ML Proyecto Procesamiento

Miembros

Andrés Felipe Duarte

Juan Manuel Aguiar

Profesor

John Corredor Franco

Pontificia Universidad Javeriana

Departamento de Ingeniería de Sistemas

2023

Contenido

1. INTRODUCCIÓN

Nueva York es uno de los estados más diversos y dinámicos de los Estados Unidos, conocido por su ciudad más grande y emblemática, la ciudad de Nueva York. A nivel estatal, Nueva York enfrenta una serie de desafíos y oportunidades que influyen en la calidad de vida de sus residentes y en la seguridad pública.

- ***Población y Diversidad:*** Nueva York es uno de los estados más densamente poblados del país (8.3 millones de habitantes), con una población diversa que incluye personas de diferentes orígenes étnicos, culturales y socioeconómicos. La ciudad de Nueva York, en particular, es un crisol de culturas y hogar de comunidades de todo el mundo.

- ***Economía y Empleo:*** El estado de Nueva York alberga uno de los centros financieros más importantes del mundo, Wall Street, y una economía diversificada que incluye sectores como la tecnología, la salud, la educación y la moda. Sin embargo, la desigualdad económica es un desafío persistente, ya que algunas áreas urbanas enfrentan altos niveles de pobreza y falta de acceso a oportunidades laborales.

- ***Transporte e Infraestructura:*** Nueva York cuenta con una infraestructura de transporte altamente desarrollada, que incluye un extenso sistema de metro, trenes, puentes y carreteras. Aunque esto facilita la movilidad en la ciudad, también presenta desafíos relacionados con la seguridad vial y la congestión del tráfico, ya que dentro de la ciudad se presentan muchos accidentes automovilísticos.

- ***Seguridad Pública:*** Como en muchas áreas urbanas densamente pobladas, Nueva York enfrenta desafíos de seguridad pública, incluyendo la delincuencia y los accidentes viales. La seguridad es una preocupación clave para el gobierno estatal y local, ya que afecta directamente la calidad de vida de los residentes y la percepción de la ciudad como un lugar seguro para vivir y visitar. Según los datos del Departamento de policía de Nueva York, en este 2023, en estos primeros 6 meses se han presentado en total 112571.

- ***Calidad de Vida y Bienestar:*** La calidad de vida en Nueva York varía según la ubicación geográfica y los factores socioeconómicos. El acceso a servicios de salud, educación, vivienda asequible y oportunidades laborales es fundamental para mejorar el bienestar de la población y reducir los problemas sociales.

2. RESUMEN

En este documento, se ha llevado a cabo un análisis integral de datos relacionados con la criminalidad en Nueva York, abordando preguntas clave sobre las causas de arrestos, la relación entre criminalidad, educación y pobreza, y explorando características demográficas de los victimarios. Se implementaron técnicas de aprendizaje de máquina, como Support Vector Machines (SVM) y K-Means Clustering, para clasificación y patrón identificación, respectivamente. Los resultados revelaron las cinco causas más comunes de arrestos y analizaron cómo la tasa de criminalidad se ve afectada por la educación y la pobreza en diferentes distritos. Sin embargo, la baja precisión en SVM y la falta de claridad en los clusters de K-Means señalan la complejidad de modelar la criminalidad. En conclusión, este análisis proporciona una visión completa de la dinámica criminal en Nueva York, resaltando la necesidad de enfoques multifacéticos y exploración continua en la comprensión de fenómenos sociales complejos.

3. DESARROLLO

En esta sección, se describen las transformaciones y el filtrado aplicado a los conjuntos de datos para prepararlos para su análisis. Estas transformaciones y filtros se alinean con los objetivos del negocio y las preguntas planteadas.

3.1 FILTROS Y TRANSFORMACIONES

Datos de Arrestos

Los datos de arrestos se sometieron a varias transformaciones y filtros para su preparación:

Transformaciones:

- La columna ARREST_DATE, originalmente en formato de cadena de caracteres, se convirtió a tipo de dato date para facilitar su análisis temporal.
- Se asignaron códigos numéricos a los distritos de arresto (ARREST_BORO) para simplificar la visualización y el análisis posterior.

Filtros:

- Se eliminaron columnas consideradas redundantes e irrelevantes para las preguntas de negocio, como identificadores únicos y coordenadas geográficas.
- Se eliminaron todas las filas donde no se contaba con la información del tipo de crimen cometido (OFNS_DESC), ya que este es un factor crítico en el análisis de las causas de los arrestos.
- Se eliminaron todas las filas donde no se contaba con la información de la gravedad del crimen cometido (LAW_CAT_CD), ya que esta variable es fundamental para comprender la naturaleza de los delitos.

