Pontificia Universidad Javeriana

Esteban Pedraza
Valentina Bustos
Mauricio Rodríguez
Sofía Torres

Entrega 2: Proyecto

Procesamiento de Datos a Gran Escala
2024-10

Índice

1. Introducción

La ciudad de Nueva York, como epicentro de innumerables oportunidades y desafíos, enfrenta una coyuntura decisiva en su desarrollo socioeconómico post-pandemia. Este escenario, marcado por una recuperación económica, plantea interrogantes críticos sobre la dirección futura de sus políticas públicas y estrategias de desarrollo.

Ante los desafíos que enfrenta la ciudad de Nueva York, particularmente en lo que respecta a la seguridad y el bienestar de sus habitantes, este proyecto se propone abordar de manera estratégica problemas críticos como son: la elevada cantidad de arrestos, los altos índices de criminalidad y los niveles de pobreza que se presentan en sus comunidades . Utilizando datos comprehensivos que abarcan desde 2018 hasta 2024, este análisis buscará entender las dinámicas detrás de estos fenómenos con el fin de proponer soluciones efectivas.

Objetivo General del Proyecto:

Desarrollar tres estrategias específicas para la ciudad de Nueva York dos destinadas a reducir la cantidad de arrestos junto con una mejora del acompañamiento policial y una estrategia orientada a revelar los aspectos que más afectan a las comunidades que se encuentran en situación de pobreza, basándose en análisis detallados de los datos correspondientes a arrestos y pobreza 2016 y 2024.

Objetivos Específicos del Proyecto:

- Utilizar el K-Means Clustering para identificar patrones latentes que sean difíciles de observar y que tengan una alta influencia en la cantidad de arrestos en la ciudad de Nueva York.
- Analizar la relación entre niveles de educación, pobreza, etnia, género y tasas de arresto en distritos de Nueva York para identificar patrones y desarrollar estrategias de intervención específicas.
- Desarrollar un modelo de clasificación utilizando un árbol de decisión que determine si una persona tiene empleo o no. Este modelo se basará en datos categóricos mayormente relacionados con la pobreza, los cuales son de fácil recolección para la mayoría de los gobiernos.

2. Filtros y transformaciones

En el análisis de datos de arrestos, es fundamental realizar transformaciones y filtros para preparar los datos adecuadamente para el análisis y la modelización. Estas transformaciones permiten normalizar los datos, agrupar categorías, y manejar valores faltantes o inconsistencias. A continuación, se presentan las transformaciones más relevantes aplicadas a los datasets de arrestos:

```python
# Diccionario de mapeo
borough_map = {
    1: 'Bronx',
    2: 'Brooklyn',
    3: 'Manhattan',
    4: 'Queens',
    5: 'Staten Island'
}

# Aplicar el mapeo a la columna 'Borough'
pov['Borough'] = pov['Borough'].map(borough_map)

# Mostrar el DataFrame resultante
print(pov[['Borough']])
```

```python
age_map = {
    '<18': '< 18',
    '25-44': '18 - 64',
    '65+': '65+',
    '45-64': '18 - 64',
    '18-24': '18 - 64'
}

# Aplicar el mapeo a la columna 'Age'
arrests['AGE_GROUP'] = arrests['AGE_GROUP'].map(age_map)
arrests['AGE_GROUP'].unique()
```

Estas transformaciones son esenciales para limpiar y preparar los datos para el análisis. Al normalizar las categorías de Borough, agrupar los grupos de edad y categorizar las descripciones de delitos, los datos se vuelven más manejables y adecuados para el análisis estadístico y la modelización. Estas transformaciones permiten identificar patrones más fácilmente y aseguran que los análisis posteriores sean precisos y

significativos a la hora de las visualizaciones. Esto se hizo con otras columnas como lo fueron el genero, estado civil, mecanismos de arrestos y estados laborales.

Para la mayoría de los filtros se realizaron durante la implementación de las visualizaciones. Sin embargo, uno de los más relevantes durante la limpieza de datos fue la identificación de columnas relevantes para realizar nuestro análisis, que también se presentaron en la entrega anterior y están detalladas en los códigos. Además, se identificaron outliers en lo que respecta a los ingresos.

- **Identificación y Remoción de Outliers en Ingresos:** Se filtraron los valores de ingresos ajustados para identificar y remover outliers.
- **Filtrado de Columnas Relevantes para el Análisis:** Se seleccionaron columnas específicas para el análisis detallado, enfocándose en variables importantes como etnicidad, edad, género, y otros factores socioeconómicos.
- **Creación de una Columna Indicadora de Hijos:** Se creó una columna que indica si el individuo tiene hijos basándose en los gastos en cuidado de niños.
- **Filtrado por Presencia de Hijos**:Se filtraron los datos para analizar la relación entre tener hijos y los ingresos laborales ajustados.

```
# mostrar outliers
pov[pov['Ingresos_Laborales_Ajustados'] < 20000]
```

