Proyecto - Procesamiento de datos a gran escala

Entrega 2 - Preparación de datos, modelado y presentación de resultados

Procesamiento de datos a gran escala

Santiago Botero Pacheco

Santiago Rueda Pineda

Santiago Avilés Tibocha

Brayan Steven Carrillo Mora

Pontificia Universidad
JAVERIANA
Colombia

Tabla de contenido

Tabla de graficas

1. Filtros y transformaciones
    a. Filtrado

    Para el caso del conjunto de datos de pobreza en Nueva York, se realizan los siguientes filtros:

    - Eliminación de registros de nulos

      Esta es la opción mas usual en el caso de filtrado, debido a que en muchas ocasiones los datos nulos no representan una cantidad relevante para el estudio, por lo cual el proceso de imputación o cualquier otra medida puede resultar costoso en términos de esfuerzo y poco productivo, lo cual no corresponde a un balance satisfactorio. Este método fue implementado para las columnas edad y educación en el conjunto de datos correspondiente a la pobreza en Nueva York, debido a que sus proporciones de 1% y 3,08% respectivamente no constituyen mayor relevancia para el estudio del proyecto. También este método es implementado en el conjunto de datos correspondientes a los arrestos en Nueva York, debido a la poca cantidad de nulos que se registran en las columnas a excepción del nivel de ofensa.
    - Utilización de algoritmo no supervisado para adaptar el campo nulo.

      Esta opción es el fruto de una ardua tarea por buscar obtener la mejor calidad de datos posible, debido a que es utilizada para hallar los valores nulos correspondientes a la columna de estado civil en el conjunto de datos de pobreza en Nueva York, por lo cual un filtrado a partir de los resultados de un algoritmo no supervisado permite obtener con cierto nivel de exactitud un valor veraz para esta columna.
    - Interés de columnas

      Algunas columnas en ambos conjuntos de datos seleccionados no representan mayor interés para las preguntas planteadas respecto a la ejecución del proyecto, por lo cual este filtro evidencia efectividad para evitar ruido durante el estudio, además de un ahorro de recursos durante las diferentes etapas de proceso que se puedan llevar a cabo en el desarrollo del proyecto.

b. Transformación

- Estandarización de edad

  Para la primera transformación en el conjunto de datos de pobreza en Nueva York, se estandariza la edad, de acuerdo con los ingresos antes de impuestos, debido a que, en algunos registros, se encuentran edades en 0, pero se cuenta con ingresos antes de impuesto, lo cual no tiene sentido. Debido a esto, se establece como información faltante la edad en dichos registros, y se convierte en nulo, para en una etapa de filtrado establecer que hacer en dichos registros.

- Conversión de horas trabajadas a semanas trabajadas:

  La segunda transformación corresponde a la conversión de horas trabajadas a semanas trabajadas para los registros los cuales contengan nulos en esta última columna, esto debido a que la columna horas trabajadas no contiene datos nulos, lo cual nos permite contar con este registro y evitar su eliminación.

