

**PROYECTO ENTREGA 2**

**Federico García Gutiérrez**

**Juan José Gómez Arenas**

**Procesamiento de Datos a Gran Escala**

**PONTIFICIA UNIVERSIDAD JAVERIANA**

**BOGOTÁ D.C**

**2023**

**RESUMEN**

**Contexto**

Nueva York, uno de los estados más grandes y poblados de los Estados Unidos, se encuentra en la región noreste del país. Esta diversidad geográfica y demográfica, que abarca desde áreas urbanas altamente desarrolladas hasta zonas rurales, presenta desafíos para la administración. La economía del estado, con un enfoque particular en la ciudad de Nueva York, es un motor importante para la economía nacional, destacándose en finanzas, negocios y cultura.

**Objetivo**

El estado de Nueva York ha contratado a nuestro equipo de consultoría con un objetivo; utilizar el procesamiento de datos para desarrollar un plan de acción que reduzca la cantidad de arrestos y accidentes viales en el estado. Estos indicadores se han identificado como áreas de preocupación prioritaria, ya que impactan directamente en la seguridad y la calidad de vida de los ciudadanos. Nuestra misión es analizar los datos disponibles, identificar patrones, factores desencadenantes y áreas críticas, y proponer medidas efectivas para mejorar estos indicadores territoriales. Este proyecto tiene como objetivo contribuir a la mejora de la seguridad y el bienestar de la población de Nueva York.

**Resultados**

En el estudio de modelos de aprendizaje de máquina orientados a la predicción de lugares con mayor incidencia de arrestos en Nueva York, se exploraron diversas condiciones relacionadas con accidentes de tráfico. La investigación se enfocó en la capacidad predictiva de estas variables para determinar tanto el número de personas heridas como las posibles afectaciones individuales en caso de un accidente. Mediante el análisis de datos detallados sobre accidentes viales en la ciudad, se identificaron patrones y correlaciones significativas entre factores como la ubicación específica del incidente, las condiciones meteorológicas, la hora del día y la severidad de las lesiones. Estos resultados proporcionan valiosa información para optimizar estrategias de seguridad vial y asignar recursos de manera más efectiva en función de las condiciones particulares de cada área en Nueva York, contribuyendo así a la mejora de la prevención y respuesta en casos de accidentes de tráfico.

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.

* **Dataset Arrestos Históricos NYPD**

El conjunto de datos no tenía un número muy elevado de datos faltantes, donde únicamente el 0.644% de los datos eran faltantes. Por esta razón, se eliminaron todos los registros que tenían algún dato faltante.

No obstante, es posible simplificar el set de datos para sintetizar la información para el análisis. Teniendo en cuenta que Nueva York se compone de 5 regiones principales, los cuales son

* The Bronx
* Brooklyn
* Manhattan
* Queens
* Staten Islands

Para términos prácticos, se va a estandarizar poniendo el nombre completo en la columna de ARREST\_BORO, de manera que se pueda saber en qué sector se realizó, y se eliminara la información relacionada con la ubicación geográfica de arrestos. Por lo cual se omitirán las columnas de 'ARREST\_KEY','PD\_CD','PD\_DESC', 'LAW\_CODE','KY\_CD','X\_COORD\_CD', 'Y\_COORD\_CD', 'Latitude', 'Longitude', 'Lon\_Lat'. Teniendo en cuenta que un número elevado de ellas hacen referencia a la posición geográfica donde ocurrió el arresto, mientras que las otras hacen referencia a llaves para el caso, en donde en realidad no nos interesa.

Por otra parte, para tener más claridad en la jurisdicción y no tener que recurrir a códigos para comprender los datos se reemplazan con lo que corresponden, esto corresponde a la columna JURISDICTION\_CODE. Donde:

1. Patrol (NYPD)
2. Transit (NYPD)
3. Housing (NYPD)
4. > NO-NYPD

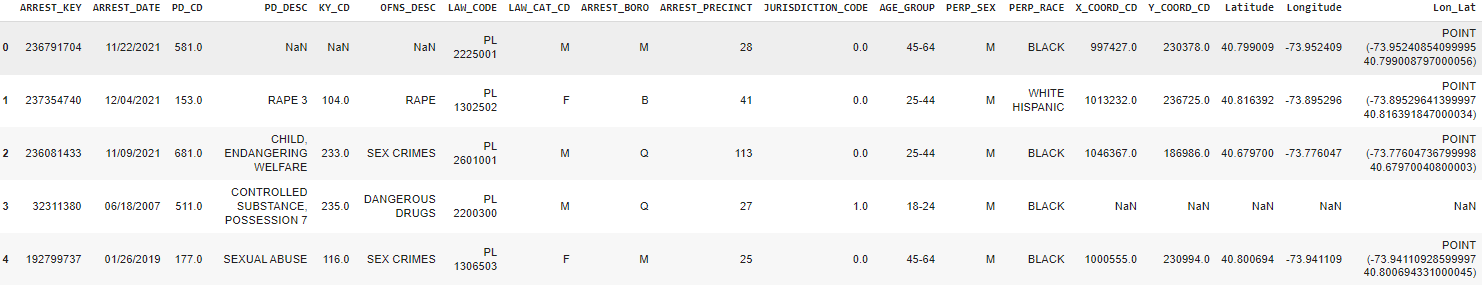
En este caso los códigos hacen referencia a la jurisdicción del arresto y si fue realizado por el departamento de policía de Nueva York o si fue realizada fuera de la jurisdicción del departamento de policía.

Además, también se estandarizaron las columnas LAW\_CAT\_CD y PERP\_SEX para que contengan la palabra que representan y no únicamente la letra inicial. De manera que quede estandarizado todo el set de datos y sea fácil de interpretar sin necesidad de recurrir a un diccionario de datos, donde   
LAW\_CAT\_CD:

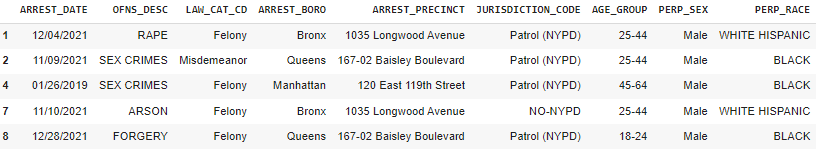
F – Felony

M – Misdemeanor

V – Violation   
Y finalmente, con respecto al ARREST\_PRECINCT, para tener claridad a que precinto corresponde el arresto explícitamente y no a su id numérico se reemplazó de manera que quede estandarizado para todo el dataset con los valores correspondientes. Finalmente, se removieron algunos datos atípicos teniendo en cuenta el grupo de edad, donde existían valores como 963 o 342, de esta manera, se cambió el valor a Unknown, un valor que ya existía dentro del conjunto de datos originalmente.

De manera que la tabla resultante paso de verse así   


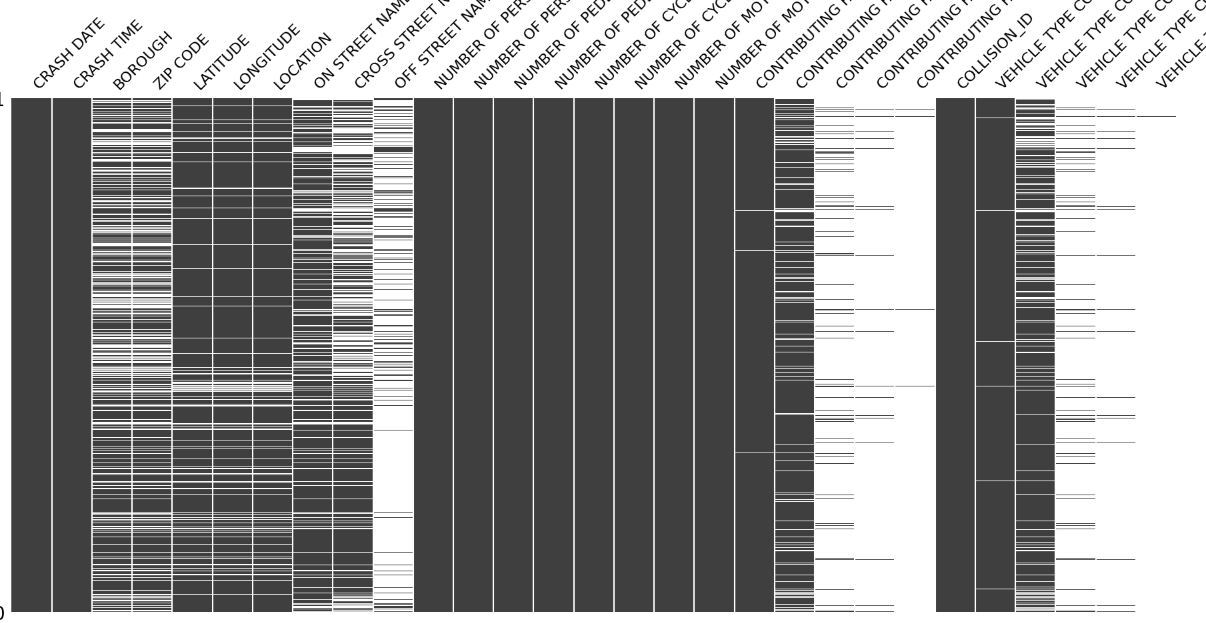
*Fig. 1.1 Arrestos Históricos antes de transformaciones y filtros.*