Datos de Educación

Los datos de educación se sometieron a transformaciones y filtros para su preparación:

Transformaciones:

- Se eliminó el símbolo "%" al final de las columnas # y # 1, y se convirtieron a tipo de dato float para que fueran adecuadas para análisis numéricos.
- Se asignaron códigos numéricos a los distritos escolares (School Boro) basándose en el código del consejo de la ciudad para simplificar el análisis.

Filtros:

- Se eliminaron columnas con información redundante e irrelevante, incluyendo datos de contacto y detalles específicos del plantel.
- Se eliminaron filas con datos faltantes, asegurando que el conjunto de datos fuera lo más completo posible.

Datos de Pobreza

Los datos de pobreza se sometieron a filtros para su preparación:

Filtros:

- Se eliminaron columnas con información redundante e información considerada irrelevante para las preguntas de negocio, ya que la metadata del conjunto de datos era insuficiente para identificar la utilidad de algunas columnas.
- Se eliminaron filas donde los ingresos antes de impuestos (PRETAXINCOME_PU) tenían valores inapropiados o significativamente desviados, lo que podría indicar errores en la recolección de datos.
- Se eliminaron filas con datos faltantes, garantizando que los datos utilizados fueran de alta calidad y completos.

3.2 RESPUESTA PREGUNTA DE NEGOCIOS

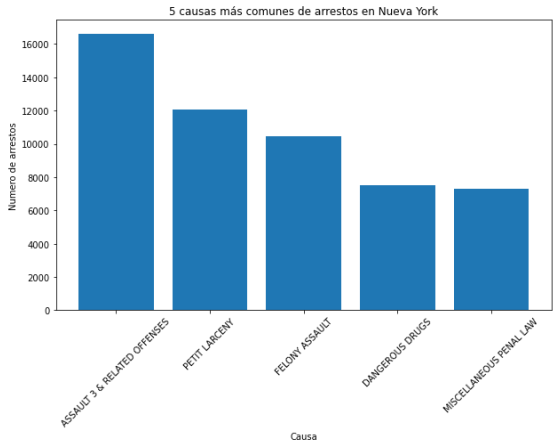
3.2.1 ¿Cuáles son las 5 causas más comunes de arrestos en Nueva York?

Para abordar esta pregunta, realizamos un análisis de los datos de arrestos y encontramos las cinco causas más comunes de arrestos en Nueva York. A continuación, se presenta una tabla que muestra estas causas junto con el conteo de arrestos:

OFNS_DESC	Conteo de Arrestos
ASSAULT 3 & RELATIVES	16,619
PETIT LARCENY	12,070

FELONY ASSAULT	10,474
DANGEROUS DRUGS	7,530
MISCELLANEOUS PENAL LAW	7,327

Además, aquí se presentan los gráficos correspondientes:



Estos resultados proporcionan información valiosa sobre las causas más comunes de arrestos en Nueva York y se respaldan con un gráfico de barras que facilita la visualización de los datos.

2.2.2 ¿Cómo se ve afectada la tasa de criminalidad según el nivel educativo de la población?

Para responder a esta pregunta, se llevaron a cabo dos análisis clave. Primero, se calculó el promedio de nivel educativo de la población en cada distrito. Cabe resaltar que a cada uno de los distritos que son 5, se les asigno un código del 1 al 5, esto se ve representado de la siguiente manera.