- Imputación de datos con la moda

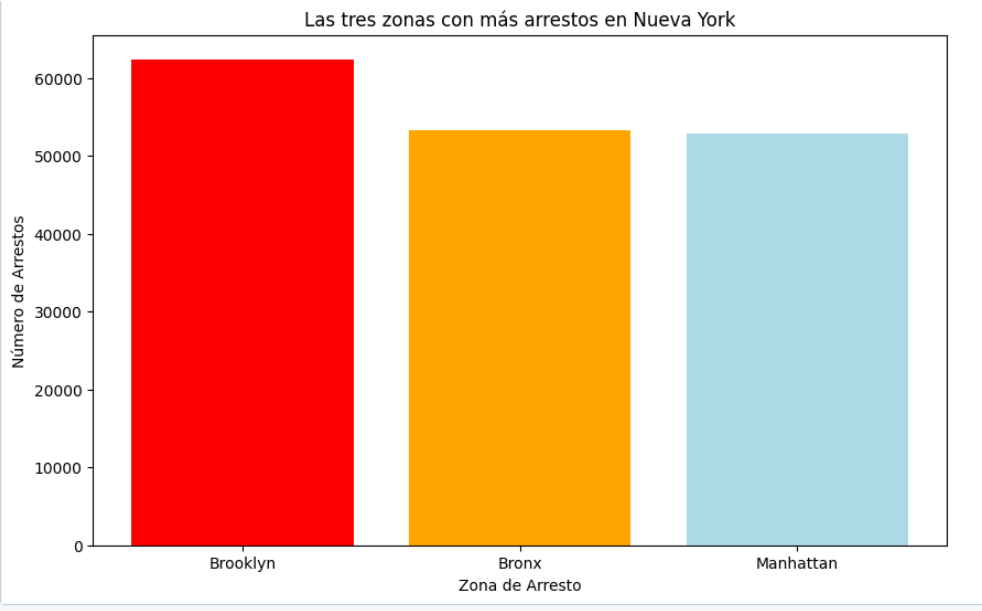
  La tercera transformación corresponde a la implementación de imputación de datos en los registros que contienen nulos en la columna nivel de ofensa o LAW_CAT_CD, esta transformación se debe a que la columna puede ser de interés para llevar a cabo la realización del proyecto, además que solo el 9% de la columna contiene nulos, y se obtiene menos ruido a partir de la imputación con respecto a la eliminación de dichos registros.

2. Respuesta a las preguntas

   Para la realización del proyecto anteriormente se habían propuesto las siguientes preguntas:

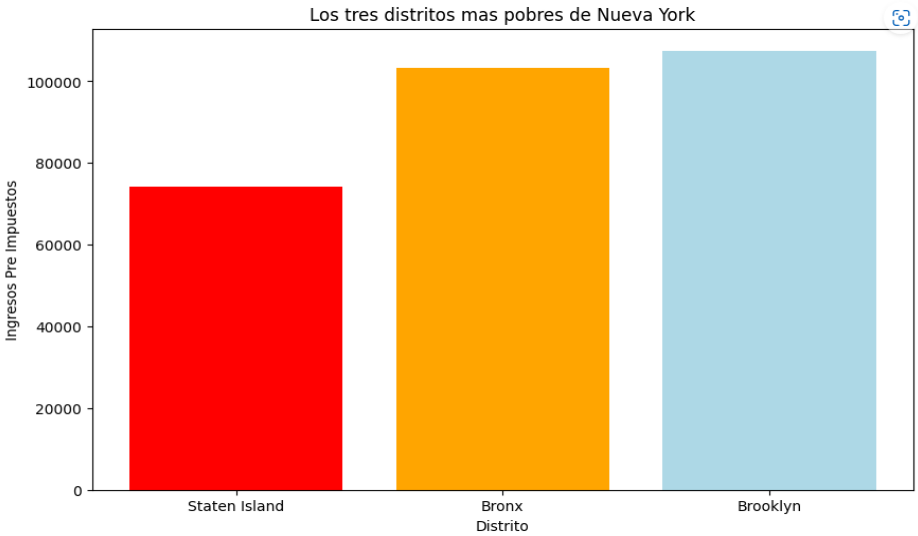
   a. ¿Son las 3 zonas más pobres de Nueva York, las mismas con más arrestos?

   Para resolver esta pregunta, se tuvo en cuenta ambos conjuntos de datos propuestos para el proyecto, por lo cual, se requirió obtener las tres zonas más pobres de acuerdo con el conjunto de datos de pobreza de Nueva York, y las 3 zonas con más arrestos, esto se logra, gracias a las graficas propuestas con este objetivo.