A verse así   


*Fig. 1.2 Arrestos Históricos después de transformaciones y filtros.*

* **Dataset Colisiones (Choques)**

El conjunto de datos cuenta con un número elevado de datos faltantes, teniendo en cuenta que muchas columnas están condicionadas a si un choque involucro más de un vehículo. El número de datos faltantes se puede apreciar mejor en la siguiente figura, donde cada espacio en blanco representa un dato faltante;

 *Fig. 1.3 Valores faltantes de Colisiones (Choques)*

Teniendo en cuenta estas columnas condicionadas, se decidió simplificar para que se cree una columna adicional y de esta manera se puedan saber cuántos vehículos estuvieron involucrados en el choque, omitiendo todas las columnas relacionadas con el tipo de vehículo al igual que los factores que contribuyeron para cada vehículo, pues, su número de valores nulos es bastante elevado, y si no lo es, únicamente es ‘Unspecified’, correspondiendo a alrededor de 700 mil registros. De manera que se borraron las columnas CONTRIBUTING FACTOR VEHICLE y también las de VEHICLE TYPE CODE.

Por otra parte, para simplificar la ubicación de los choques, se opto por crear una función donde dadas las coordenadas de latitud y longitud se pueda identificar el ‘BOUROUGH’ en donde ocurrió, es decir, en cuál de las grandes 5 regiones de NY ocurrió.

Otro de los grandes problemas del dataset es que en muchos de sus campos de calles es posible que no tenga datos, de manera que se decidió simplificar el ON STREET, CROSS STREET y OFF STREET a una única columna que se llame Reference Street, donde se tomara una calle como referencia para saber en donde ocurrió el choque. Por lo cual se priorizará la información de ON STREET, luego, si no hay un campo ahí se procede a la información de CROSS STREET y finalmente se muestra la información de OFF STREET, si no hay información en ninguna, se muestra que se desconoce en qué calle ocurrió.

Tras esto se borran las columnas que no aportan como los IDs, las calles por lo explicado anteriormente y todas las columnas que contaban con un número significativo de datos faltantes. Resultando con esta matriz donde los puntos blancos representan datos faltantes

Imagen que contiene Patrón de fondo

Descripción generada automáticamente*Fig. 1.4 Valores faltantes después de transformación Colisiones (Choques)*

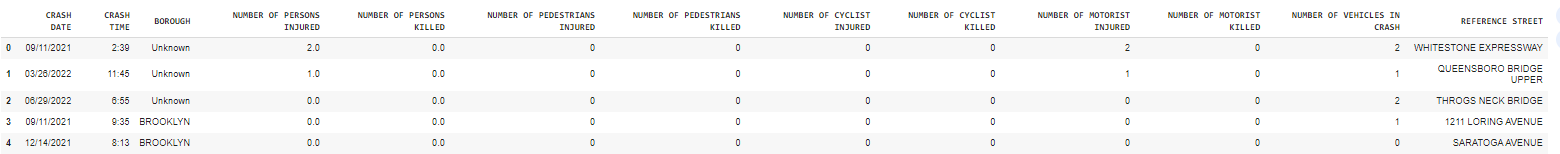
De manera que la tabla resultante paso de