	Borough	nivel_educativo	Etnicidad	Tiempo_Completo_Parcial	puesto_familiar	Horas_
2	Manhattan	3.0	White	3	1	
5	Queens	4.0	Hispanic	1	0	
6	Queens	4.0	Hispanic	1	1	
8	Brooklyn	4.0	Black	3	0	
10	Brooklyn	3.0	Asian	2	1	
...	...	...	...	...	...	
68263	Brooklyn	3.0	Hispanic	3	6	
68265	Bronx	1.0	Hispanic	3	4	
68266	Bronx	1.0	Hispanic	2	4	
68269	Brooklyn	NaN	Other	3	2	
68272	Staten Island	4.0	White	3	4	

```
# Calcula el promedio de ingresos y gastos
ingresos_promedio = pov['Ingresos_Laborales_Ajustados'].mean()
gastos_promedio = pov['Gastos_Cuidado_Ninos'].mean()

# Definir las categorías y los valores
categorias = ['Ingresos', 'Gastos']
valores = [ingresos_promedio, gastos_promedio]

# Graficar las barras
plt.bar(categorias, valores, color=['green', 'red'])

# Agregar etiquetas, título y leyenda
plt.xlabel('Categoría')
plt.ylabel('Promedio')
plt.title('Promedio de ingresos y gastos')
plt.show()
```

```
# box plot de Ingresos_Laborales_Ajustados
pov.boxplot(column='Ingresos_Laborales_Ajustados',
            by='Borough',
            grid=False,
            rot=45, fontsize=15)
```

```
# crear columna que si tiene hijos. Si el gasto en cuidado de niños es mayor a 0, entonces tiene hijos. 1 si tiene hijos, 0 si no tiene hijos
pov['Tiene_Hijos'] = pov['Gastos_Cuidado_Ninos'].apply(lambda x: 1 if x > 0 else 0)
pov['Tiene_Hijos'].sum()
```

```
6930
```

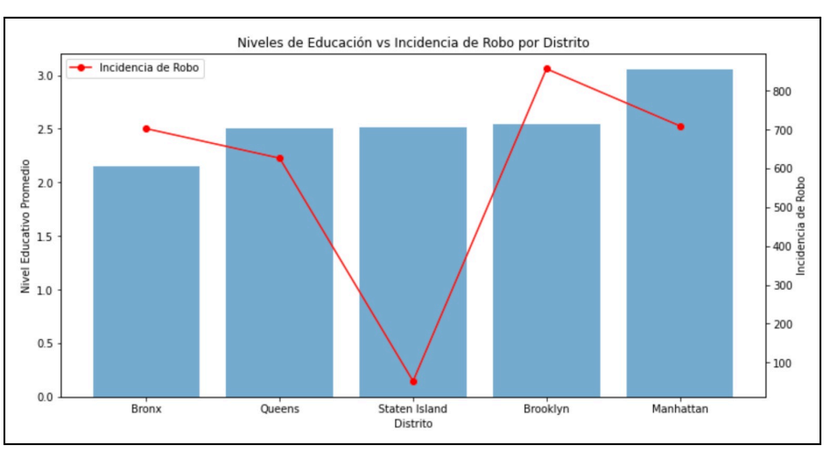
```
pov['Tiene_trabajo'] = pov['Ingresos_Laborales_Ajustados'].apply(lambda x: 1 if x > 0 else 0)
pov['Tiene_trabajo'].sum()
```

```
34731
```

```
#box plot de Ingresos_Laborales_Ajustados segun si tiene hijos o no
import seaborn as sns
import matplotlib.pyplot as plt
sns.boxplot(x='Tiene_Hijos', y='Ingresos_Laborales_Ajustados', data=pov)
plt.show()
```

3. Solución de la de las preguntas de negocio planteadas

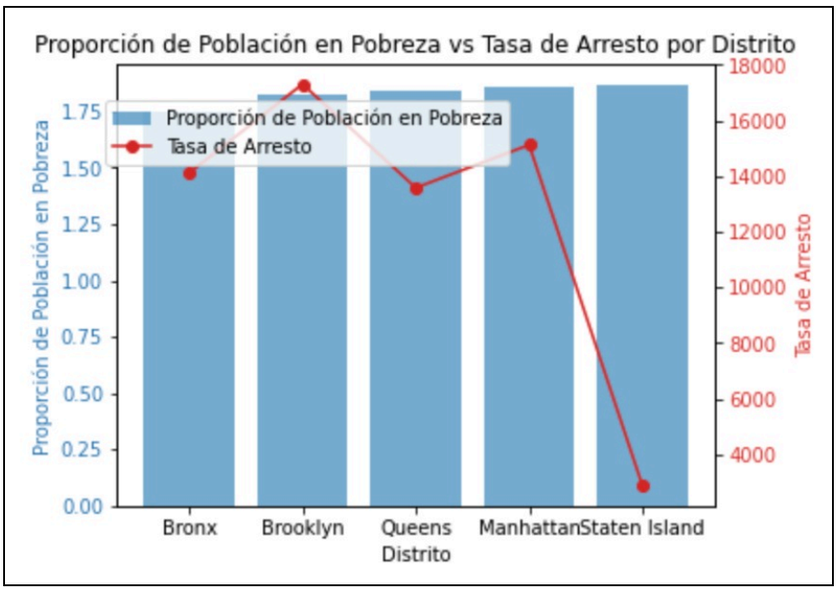
- *¿Cuáles son los distritos que muestran niveles más bajos de educación y coinciden estos con áreas con mayor incidencia de robo?*



El gráfico muestra la relación entre los niveles educativos promedio y la incidencia de robos por distrito. Los distritos con niveles educativos más bajos son el Bronx, Queens y Staten Island, mientras que Brooklyn y Manhattan tienen