Grafica 1 - Top 3 zonas con más arrestos en Nueva York

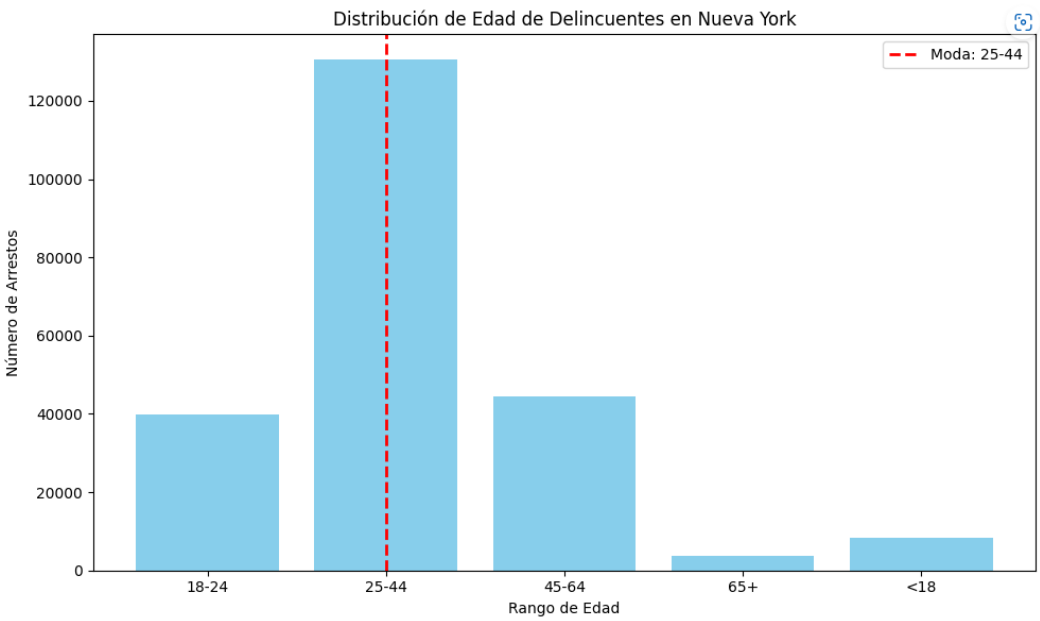
Como se presenta en la Grafica 1 - Top 3 zonas con más arrestos en Nueva York y Grafica 2 - Zonas con menores ingresos antes de impuestos en Nueva York, las cuales nos muestran las zonas mencionadas en la pregunta e identificamos que si bien tanto Bronx, como Brooklyn aparecen en ambas gráficas, el top 3, se complementa con otra distrito en cada uno de los casos, siendo Manhattan y Staten Island para cada una de las gráficas respectivamente.



Grafica 2 - Zonas con menores ingresos antes de impuestos en Nueva York

b. ¿Cuál es la moda para el rango de edad del delincuente en Nueva York?

Para el caso del rango de edad del delincuente en Nueva York, buscamos cual es el grupo etario mas propenso a cometer delitos, esto con el objetivo de ofrecer posibilidades a los ciudadanos presentes en dicho grupo, y así reducir los índices de criminalidad.