Captura de pantalla de computadora

Descripción generada automáticamente con confianza media

*Fig. 1.5 Colisiones (Choques) antes de transformaciones y filtros.*

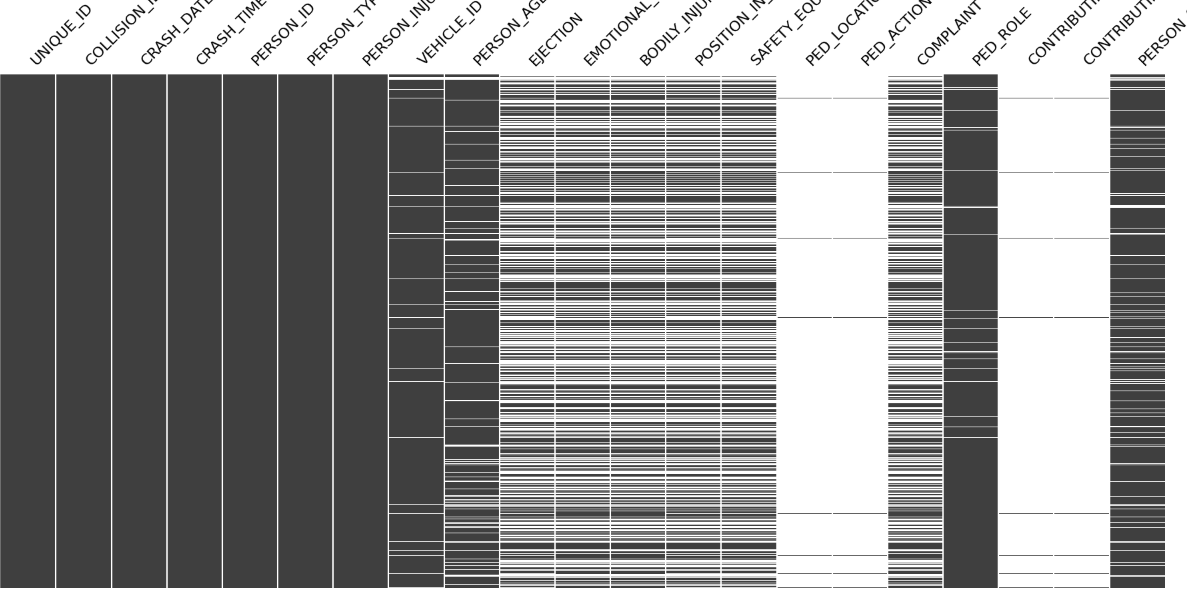
A esta



*Fig. 1.6 Colisiones (Choques) después de transformaciones y filtros.*

* **Dataset Colisiones (Personas)**

Al igual que el Dataset anterior, este tiene un número elevado de datos vacíos. Esto se puede evidenciar en la siguiente matriz, donde cada espacio en blanco representa algún dato faltante

 *Fig. 1.7 Valores faltantes de Colisiones (Personas)*

Teniendo el gran número de datos faltantes en algunas columnas, y que estas no contribuyen en gran medida al análisis que se desea realizar, por lo cual columnas de 'UNIQUE\_ID', 'COLLISION\_ID', 'PERSON\_ID', 'VEHICLE\_ID', 'EJECTION', 'EMOTIONAL\_STATUS', 'PED\_LOCATION', 'PED\_ACTION', 'CONTRIBUTING\_FACTOR\_1', 'PED\_ROLE', 'CONTRIBUTING\_FACTOR\_2' se eliminaran.

Por otra parte, en algunas columnas se tienen datos que entre sus parámetros se tiene el valor ‘Unknown’, pero que pueden ser relevantes para un posible análisis, por lo cual se opto por reemplazar estos valores en las columnas y finalmente por estandarizar la columna referente al sexo de la persona, para que corresponda con las tablas anteriores. De manera que la tabla paso de    
Interfaz de usuario gráfica, Texto, Aplicación, Chat o mensaje de texto

Descripción generada automáticamente

*Fig. 1.8 Colisiones (Personas) antes de transformaciones y filtros.*

A la tabla resultante

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente *Fig. 1.9 Colisiones (Personas) después de transformaciones y filtros.*

1. **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.

El estado de Nueva York ha contratado a nuestro equipo de consultoría con un objetivo; utilizar el procesamiento de datos para desarrollar un plan de acción que reduzca la cantidad de arrestos y accidentes viales en el estado. Estos indicadores se han identificado como áreas de preocupación prioritaria, ya que impactan directamente en la seguridad y la calidad de vida de los ciudadanos. Nuestra misión es analizar los datos disponibles, identificar patrones, factores desencadenantes y áreas críticas, y proponer medidas efectivas para mejorar estos indicadores territoriales. Este proyecto tiene como objetivo contribuir a la mejora de la seguridad y el bienestar de la población de Nueva York.

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

Gráfico, Gráfico de barras

Descripción generada automáticamente

*Fig. 2.1. Distribución de accidentes de tráfico con respecto a las regiones principales de Nueva York*

Sí, existen patrones geográficos en los accidentes de tráfico que sugieren la necesidad de medidas específicas en ciertas áreas. La imagen que me has enviado muestra que el número de personas heridas en accidentes de tráfico varía significativamente entre los diferentes distritos de la ciudad de Nueva York. Brooklyn, Queens y Manhattan tienen el mayor número de personas heridas, mientras que el Bronx y Staten Island tienen el menor.