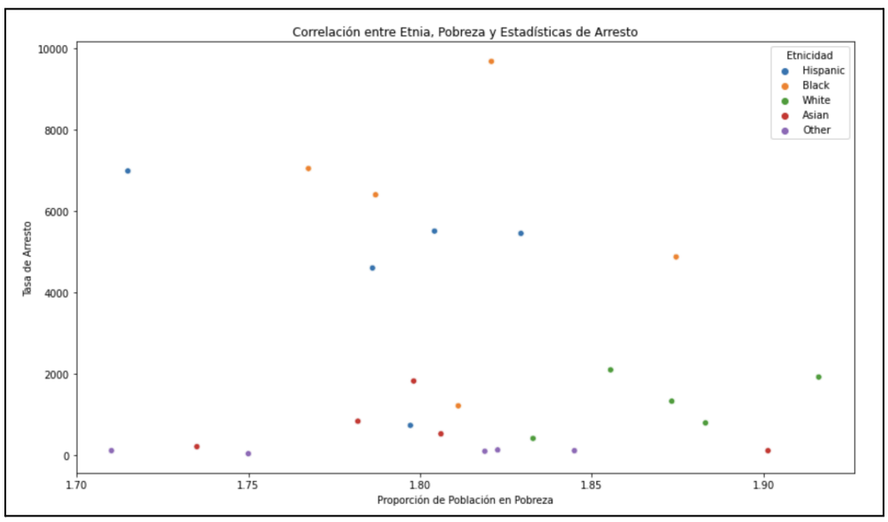
niveles educativos más altos en comparación. **Podemos observar que el Bronx, con uno de los niveles educativos más bajos, tiene una alta tasa de robos. Brooklyn y Manhattan, aunque tienen niveles educativos altos también muestran altas tasas de robos.**

- *¿Los distritos con una mayor proporción de población en situación de pobreza presentan tasas de arresto más elevadas?*



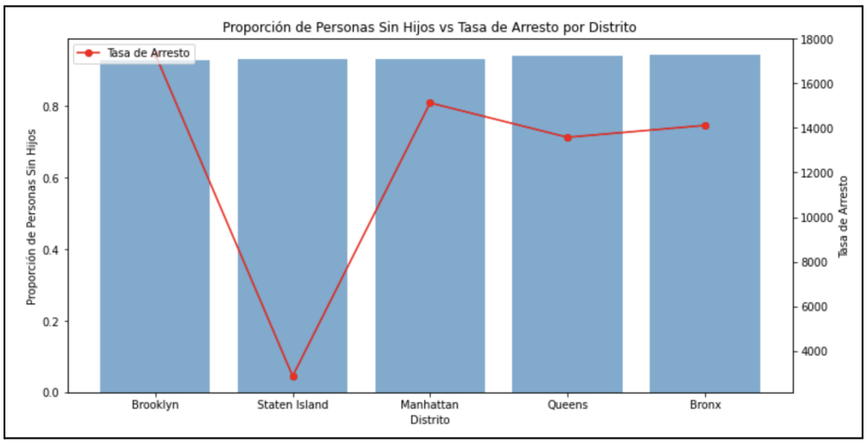
El gráfico muestra la relación entre la proporción de población en situación de pobreza y la tasa de arresto por distrito. Los datos indican que el Bronx y Brooklyn, con las proporciones de pobreza más bajas (1.74 y 1.82 respectivamente), presentan tasas de arresto elevadas. Queens y Manhattan, aunque tienen proporciones de pobreza superiores, también muestran altas tasas de arresto. Staten Island, a pesar de tener una proporción de pobreza parecida a la de Manhattan, muestra una tasa de arresto significativamente más baja. **Esto muestra que la proporción de población en situación de pobreza no es el único factor que influye en las tasas de arresto y que pueden existir otros factores que también afectan las estadísticas de arresto.**

- *¿Existe alguna correlación entre la etnia y la pobreza, y se refleja esta relación en las estadísticas de arresto?*



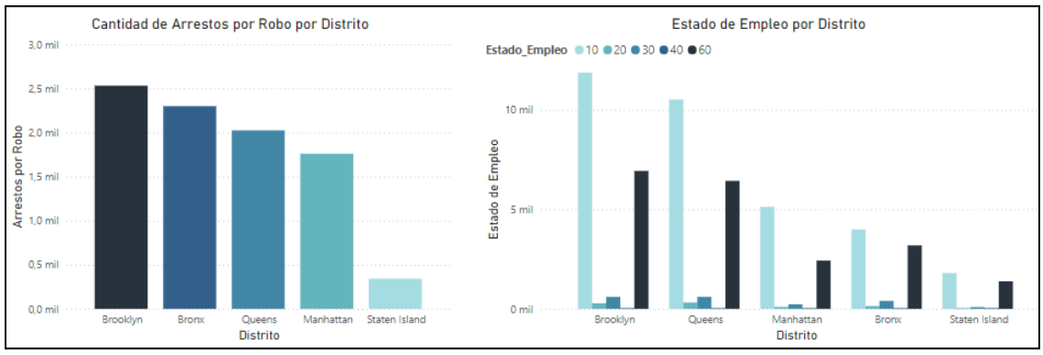
El gráfico muestra la relación entre la proporción de población en situación de pobreza y la tasa de arresto, diferenciada por etnia. **Se puede observar que las tasas de arresto tienden a ser más altas para ciertos grupos étnicos (como los afroamericanos y los hispanos) en los distritos con mayor proporción de población en situación de pobreza.**

- *¿Las personas que no tienen hijos tienden a residir en áreas con índices de arresto más altos o más bajos?*



El gráfico muestra la relación entre la proporción de personas sin hijos y la tasa de arresto por distrito. Vemos que los distritos con mayor proporción de personas sin hijos, como el Bronx y Queens, también tienen tasas de arresto elevadas. Staten Island, que tiene una menor proporción de personas sin hijos, muestra una tasa de arresto significativamente más baja en comparación con otros distritos. **Parece haber una ligera tendencia a que las personas sin hijos vivan en áreas con tasas de arresto más altas, aunque la relación no es lo suficientemente marcada y varía entre los distritos.**

- *¿Los lugares con una menor proporción de trabajadores formales u operativos muestran índices de robo más altos o más bajos?*