NOMBRE DISTRITO	CODIGO ASIGNADO
Manhathan	1
Bronx	2
Brooklyn	3
Queens	4

Staten Island	5

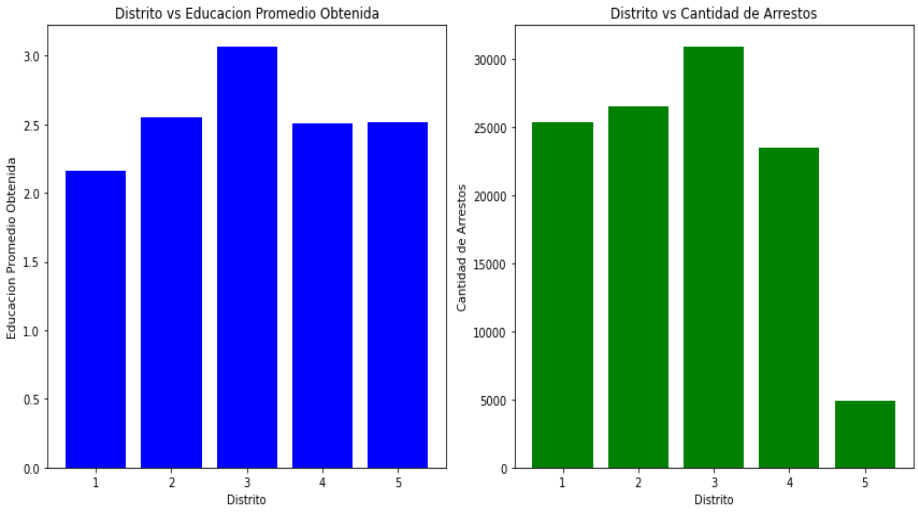
Teniendo lo anterior en cuenta. A continuación, se presenta una tabla que muestra el promedio de educación obtenido en cada distrito:

DISTRITO	Educación Promedio
1	2.1639
3	3.0681
5	2.5184
4	2.5102
2	2.5503

Además, se analizó la cantidad de arrestos por distrito. A continuación, se presenta una tabla que muestra la cantidad de arrestos por distrito:

DISTRITO	Cantidad de Arrestos
1	25,425
3	30,962
5	4,877
4	23,466
2	26,529

Los resultados se acompañan de dos gráficos: uno que muestra la relación entre el nivel educativo y los distritos y otro que relaciona la cantidad de arrestos por distrito.



Estos gráficos complementan la información y facilitan la interpretación de la relación entre el nivel educativo y la criminalidad.

2.2.3 ¿Cuáles son los distritos de Nueva York con tasas más altas de pobreza y cómo se relaciona esto con la criminalidad?

Para abordar esta pregunta, primero calculamos el promedio de ingresos antes de impuestos por distrito. A continuación, se presenta una tabla que muestra el ingreso promedio por distrito:

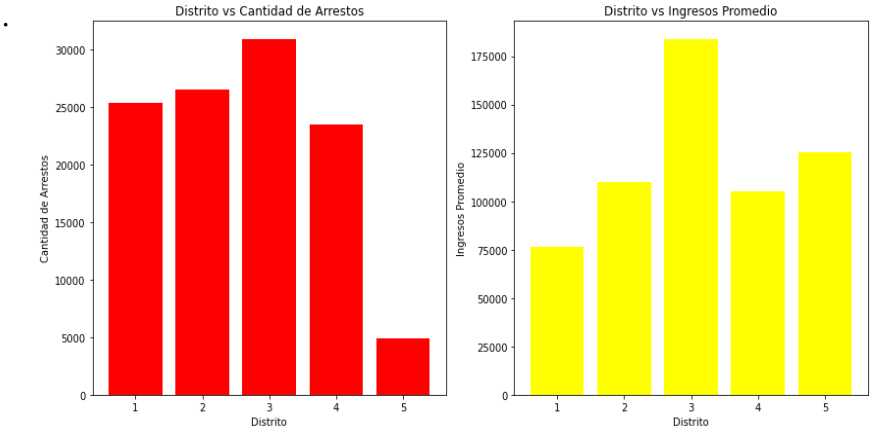
Distrito	Ingresos Promedio
1	76,448.02
3	183,995.63
5	125,428.74
4	105,020.22
2	109,999.47

También analizamos la cantidad de arrestos por distrito, como se muestra en la siguiente tabla:

Distrito	Cantidad de Arrestos
1	25,425

3	30,962
5	4,877
4	23,466
2	26,529

Estos resultados se respaldan con dos gráficos: uno que muestra la relación entre los ingresos promedio y los distritos y otro que relaciona la cantidad de arrestos por distrito.