Estos patrones sugieren que es posible que haya factores específicos que contribuyan a la alta tasa de accidentes en estos distritos

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

Gráfico, Gráfico de barras

Descripción generada automáticamente

La imagen muestra que los hombres negros son el grupo demográfico más propenso a estar involucrado en accidentes de tráfico y arrestos. Los hombres negros representan el 35% de todos los arrestos en la ciudad de Nueva York.

Hay una serie de factores subyacentes que pueden contribuir a esta tendencia. Estos factores incluyen:

* Algunas personas creen que la policía es más propensa a detener y arrestar a las personas de color que a las personas blancas. Esto puede conducir a una mayor representación de las personas de color en las estadísticas de arrestos.

Por otra parte, con respecto a los accidentes de tráfico,

Gráfico, Histograma

Descripción generada automáticamente

La grafica muestra que el grupo que tiene más número de accidentes son los hombres que están ocupando un vehículo, seguido por las mujeres que utilizan el mismo medio, esto indica que los accidentes de tráfico en Nueva York se dan principalmente entre vehículos, por lo cual los enfoques deben orientarse a tratar de evitar que ocurran accidentes entre carros, camionetas y vehículos de este tipo.

1. ¿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?

Según la gráfica, sí hay una correlación entre la hora del día y la incidencia de accidentes de tráfico y arrestos. La gráfica muestra que el número de personas lesionadas en accidentes de tráfico es mayor durante las horas pico de la mañana y de la tarde, que son de 7 a 9 de la mañana y de 5 a 7 de la tarde, respectivamente. Esto se debe a que hay más tráfico en las carreteras durante estas horas, lo que aumenta el riesgo de accidentes.

Gráfico, Gráfico de líneas

Descripción generada automáticamente

En resumen, la respuesta a la pregunta es que sí hay una correlación entre la hora del día y la incidencia de accidentes de tráfico y arrestos.

Aquí hay una explicación más detallada de la correlación entre la hora del día y la incidencia de accidentes de tráfico y arrestos:

* Horas pico de la mañana: El número de personas lesionadas en accidentes de tráfico es mayor durante las horas pico de la mañana, que son de 7 a 9 de la mañana. Esto se debe a que más personas están viajando al trabajo o a la escuela durante estas horas, lo que aumenta el tráfico en las carreteras.
* Horas pico de la tarde: El número de personas lesionadas en accidentes de tráfico es mayor durante las horas pico de la tarde, que son de 5 a 7 de la tarde. Esto se debe a que más personas están regresando a casa del trabajo o de la escuela durante estas horas, lo que aumenta el tráfico en las carreteras.

1. **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.

Las técnicas de aprendizaje de máquina supervisada que se seleccionaron para alcanzar el objetivo del negocio fueron:

* Técnicas de regresión, como regresión logística, para predecir tasas de accidentes tráfico, el número de víctimas en accidentes o los lugares con mayor número de arrestos. Ayudando así, a identificar factores que contribuyen a la seguridad vial y permitiendo una asignación de recursos más efectiva.

La técnica de aprendizaje de máquina no supervisada que se seleccionaron para alcanzar el objetivo del negocio fue:

* Técnicas de agrupamiento o *clustering,* como *K-means,* para identificar patrones y agrupar áreas o calles con características de seguridad similares, ayudando a identificar zonas de alto riesgo o definir zonas que necesitan un mayor patrullaje.

1. **Preparación de datos para modelado: este apartado se divide en las siguientes tareas.**  
   · Eliminar características 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.  
   · Normalización de variables numéricas: en este apartado se espera que se realice una normalización de datos si es el caso.  
   · 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.

Con el objetivo de preparar nuestros datos para un análisis profundo y su posterior implementación en un modelo de machine learning, se han llevado a cabo transformaciones clave en nuestros DataFrames. Estas transformaciones han permitido convertir las variables categóricas en representaciones numéricas, facilitando así su interpretación y análisis.

*Transformación en Valores Clave-Valor:*

Para facilitar la interpretación y el procesamiento de los datos, hemos asignado valores numéricos a las categorías presentes en las columnas de nuestros DataFrames. Esto se hizo mediante la creación de diccionarios de mapeo que asignan un valor numérico único a cada categoría. Por ejemplo, en la columna 'ARREST\_BORO', las categorías 'Bronx', 'Queens', 'Manhattan', 'Brooklyn' y 'Staten Island' se han asignado valores numéricos del 1 al 5. Este enfoque de clave-valor permite que el modelo de Machine Learning identifique y comprenda estas categorías de manera más efectiva.