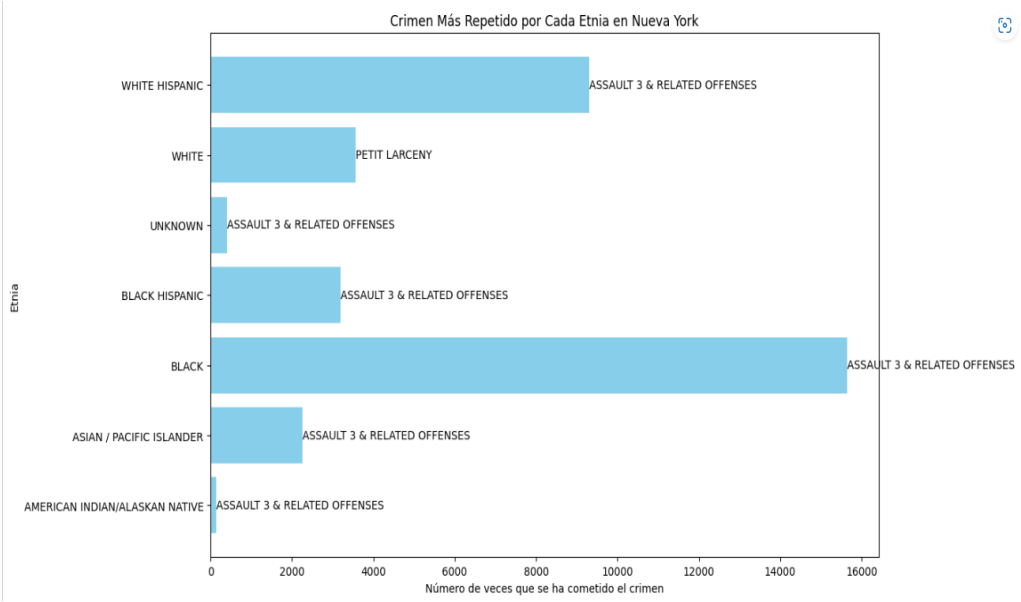


Grafica 3 - Distribución de rangos de edad en los crímenes en Nueva York

Como logramos observar en la Grafica 3 - Distribución de rangos de edad en los crímenes en Nueva York, encontramos que el grupo etario el cual comprende los rangos de edad entre 25 y 44 años es la población más susceptible a cometer delitos, por lo cual, este grupo es de atención prioritaria al momento de implementar medidas sociales que busquen disminuir los índices de criminalidad. También es interesante ver una participación significativa de menores de edad, por lo cual podrían implementarse algunas medidas con el fin de erradicar los índices de arrestos en dicho grupo etario.

c. ¿Cuál es el crimen más repetido por cada etnia?

Esta pregunta busca identificar cuales son las razones que pueden llevar a un individuo a cometer posibles crímenes, según las necesidades de su grupo poblacional y su realidad, sin embargo, a continuación, se presentan los delitos mas cometidos por cada etnia.

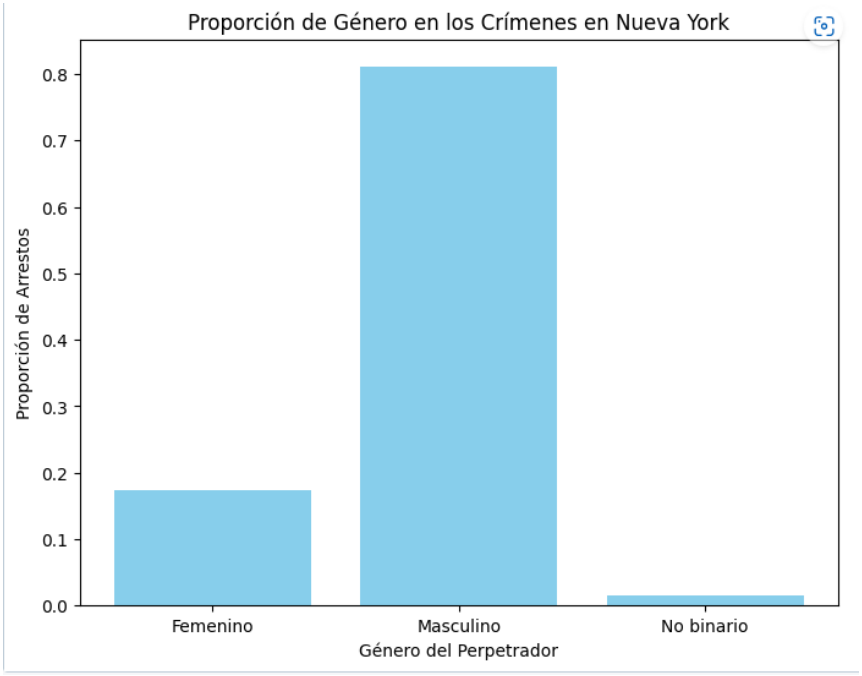


Grafica 4 - Delito más cometido por etnia

En la Grafica 4 - Delito más cometido por etnia, podemos observar que en casi todas las etnias a excepción del grupo étnico caucásico, el delito más común es asalto y ofensas relacionadas, esto traduce en un crimen como lo es el robo a mano armada, esto implica que la mayoría de las personas en estos grupos poblacionales cometen delitos por necesidad, por lo cual la implementación de algunas políticas sociales sería ideal para la reducción de los crímenes cometidos por individuos pertenecientes a dichos grupos étnicos.

d. ¿Cuál es la proporción de género en los crímenes en Nueva York?