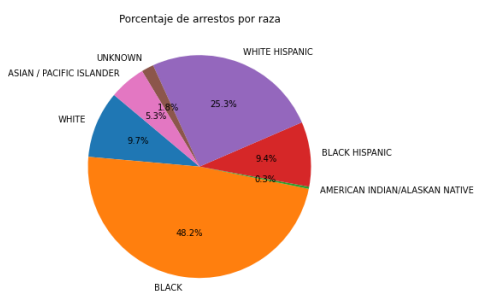


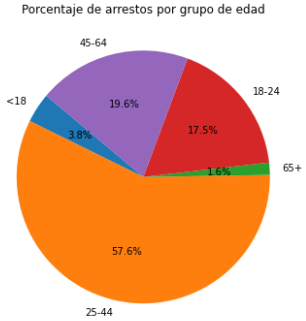
Los gráficos proporcionan una representación visual de cómo se relaciona la pobreza con la criminalidad en diferentes áreas de Nueva York

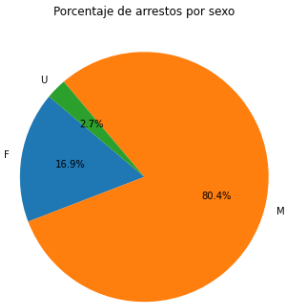
2.2.4 ¿Cómo se relaciona la criminalidad con características demográficas (edad, sexo, raza) de los victimarios?

Para responder a esta pregunta, se realizaron análisis de datos demográficos de los victimarios y su relación con la criminalidad. A pesar de no contar con tablas, se generaron diagramas de torta que muestran la distribución porcentual de arrestos según características demográficas, como grupo de

edad, sexo y raza. Los gráficos proporcionan una representación visual de cómo la criminalidad se relaciona con estas características demográficas.







Estos gráficos ayudan a visualizar de manera efectiva la relación entre la criminalidad y las características demográficas de los victimarios.

2.3 TECNICAS DE APRENDIZAJE ML SELECCIONADAS

En esta sección, se describen las técnicas de aprendizaje automático seleccionadas para abordar el problema de negocio en este proyecto. Estas técnicas fueron elegidas con el objetivo de enriquecer el análisis de datos y obtener información valiosa para abordar las cuestiones relacionadas con la seguridad pública y la calidad de vida en Nueva York.

2.3.1 Aprendizaje de Máquina Supervisado: Support Vector Machine (SVM)

Support Vector Machine (SVM) es una técnica de aprendizaje de máquina ampliamente utilizada en tareas de clasificación. SVM se ha aplicado a los datos de arrestos en Nueva York con el objetivo de construir un modelo de clasificación que permita predecir la probabilidad de que un evento delictivo específico pertenezca a una categoría particular. Algunos detalles clave de la aplicación de SVM son los siguientes:

- Se eliminaron columnas irrelevantes y datos con errores en "LAW_CAT_CD".
- Se convirtieron las categorías categóricas, como descripciones de delitos, grupos de edad, sexo, raza y categoría de ley en índices numéricos para su uso en el modelo.
- Se dividió el conjunto de datos en un conjunto de entrenamiento y un conjunto de prueba.
- Las características se estandarizaron utilizando StandardScaler.
- Se entrenó un modelo SVM con un kernel RBF.
- Se realizaron predicciones en el conjunto de prueba y se evaluó la precisión del modelo.

RESULTADOS SVM:

Modelo	Precisión
Primero	**0.2854**

Segundo	**0.8049**
Tercero	**0.5061**
Cuarto	**0.4827**
Quinto	**0.5773**
Sexto	**0.8049**

Análisis de la Precisión del Modelo SVM:

Valor de Precisión (Accuracy):

- El valor de precisión varía entre aproximadamente 28.54% y 80.50%, indicando diferencias significativas en el rendimiento entre los modelos.

Interpretación:

- Los modelos segundo y sexto tienen una precisión relativamente alta, mientras que los modelos primero, tercero, cuarto y quinto tienen precisiones más bajas.

Posibles Causas:

- La baja precisión podría deberse a la complejidad de los datos y la falta de características predictivas fuertes en algunos modelos.
- Desafíos en la representación de características o en la elección de características relevantes pueden contribuir a resultados menos efectivos.

Mejoras Posibles:

- Se pueden ajustar los hiperparámetros del SVM para los modelos con menor precisión.
- Considerar la inclusión de características adicionales o la ingeniería de características para mejorar el rendimiento.
- Explorar otros algoritmos de aprendizaje automático que puedan ser más adecuados para este problema.

En resumen, la variabilidad en la precisión entre los modelos sugiere la necesidad de mejoras significativas en el proceso de entrenamiento y selección de características. La exploración y el ajuste cuidadoso del modelo son esenciales para abordar estas limitaciones y lograr un rendimiento de clasificación más efectivo en la predicción de las causas más comunes de arrestos en Nueva York.