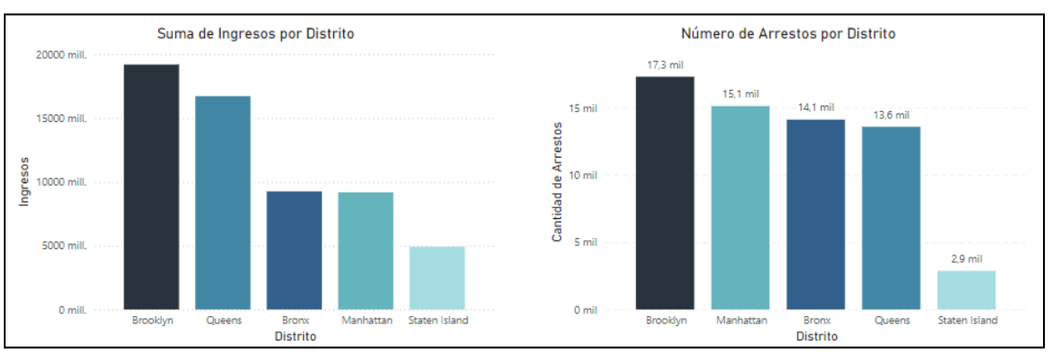


En la primera gráfica se muestra la cantidad de trabajadores de cada tipo (10-60) en cada condado, siendo de interés el 10, civilian with employment at work, y 20, civilian employed with a job but not at work. En la segunda gráfica se muestra la cantidad de arrestos tipificados como robo (‘assault 3 & related offenses’ y ‘robbery’) por distrito.

Los condados de interés son Staten Island y Bronx que son las cuentan con un menor número de trabajadores formales; en Staten Island el índice de robo es más bajo que en el resto de condados, por lo que no hay una fuerte relación entre la proporción de trabajadores formales y los arrestos por robo. Sin embargo, en Bronx siendo el segundo condado con menos trabajadores formales, es el segundo condado con mayor índice de robo, por lo que **se puede deducir que hay una fuerte relación entre el gran número de arrestos por robo y la poca cantidad de trabajadores formales.**

El resultado inesperado de Staten Island puede ser porque del total de registros del dataset de arrestos, solo el 4,56% se encuentra en este condado.

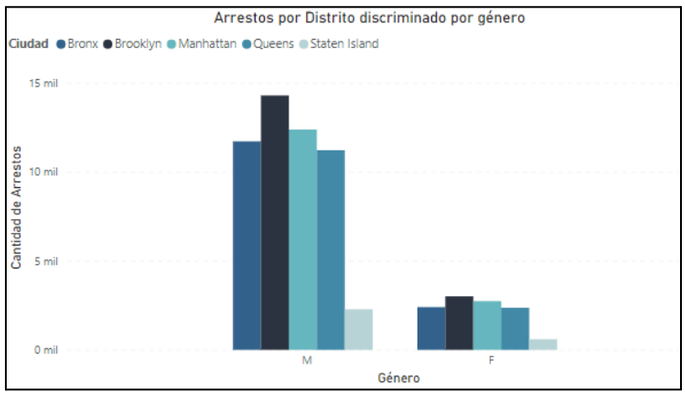
- *¿Los distritos con ingresos más altos presentan tasas de criminalidad mayores o menores?*



En la primera gráfica se encuentra la suma de ingresos por condado y en la segunda gráfica la cantidad de arrestos por ciudad.

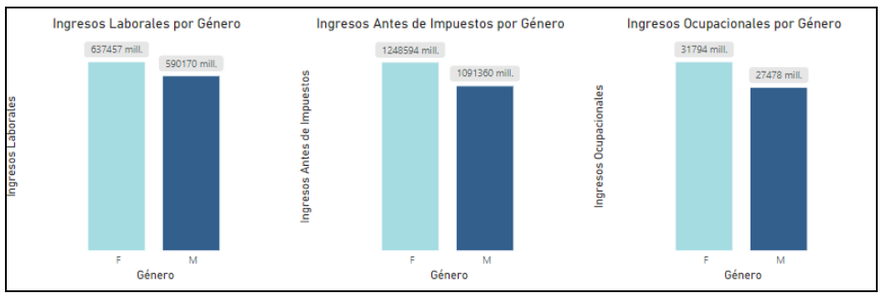
Se observa que el distrito de Brooklyn es el condado con mayores ingresos pero a su vez es el condado con mayor tasa de criminalidad, no obstante, Queens, el segundo distrito con mayores ingresos no tiene una tasa de criminalidad tan alta, de hecho es el cuarto distrito con mayor cantidad de arrestos, por lo que **no hay una tendencia clara que se pueda identificar.**

- *¿Se observa una tendencia en ciertas áreas hacia un mayor número de arrestos de mujeres en comparación con hombres?*



No se observa en ningún área metropolitana una tendencia hacia un mayor número de arrestos de mujeres en comparación con hombres, por el contrario, **es indudable que en los 5 condados de los que se tiene registro hay una cifra muy superior de arrestos de hombres en comparación con mujeres.**

- *¿Existe una disparidad de género en los ingresos reportados en el conjunto de datos sobre pobreza?*



En toda la declaración sobre los ingresos que se encuentran en el dataset de pobreza, **es evidente que existe hay una diferencia entre los ingresos de hombres y mujeres**, que para los casos de ingreso laboral, ingresos antes de impuestos e ingresos ocupacionales es mayor para las mujeres, **sin embargo, esta diferencia es relativamente poca.**

Es probable que este resultado se deba a que del total de registros hay un 52,92% de mujeres y un 47,08% de hombres lo que influye en los resultados que muestran las gráficas.