*Exploración de Relaciones con la Matriz de Correlación:*

Además de la transformación de categorías en valores numéricos, hemos explorado las relaciones entre las variables en nuestros DataFrames. Una herramienta valiosa para esta exploración es la matriz de correlación. Esta matriz nos permite visualizar las relaciones entre todas las variables del DataFrame y determinar si existen correlaciones significativas. Al calcular la correlación entre pares de variables, podemos identificar patrones y dependencias que pueden ser útiles para el modelado de Machine Learning. Por ejemplo, podríamos descubrir que ciertas variables tienen una fuerte correlación positiva o negativa, lo que sugiere una influencia mutua.

*Finalidad de las Transformaciones:*

La finalidad de estas transformaciones es permitir que nuestro modelo de Machine Learning procese los datos de manera más efectiva y encuentre relaciones significativas entre las variables. Al convertir categorías en valores numéricos y visualizar las correlaciones, estamos proporcionando al modelo información estructurada y significativa que puede utilizar para realizar predicciones precisas. Estas transformaciones no solo mejoran la capacidad de los algoritmos de ML para comprender los datos, sino que también pueden llevar a una mayor precisión y rendimiento en la predicción de resultados.

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

Este punto se encontrará en los *Notebooks.*

1. **Evaluación:**

Los modelos de aprendizaje supervisado tuvieron métricas de rendimiento mucho más altas y precisas prediciendo las zonas con mayor número de arrestos (Bronx, Queens, Manhattan, Brooklyn, Staten Island) y cuáles son los tipos de heridas más frecuentes (no especificadas, heridas o muerte). En cambio, los aprendizajes no supervisados tendieron a desviarse y a arrojar resultados con métricas bajas.

En base a las observaciones proporcionadas para las métricas de precisión, recall, F1-score y accuracy para cada uno de los tres conjuntos de datos, podemos realizar evaluaciones generales para cada técnica de clasificación. A continuación, proporciono algunas evaluaciones y consideraciones:

1. Técnica de Clasificación 1 (Clases: 1, 2, 3):

- Precisión:

- Ventaja: Buena precisión para las clases '1' y '2'.

- Desafío: Dificultades en la precisión para la clase '3'.

- Recall:

- Ventaja: Recall alto para la clase '1'.

- Desafío: Recall bajo para la clase '3'.

- F1-Score:

- Ventaja: Buen equilibrio para las clases '1' y '2'.

- Desafío: F1-score bajo para la clase '3'.

- Accuracy:

- Ventaja: Alta accuracy global.

- Desafío: Posible impacto del desbalance de clases.

2. Técnica de Clasificación 2 (Clases: Bronx, Queens, Manhattan, Brooklyn, Staten Island):

- Precisión:

- Ventaja: Buena precisión para la clase '3' (Manhattan).

- Desafío: Dificultades en la precisión para la clase '5' (Staten Island).

- Recall:

- Ventaja: Recall alto para la clase '3' (Manhattan).

- Desafío: Recall bajo para la clase '5' (Staten Island).

- F1-Score:

- Ventaja: Buen equilibrio para la clase '3' (Manhattan).

- Desafío: F1-score bajo para la clase '5' (Staten Island).

- Accuracy:

- Ventaja: Moderadamente alta accuracy global.

- Desafío: Posible impacto del desbalance de clases.

3. Técnica de Clasificación 3 (Clases: Unspecified, Injured, Killed):

- Precisión:

- Ventaja: Buena precisión para las clases '1' (Unspecified) y '2' (Injured).

- Desafío: Precisión baja para la clase '3' (Killed).

- Recall:

- Ventaja: Recall alto para la clase '1' (Unspecified) y '2' (Injured).

- Desafío: Recall bajo para la clase '3' (Killed).

- F1-Score:

-Ventaja: Buen equilibrio para las clases '1' (Unspecified) y '2' (Injured).

- Desafío: F1-score bajo para la clase '3' (Killed).

- Accuracy:

- Ventaja: Alta accuracy global.

- Desafío: Posible impacto del desbalance de clases.

- En todos los conjuntos de datos, hay desbalance de clases, lo que puede afectar las métricas de evaluación y la interpretación de los resultados.