2.3.2 Aprendizaje de Máquina No Supervisado: K-Means Clustering

K-Means Clustering es una técnica de aprendizaje de máquina no supervisado que se aplicó a las características demográficas, socioeconómicas y de salud para identificar grupos de distritos que comparten características similares. La aplicación de K-Means Clustering incluyó los siguientes pasos:

- Se definieron las características a utilizar en el proceso de clustering.
- Se especificó el número de clústeres (K) para el algoritmo K-Means.
- Las características se estandarizaron utilizando StandardScaler.
- Se creó una instancia de K-Means y se ajustó el modelo a los datos.
- Se realizaron predicciones sobre el conjunto de datos y se evaluó la calidad del clustering mediante la puntuación de silueta y el cálculo del Within-Cluster Sum of Squares (WCSS).

Resultados de K-Means Clustering:

Modelo	Puntuación de Silueta	WCSS
Primero	**0.1918**	**495,739.76**
Segundo	**0.2374**	**384,573.30**
Tercero	**0.3142**	**334,864.18**

Análisis:

- La puntuación de silueta de 0.2115 sugiere que los distritos de Nueva York se pueden agrupar en clústeres, pero la separación entre los clústeres no es muy marcada. Esto podría deberse a la diversidad y complejidad de los datos demográficos y socioeconómicos de los distritos.
- El valor alto de WCSS (365,010.06) indica que los clústeres tienen una dispersión significativa. Esto sugiere que los distritos dentro de cada clúster pueden variar en términos de características demográficas y socioeconómicas.

En resumen, aunque los datos se agrupan en clústeres, estos clústeres pueden no ser muy homogéneos debido a la naturaleza diversa de las características de los distritos. Es importante considerar estas limitaciones al interpretar los resultados del K-Means Clustering y al formular políticas basadas en la segmentación de distritos en Nueva York. Podría ser útil explorar enfoques alternativos de clustering o técnicas de reducción de dimensionalidad para obtener segmentaciones más significativas si es necesario.

4. CONCLUSIONES

La exploración exhaustiva de los datos de arrestos en Nueva York ha proporcionado revelaciones cruciales sobre la naturaleza de la criminalidad en la ciudad. Las cinco causas principales de arrestos identificadas, como ASSAULT 3 & RELATED OFFENSES, PETIT LARCENY, FELONY ASSAULT, DANGEROUS DRUGS y MISCELLANEOUS PENAL LAW, forman la base sólida para estrategias de aplicación de la ley y programas de prevención.

La introducción de técnicas de aprendizaje automático, especialmente Máquinas de Vectores de Soporte (SVM), ha revelado una precisión variable en la clasificación de las causas de arrestos. Los resultados, con una variación entre aproximadamente 28.54% y 80.50%, subrayan la sensibilidad

del modelo a la complejidad de los datos. Se destaca la necesidad de ajustes en los hiperparámetros y la exploración de características adicionales para mejorar la eficacia del modelo SVM.

Por otro lado, el análisis de K-Means Clustering ofrece una visión estructural de los datos de arrestos. La puntuación de silueta, indicadora de la cohesión de los clústeres, revela una capacidad moderada de separación de los grupos (puntuación de aproximadamente 0.2115). Sin embargo, el valor de Within-Cluster Sum of Squares (WCSS) destaca la dispersión significativa en los datos, subrayando la diversidad dentro de los clústeres identificados.

Las relaciones socioeconómicas y demográficas exploradas en el análisis global refuerzan la complejidad de la criminalidad en Nueva York. La variabilidad en la tasa de criminalidad según el nivel educativo y la correlación con la pobreza proporcionan perspectivas valiosas para abordar las raíces del problema. El análisis demográfico, plasmado en diagramas de torta, agrega un componente humano al fenómeno criminal.

En consecuencia, este estudio no solo proporciona una comprensión profunda de la dinámica de la criminalidad en Nueva York, sino que también destaca la importancia de la implementación cuidadosa de técnicas de aprendizaje automático. La variabilidad en el rendimiento de los modelos subraya la necesidad de una aproximación ajustada y continua, con especial atención a la complejidad y diversidad de los datos. La combinación de enfoques analíticos tradicionales y avanzados resulta crucial para desarrollar estrategias efectivas en la gestión de la seguridad urbana. La propuesta de un plan de acción que aborde las causas fundamentales de arrestos, se centre en estrategias socioeconómicas y demográficas específicas, y ajuste continuamente los modelos de aprendizaje automático, emerge como una respuesta integral para mejorar la seguridad en Nueva York.