Esta pregunta va encaminada a verificar si la implementación de políticas de género surge efecto y reduce la necesidad del genero femenino en exponer su integridad al momento de cometer un delito, gracias a la priorización que obtiene al interior de las políticas gubernamentales adoptadas por las grandes urbes alrededor del mundo, en especial Nueva York.



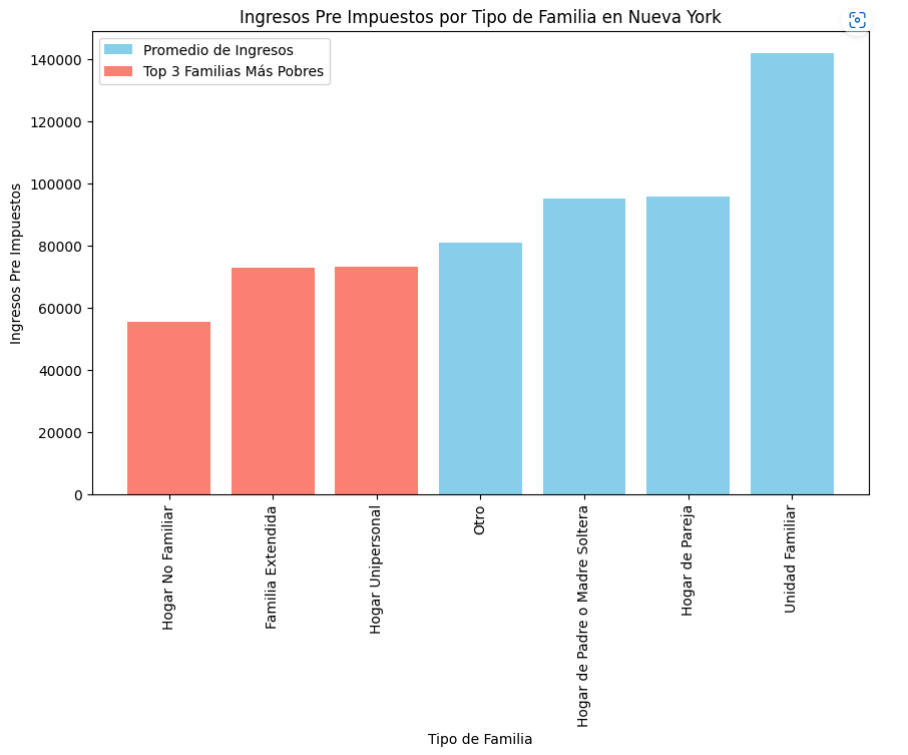
Grafica 5 - Distribución de género en los crímenes cometidos en Nueva York

En la Grafica 5 - Distribución de género en los crímenes cometidos en Nueva York podemos observar una predominancia del genero masculino en la participación de delitos, lo cual comprueba la afectividad de las políticas adoptadas para generar un cambio en las posibilidades de las personas identificadas al interior del genero femenino, a su vez, la ciudad propone un espacio de inclusión para las personas que se identifican como no binarias, y como se refleja, una participación relativamente significativa.

e. ¿Cuáles son los 3 tipos de familia más pobres en Nueva York?

A continuación, se presentan los tipos de familias presentes en el conjunto de datos de pobreza en Nueva York, el cual fue propuesto para el desarrollo del actual parcial, esta grafica se concibe a través de los ingresos medios antes de impuestos de los integrantes que conforman el grupo familiar

.