4. Técnicas de Aprendizaje de Máquina

Supervisada

Este modelo se centra en la aplicación de modelos supervisados para predecir si una persona tiene empleo, utilizando un dataset de pobreza en Nueva York del año 2016. El objetivo es identificar las características que más influyen en la empleabilidad y determinar qué modelo predictivo ofrece los mejores resultados a partir de la librería ML-lib suministrada por databricks.

Variables Utilizadas

Las variables seleccionadas para este estudio son principalmente categóricas y se consideran básicas en la recolección de datos por parte de los gobiernos. Estas variables son:

- Etnicidad
- Nivel educativo
- Lugar de nacimiento
- Estado civil
- Género
- Tiene trabajo (variable objetivo)
- Borough (distrito)
- Grupo de edad
- Tiene hijos

Estas variables s**on elegidas debido a su disponibilidad y relevancia en estudios socioeconómicos, lo que podría darnos un acercamiento a estos indicadores considerando que se pueden aplicar años después**. Además, evaluar la empleabilidad formal de los individuos nos permitirá identificar patrones y tendencias a lo largo del tiempo. Este enfoque proporciona una comprensión integral de los factores demográficos y sociales que pueden influir en la empleabilidad de los individuos, ayudando a formular políticas y programas más efectivos para mejorar las oportunidades laborales y reducir la pobreza en Nueva York.

Se usaron 3 técnicas diferentes con distintos parámetros para evaluar los resultados:

1. **Árbol de Decisión (Decision Tree):**

Un árbol de decisión es un modelo que segmenta los datos en función de características específicas, creando reglas de decisión en nodos que conducen a una predicción final.

Máxima profundidad	Accuracy
5	~ 0.76
3	~ 0.76

El árbol de decisión con profundidades máximas de 3 y 5 ha logrado una precisión (accuracy) de aproximadamente 0.76. Esto sugiere que ambos árboles están capturando de manera similar la estructura de los datos y realizando predicciones con una precisión decente. No hay una mejora significativa en la precisión al aumentar la profundidad, lo cual podría indicar que la estructura de los datos no requiere una mayor complejidad para obtener buenos resultados, o que los datos están bien balanceados y no sufren de sobreajuste incluso con una profundidad mayor.

Sin embargo, es mejor que el azar.

2. **Bosque Aleatorio (Random Forest):**

Un bosque aleatorio es un conjunto de árboles de decisión que trabajan juntos para mejorar la precisión de la predicción. Combina múltiples árboles mediante el uso de técnicas de bootstrap y agregación.Mejora la robustez y la generalización del **modelo al reducir el riesgo de sobreajuste** que puede ocurrir con un único árbol de decisión.

Máxima profundidad	Arboles	Accuracy
500	5	~ 0.76
500	3	~ 0.76
100	5	~ 0.76

El bosque aleatorio ha sido evaluado con distintas configuraciones de profundidad máxima y número de árboles, alcanzando una precisión de aproximadamente 0.76 en todos los casos. La consistencia en la precisión sugiere que el modelo es robusto y generaliza bien en diferentes configuraciones, lo cual es una ventaja de los bosques aleatorios sobre los árboles de decisión individuales. A pesar de la alta profundidad máxima, el modelo no parece estar sobreajustado, lo que indica que la técnica de bootstrap y la agregación de múltiples árboles están funcionando eficazmente para mitigar el sobreajuste.

3. **Regresión Logística (Logistic Regression):**

La regresión logística es un modelo estadístico utilizado para predecir la probabilidad de una variable binaria en función de una o más variables independientes. Este adecuado para problemas de clasificación binaria. Sin embargo**, puede no funcionar tan bien con datos de alta dimensionalidad y relaciones no lineales**, lo que podría limitar su efectividad en este contexto específico.

Accuracy: 0.61

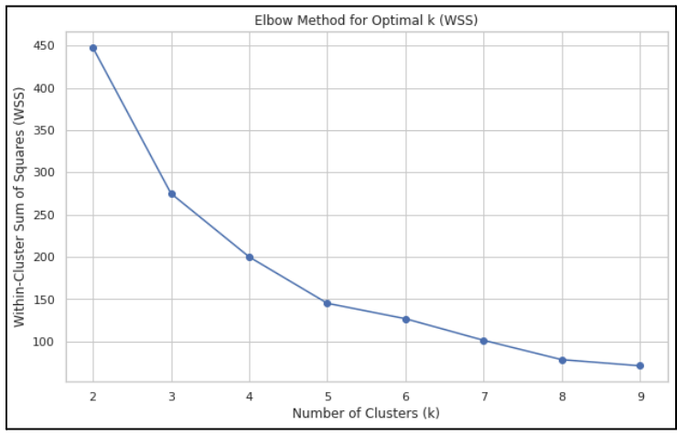
```
Coefficients: (8,[],[])
Intercept: 0.4666029327501598
Feature: Etnicidad_index, Coefficient: 0.0
Feature: nivel_educativo_index, Coefficient: 0.0
Feature: Lugar_Nacimiento_index, Coefficient: 0.0
Feature: Estado_Civil_index, Coefficient: 0.0
Feature: Genero_index, Coefficient: 0.0
Feature: Borough_index, Coefficient: 0.0
Feature: Grupo_Edad_index, Coefficient: 0.0
Feature: Tiene_Hijos_index, Coefficient: 0.0
```

El modelo de regresión logística ha logrado una precisión de 0.61, que es inferior a la obtenida por los árboles de decisión y los bosques aleatorios. Todos los coeficientes del modelo son 0.0, lo que indica que ninguna de las variables incluidas en el modelo tiene un impacto significativo en la predicción de la variable objetivo Tiene_trabajo. Esto podría deberse a varias razones, principalmente por la presencia de muchas variables de tipo factor.

