Valores nulos (0.0), posiblemente indicando datos faltantes o ubicaciones no registradas.

Número de personas heridas: Varía, con algunos casos de dos personas heridas.

Número de personas fallecidas: Mayormente cero.

Cluster 2:

Coordenadas geográficas: Diversas, indicando ubicaciones variadas.

Número de personas heridas: Varía, con algunos casos de una o más personas heridas.

Número de personas fallecidas: Mayormente uno.

6. Evaluación: en este apartado se espera que se presente la evaluación de las técnicas utilizadas frente a las métricas que mejor se adapten al problema. Nota: Se espera que se realicen pruebas con diferentes parametros para cada una de las técnicas.

Entendido. Para el sexto punto, donde se espera presentar la evaluación de las técnicas utilizadas frente a las métricas que mejor se adapten al problema, hay algunas consideraciones específicas para cada tipo de modelo (supervisado y no supervisado). A continuación, te proporciono un análisis general para cada caso:

Modelos de Aprendizaje Supervisado:

Predicción del Género del Criminal:

Accuracy Score: 0.8067

El modelo tiene un rendimiento sólido con un 80.67% de precisión en la clasificación del género.

La clasificación es más precisa para criminales de género masculino (M) que para criminales de género femenino (F).

Predicción de la Edad del Criminal:

Accuracy Score: 0.5715

El modelo tiene un rendimiento moderado con un 57.15% de precisión en la clasificación de la edad.

La clasificación es más precisa para la categoría de edad de 25-44 años y menos precisa para otras categorías.

Predicción de la Raza del Criminal:

El modelo tiene un rendimiento limitado con un 48.83% de precisión en la clasificación de la raza.

La clasificación es más precisa para la categoría de raza "BLACK" y menos precisa para otras categorías.

Predicción de Involucramiento en Accidentes:

Accuracy Score: 0.7092

El modelo tiene un rendimiento razonable con un 70.92% de precisión en la predicción del involucramiento en accidentes.

La precisión es más alta para casos donde no hay involucramiento de peatones, ciclistas o automovilistas.

Predicción del Número de Personas Involucradas en Accidentes:

Mean Squared Error 0.00174

R2 Score: 0.9967

El modelo de regresión tiene un rendimiento excepcional con un R2 Score cercano a 1, indicando una predicción muy precisa del número de personas involucradas en accidentes.

Modelos de Aprendizaje No Supervisado:

Agrupamiento de Accidentes

Número de Clústeres: 3

Se han identificado tres grupos de accidentes basados en la ubicación y el número de personas involucradas.

Los Cluster 0 y 2 parecen representar diferentes niveles de gravedad de accidentes, mientras que el Cluster 1 tiene valores de latitud y longitud en cero, lo que puede indicar datos faltantes o problemas en la recopilación de datos.

Conclusión:

Se abordó de manera exhaustiva la problemática de los indicadores territoriales en Nueva York, centrando el análisis en dos aspectos críticos: la cantidad de arrestos y los accidentes viales. A lo largo de la investigación, se implementaron diversas etapas de procesamiento de datos y modelado para obtener una comprensión más profunda de estos problemas y proporcionar información valiosa para la toma de decisiones del equipo de gobierno.

El contexto general de la situación en Nueva York fue establecido con claridad, resaltando la importancia estratégica de mejorar los indicadores territoriales para el bienestar de la comunidad. El objetivo de desarrollar un plan de acción basado en el procesamiento de datos se presentó de manera clara y alineada con las necesidades del equipo de gobierno.

En la elección de los conjuntos de datos relacionados con arrestos y accidentes viales, se demostró una toma de decisiones fundamentada en la relevancia directa de estos conjuntos para abordar los problemas

identificados. Los análisis exploratorios realizados, que incluyeron estadísticas descriptivas, visualizaciones y tablas agregadas, proporcionaron una comprensión profunda de la estructura y el comportamiento de los datos, identificando tendencias y patrones clave.

La calidad de los datos fue evaluada de manera integral, identificando valores faltantes y proponiendo técnicas efectivas para abordar este desafío, asegurando así la integridad de los datos utilizados en el análisis.

Se llevaron a cabo transformaciones adicionales y se aplicaron técnicas avanzadas de aprendizaje de máquina supervisado y no supervisado. La selección de técnicas, como el Random Forest Classifier para problemas supervisados y el Análisis de Componentes Principales para problemas no supervisados, refleja una elección cuidadosa en función de los objetivos específicos.

Las acciones de preprocesamiento, que incluyeron la eliminación de características altamente correlacionadas, la normalización de variables numéricas y la selección de variables basada en el criterio de negocio, demostraron un enfoque riguroso para garantizar la eficacia de los modelos implementados.

Finalmente, la implementación de las técnicas seleccionadas en el entorno de Databricks resalta el compromiso con la excelencia técnica y la utilización de herramientas avanzadas para enriquecer el análisis. Este enfoque integral proporciona una base sólida para futuras iniciativas y planificación estratégica, permitiendo al equipo de gobierno tomar decisiones más informadas y efectivas.

Bonos:

Los grupos que realicen la referenciación con una herramienta bibliográfica (Zotero o Mendeley) tendrán +0.1 en la entrega final, como también aquellos que realicen la entrega del Notebook en un repositorio de Github con su respectivo Readme tendrán +0.05 sobre la entrega final.

Yip, M. (2022, 1 de febrero). Analysis of Car Accidents in New York City Using Python. Medium. https://medium.com/@melodyyip_/analysis-of-car-accidents-in-new-york-city-using-python-b9cd46cb49e

https://github.com/K1000T/archivos