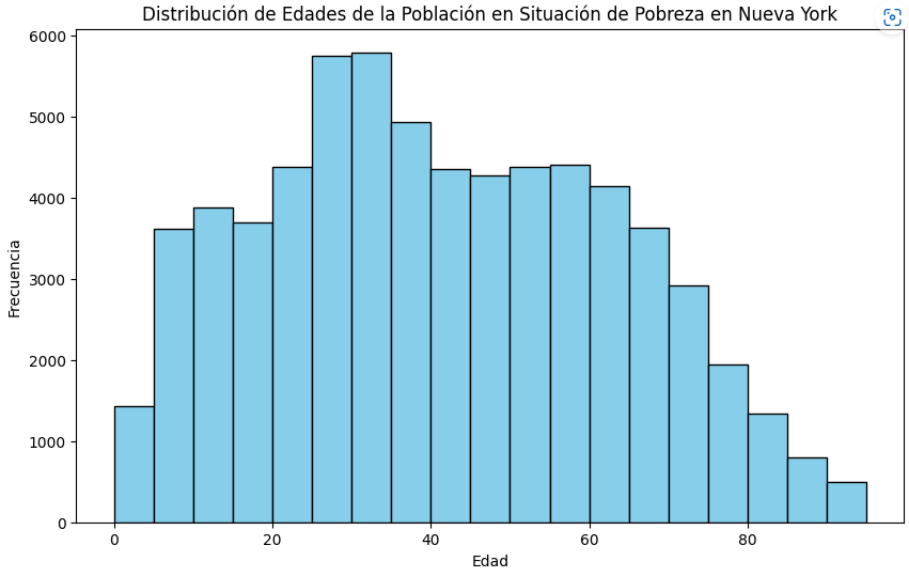
Grafica 6 - Tipos de familia en pobreza en Nueva York

En la Grafica 6 - Tipos de familia en pobreza en Nueva York, encontramos que los tipos familiares más pobres en la ciudad de Nueva York, son el Hogar no familiar, el cual comprende las edificaciones en donde conviven varias personas las cuales no sostienen algún tipo de relación en particular, el segundo tipo de hogar más pobre es la familia extendida, debido a que su naturaleza numerosa, equivale a una mayor cifra de gastos. Por último, se encuentra en tercer lugar de los tipos de familia más pobres en Nueva York, los hogares unificados, o los conformados por una sola persona, esta cifra es aumentada debido al periodo de adaptación de las personas al lograr emanciparse de la custodia parental, sin embargo, encuentra cierto balance en los individuos profesionales, los cuales no tienen otro ser humano a su cargo o responsabilidad.

f. ¿Cuál es la distribución de edades de la población en situación de pobreza en Nueva York?

Esta pregunta esta encaminada a la identificación e grupos etarios, los cuales puedan encontrarse en riesgo de caer en pobreza, sin embargo y como se refleja en la Grafica 7 - Distribución de las edades de la población en pobreza en Nueva York Encontramos que la distribución obedece a una forma casi equitativa, la cual se encuentra un poco sesgada a la izquierda, presentando picos en el rango

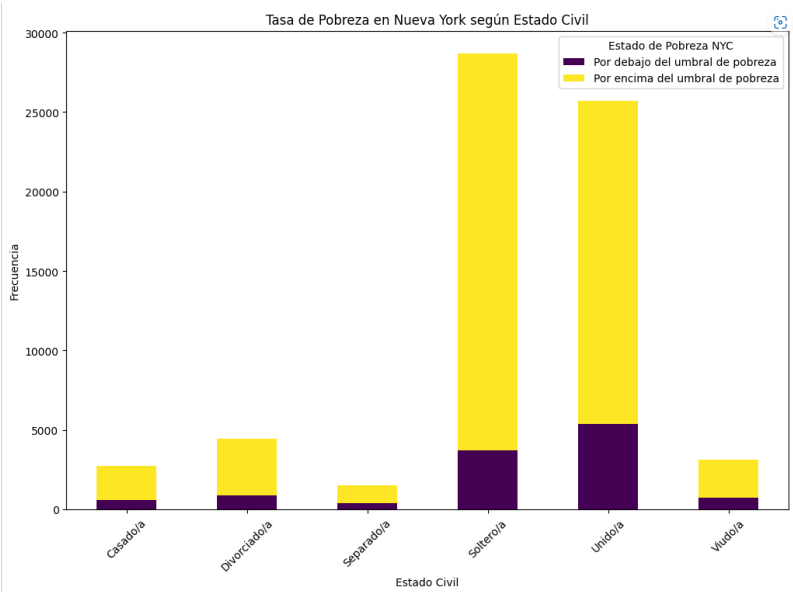
poblacional entre 25 y 35 años, el cual curiosamente también se encuentra en el grupo etario con mayor cantidad de crímenes registrados en la ciudad de Nueva York.