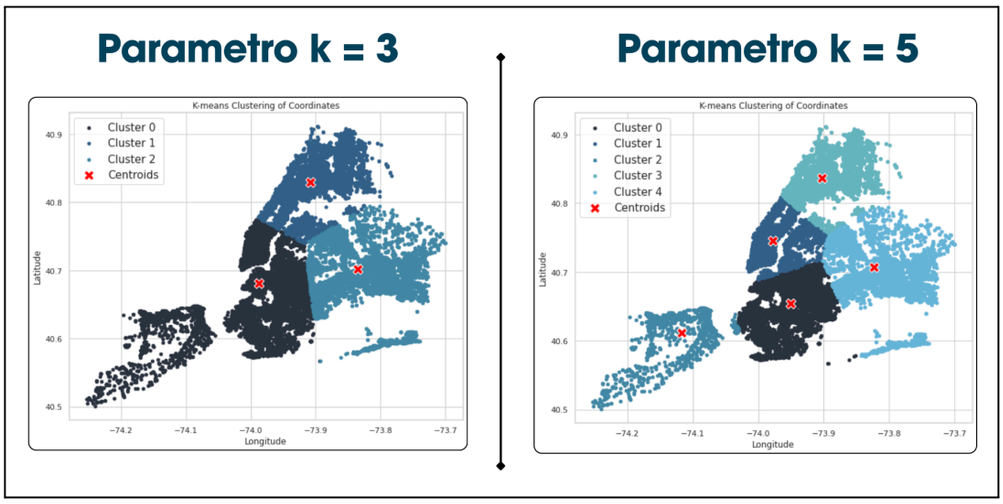
- ## No supervisada

El modelo de aprendizaje de máquina no supervisado tiene como propósito identificar **patrones geoespaciales de incidencia de robos, lo que permite a las autoridades focalizar recursos de seguridad** así como mejorar la prevención en áreas específicas, optimizar patrullajes y estrategias de vigilancia, y demás posibles estrategias para ayudar en zonas de alta incidencia, mejorando así la eficiencia en la respuesta y prevención de robos y aumentando la seguridad pública en la ciudad.

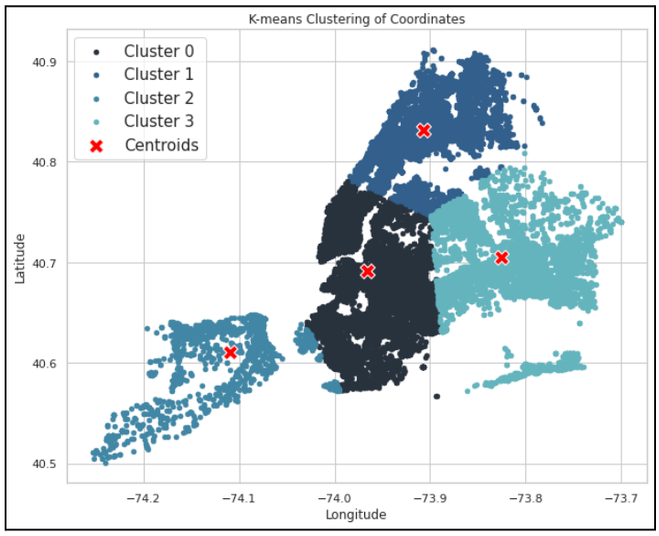
Nuestro modelo **K-Means** se desarrolló desde la librería de ML-lib de spark, en el dataset de robos en la ciudad de nueva york ya que por medio del clustering se proporciona una visión clara de cómo se distribuyen los arrestos en la ciudad de Nueva York, para lograrlo usamos unas variables características que serían nuestra entrada, elegimos entonces con el fin de tener una visión **geoespacial** la **latitud y la longitud**, para ello borramos algunos valores que no tenían sentido en la base de datos ya que por una razón u otra están por fuera de lo que serían los rangos de la ciudad que nunca duerme, finalmente para escoger el mejor ***hiperparametro k*** el método del codo fue quien nos ayudaría a tomar la decisión, con lo cuál el gráfico quedó así:



Sin embargo decidimos probar con k = 3 y K = 4 para jugar un poco con el hiperparámetro con lo que nos quedarón los siguientes clusters y centroides definidos:



Empero dada la información que nos proporciona el método del codo vemos que el mejor k posible es ***4*** por lo que fue nuestro elegido para realizar el **K-Means** final, una vez ajustado el modelo asignamos cada punto a su respectivo cluster, con lo que nuestro resultado final fue:



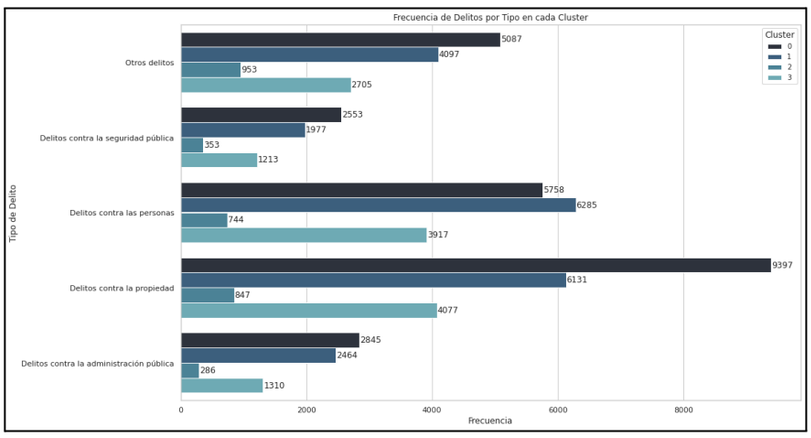
Con respecto al ***Cluster 0*** vemos que se encuentra principalmente al suroeste de la ciudad, con lo cuál esta área podría corresponder a partes del distrito financiero y vecindarios residenciales adyacentes. En cuanto al ***Cluster 1*** predomina en la parte central y sureste de la ciudad, posiblemente incluyendo áreas de Manhattan, por tanto es probable que abarque áreas altamente urbanizadas y de alta actividad comercial. Frente al ***Cluster 2*** se extiende a lo largo de la parte central de la ciudad hacia el noreste, puede representar áreas residenciales densamente pobladas como las que puede contener el distrito de Queens. Finalmente el ***Cluster 3*** está situado al sureste de la ciudad lo que en su mayoría incluye áreas mixtas de residencial y comercial, además de puntos de transporte importantes.

Análisis De Cada Cluster

Ya que K-means ayuda a identificar patrones ocultos en los datos de arrestos que no serían fácilmente visibles de otra manera debido a que agrupa los datos geográficamente, analizar las características específicas de cada cluster permite desarrollar estrategias de prevención del crimen personalizadas por tanto la descripción de cada cluster es vital.

- **Frecuencia De Delitos Por Cluster**

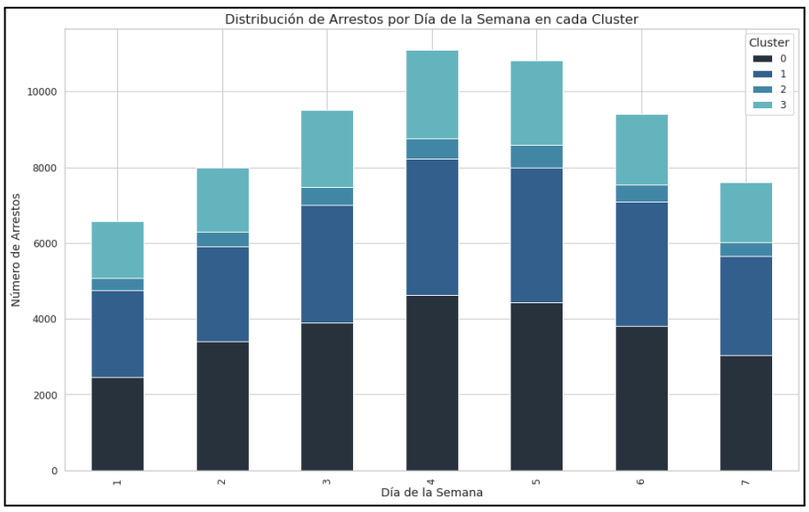
La frecuencia de diversos tipos de delitos en cada uno de los clusters identificados mediante el análisis de K-means en la ciudad de Nueva York es fundamental debido a que con esa información permite que las autoridades puedan planificar e implementar estrategias específicas de seguridad y prevención del crimen basadas en datos concretos, a través de un entendimiento claro de los tipos de delitos predominantes en cada cluster, se pueden asignar recursos de manera más eficiente y mejorar la respuesta policial.



Como aspectos importantes que vemos en la gráfica tenemos que el ***Cluster 0*** muestra una alta incidencia de delitos contra la propiedad, indicando la necesidad de fortalecer las medidas de seguridad física en estas áreas, de igual manera la alta frecuencia de delitos contra las personas en los ***Clusters 0 y 1*** sugiere una aproximación a la comunidad con medidas como programas de intervención comunitaria, apoyo psicológico o mayor vigilancia a comportamientos inusuales.

- **Distribución De Arrestos Por Día De La Semana**

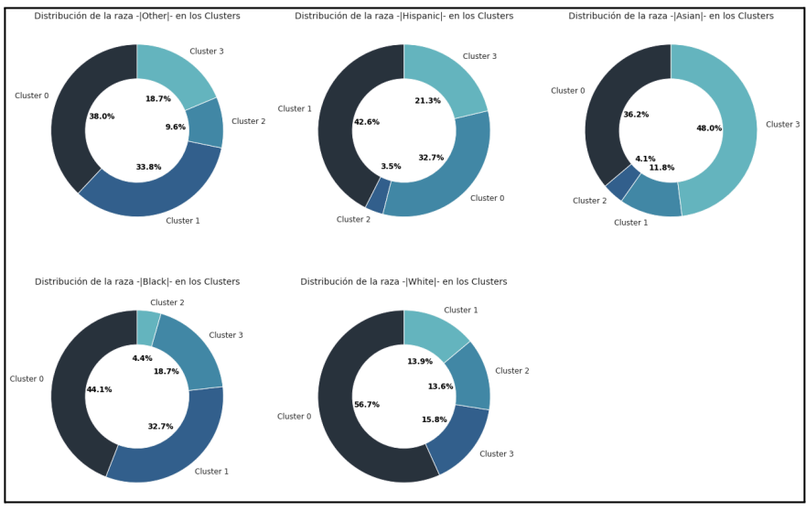
Analizar la distribución de arrestos por día de la semana en cada cluster de la ciudad de Nueva York es crucial para identificar patrones temporales en la actividad delictiva y planificar la asignación de recursos policiales de manera más efectiva.



Por medio de la gráfica revela que los arrestos alcanzan su pico máximo los miércoles y jueves en todos los clusters, indicando un aumento en la actividad delictiva a mediados de la semana, el ***Cluster 0*** es quien parece tener más arrestos en estos días de mitad de semana, por último es notable ver que los fines de semana, especialmente el domingo, tienen una menor frecuencia de arrestos, a través de lo anterior vemos que se puede asignar mucho más personal en especial en el ***Cluster 0*** para la mitad de la semana, mientras que para los fines de semana se requiere menos personal.