Grafica 7 - Distribución de las edades de la población en pobreza en Nueva York

g. ¿Cómo varía la tasa de pobreza en función del estado civil de los residentes de Nueva York?

En esta pregunta, buscamos validar la condición de las personas en pobreza de Nueva York, y su afectación en los individuos directamente relacionados con dicha persona.

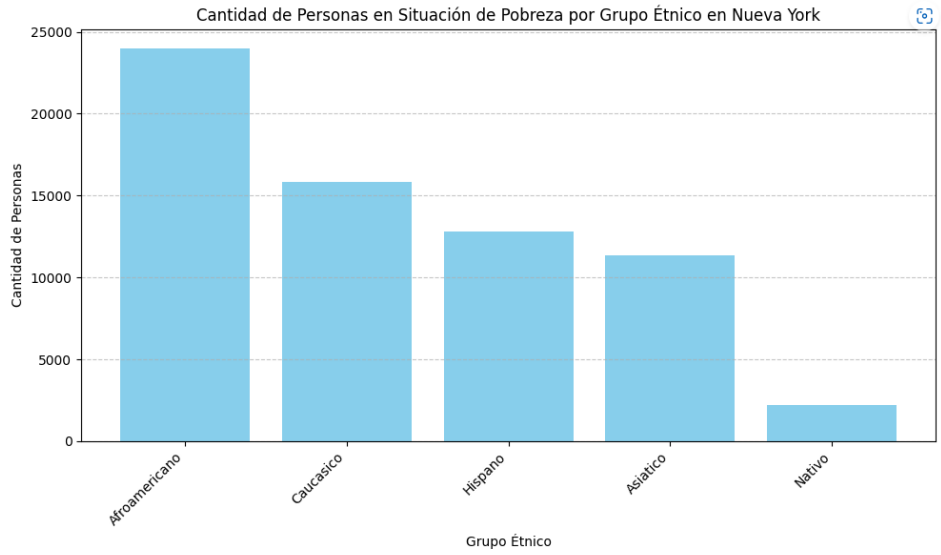


Grafica 8 - Tasa de pobreza según estado civil en Nueva York

En la Grafica 8 - Tasa de pobreza según estado civil en Nueva York, se muestra un grafico de barras apiladas donde encontramos que si bien el estado civil de los habitantes no juega un papel importante en la situación de pobreza en Nueva York, si predominan los estados civil: Soltero/a y Unido/a en la información contenida en el conjunto de datos provisto para la realización del proyecto, en este apartado se pueden generar políticas para que las personas en unión libre, busquen formalizar su unión ante la ley, a pesar de su situación económica.

h. ¿Cuál es la distribución de la etnia en el nivel de pobreza en Nueva York?

Esta pregunta, busca identificar cuales son los grupos étnicos con mayor peligro para caer en situación de pobreza, lo cual puede servir como indicador para una posible intervención al interior de dichas comunidades, intervenciones las cuales pueden suponer un apoyo realmente necesario para los individuos que conforman dichos grupos poblacionales.