- **Distribución De Las Razas En Los Clusters**

La distribución racial de los arrestos en cada cluster es crucial para comprender cómo diferentes comunidades son afectadas por la actividad delictiva y planificar estrategias de intervención y prevención más equitativas, así mismo la identificación de disparidades raciales en los arrestos permite a las autoridades abordar posibles sesgos y mejorar las relaciones comunitarias, lo que adquiere mayor importancia al ser Nueva York una ciudad tan diversa.



En cuanto al análisis vemos que la gráfica se destaca el ***Cluster 3*** por una alta representación de la raza Asian que tiene un 48%, a través de las distribuciones se logra que las autoridades puedan focalizar recursos y programas específicos en cada comunidad, promoviendo sobretodo la equidad y justicia en la aplicación de la ley.

Para finalizar el clustering por medio de K-Means el modelo es fundamental en el contexto del **análisis de datos geoespaciales y de criminalidad en la ciudad de Nueva York** por ya que ayuda a las autoridades correspondientes en aspectos que ya hemos mencionado, por tanto y ya teniendo en mente el análisis que el K-Means permitió realizar las dos estrategías propuestas destinadas a reducir la cantidad de arrestos son:

1. **Desarrollar iniciativas educativas y de inclusión** para las comunidades con alta incidencia de arrestos, como las razas Black y Hispanic en los Clusters 0 y 1, **para reducir los sesgos raciales y mejorar las relaciones comunitarias**

2. **Aumentar el acompañamiento durante los picos de actividad delictiva a mediados de la semana**, especialmente en el Cluster 0, y redistribuir el personal durante los fines de semana **para optimizar la respuesta policial.**

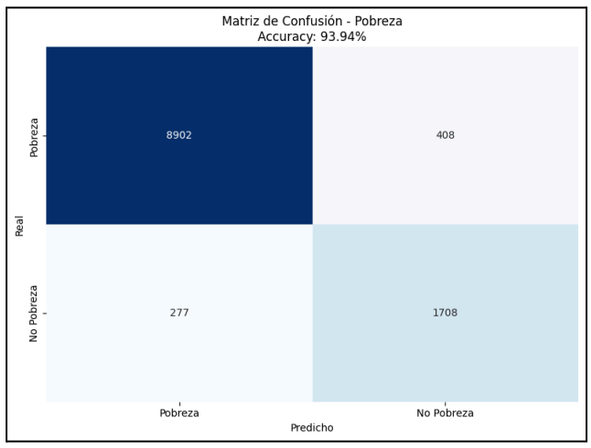
En conclusión implementar programas de equidad y diversidad, junto con una asignación eficiente de recursos policiales, permitirá abordar los sesgos raciales y optimizar la respuesta a la actividad delictiva, mejorando así la seguridad y la justicia en Nueva York.

5. Red Neuronal Bono

Para la red neuronal, se utilizó el entorno de colab, ya que databricks generó problemas para encender los clusters, a través de la red neuronal se busco predecir la categoría de ***Estatus_Pobreza*** con las demás variables presentes en el dataset que son en total 32 columnas entre las que se encuentra toas las presentes del dataset curado junto con otras nueva generadas de variables dummies que fueron necesaria crear para las variables categóricas de ***Borough***, ***Etnicidad***, ***Grupo_Edad*** y ***Genero***, debido a que la red neuronal no recibe categóricas en texto, finalmente entre parámetros y más ajuste iniciales se tuvieron en cuenta:

- **Capa De Entrada:** 64 neuronas y función de activación ReLU.
- **Capa Intermedia:** 32 neuronas y también ReLU.
- **Capa De Salida**: 2 Neuronas ya que tenemos 2 posibles valores y una activación llamada **“Softmax”** que ayuda para muchas clases.
- **Función De Pérdida:** tipo *“categorical_crossentropy”* que ayuda para múltiples clases.
- **Optimizador Aprendizaje:** tipo *“Adam”*.
- **Epochs:** 25 para que no sea tan demorado pero sea un buen número.

Finalmente después de ajustar el modelo salió la siguiente matriz de confusión:



Con lo que obtuvimos un ***Accuracy 93.94%*** obteniendo lo siguiente:

- **Verdaderos Positivos (Pobreza correcta):** 1708 de 1985 (86.05%)
- **Falsos Negativos (Pobreza no correcta):** 277 de 1985 (13.95%)
- **Verdaderos Negativos (No Pobreza correcta):** 8902 de 9310 (95.62%)
- **Falsos Positivos (Pobreza no correcta):** 408 de 9310 (4.38%)

Concluyendo vemos que su alta precisión permite asignar recursos de manera más eficiente, sin embargo, el margen de error subraya la necesidad de combinar la tecnología con la realidad de mirar y corroborar esos pequeños casos que escapan del modelo, pero en general ayuda mucho para hacernos una idea de cómo con las variables indicadas pueden ayudar a mejorar y desde ahí empezar a implementar políticas que ayuden a mejorar estas mismas.

6. Anexos

- Link Al Video De Sustentación: ***Video***
- Link Al Código De Preguntas 1 - 4: ***Preguntas Proy 2-1-4.ipynb***
- Link Al Código De Preguntas 4 - 8: ***Preguntas 5-8.pbix***
- Link Al Código De Aprendizaje Supervisado: ***Modelo-Supervisado.ipynb***
- Link Al Código De Aprendizaje No Supervisado: ***Modelo K-Means - Aprendizaje No Supervisado-2.ipynb***
- Evidencia De Ejecución Databricks: ***Evidencia_No_Supervisado*** **y** ***Evidencia_Supervisado***
- Link Al Colab Del Bono: ***Bono-Procesamiento-2-Pobreza***
- Presentación: ***Plan_Estratégico_Nueva_York.pdf***