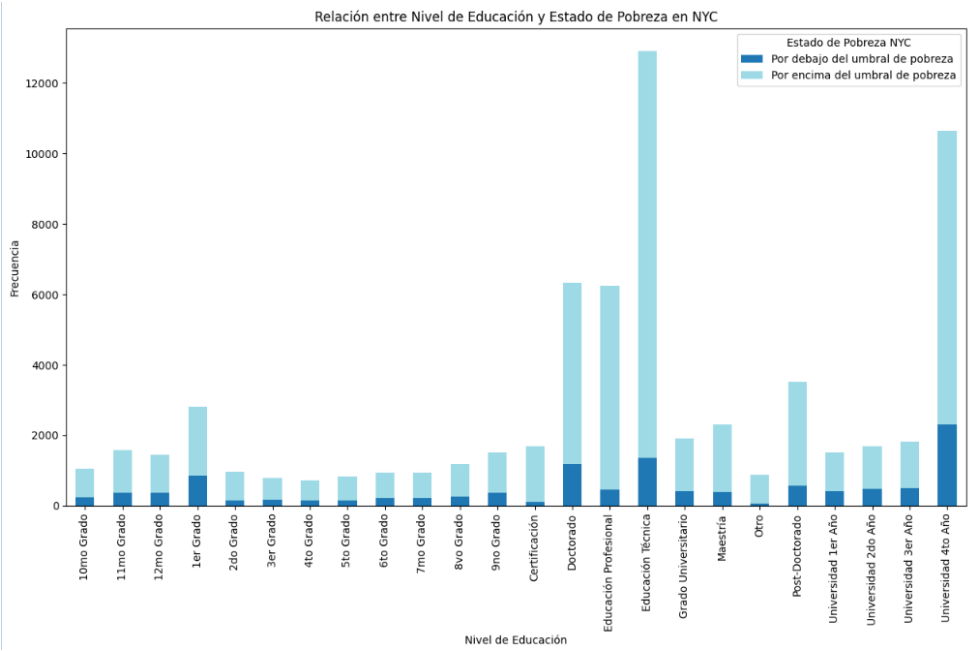


Grafica 9 - Distribución de la pobreza según el grupo étnico

Como se observa en la Grafica 9 - Distribución de la pobreza según el grupo étnico, se debe prestar especial atención a la comunidad afroamericana, ya que dicha comunidad es la más susceptible de caer en situación de pobreza, esto puede deberse a diversos factores, incluido el racismo naturalizado en algunas comunidades alrededor de Nueva York, lo cual retrocede en los esfuerzos de conformar una única comunidad con el propósito de progresar constantemente a través de la unión.

i. ¿Cuál es la relación entre el nivel educativo y el riesgo de caer en la pobreza en Nueva York?

En esta gráfica, se busca comparar si hay una incidencia significativa en el nivel de educación poseído por una persona, con respecto la clasificación de pobreza en el cual se encuentra, este indicador se consigue a través de un grafico de barras apiladas como se puede observar a continuación.



Grafica 10 - Relación entre el nivel de educación y el estado de pobreza en la ciudad de Nueva York

Como se visualiza en la Grafica 10 - Relación entre el nivel de educación y el estado de pobreza en la ciudad de Nueva York, no hay un índice que justifique una relación directa entre el nivel educativo poseído por un individuo y su clasificación en el umbral de pobreza, ya que se observan grandes proporciones en cada uno de los niveles educativos, sin embargo, si se observa una cantidad alarmante de deserción en algunos niveles escolares alimentarios, lo cual pudo contribuir directamente a la situación de pobreza al interior de dicho grupo demográfico.

j. ¿Qué factor tiene mayor incidencia en la delincuencia? ¿Falta de educación o pobreza?

Para esta pregunta, no se generó una grafica especifica, ya que se puede llegar a una conclusión al respecto a través del análisis exhaustivo de la Grafica 3 - Distribución de rangos de edad en los crímenes en Nueva York, y la Grafica 7 - Distribución de las edades de la población en pobreza en Nueva York, donde se puede observar un pico en el grupo etario entre 25 y 35 años en ambas gráficas, por lo cual se puede deducir una relación directa, en la cual la pobreza influye en gran medida a los individuos que pueden llegar a cometer crímenes

exclusivamente por las necesidades que puedan presentar. Adicionalmente, la Grafica 10 - Relación entre el nivel de educación y el estado de pobreza en la ciudad de Nueva York, no muestra un índice de relación directa en la pobreza, y siguiendo el hilo anterior de la investigación no se puede concluir a partir de los datos, que presente una relación directa sobre los índices de criminalidad en la ciudad de Nueva York, por lo cual, a partir de los datos procesados, la pobreza incide más sobre la criminalidad que la falta de educación.

3. Selección de técnicas de aprendizaje

    a. Aprendizaje Supervisado: Random Forest

    El aprendizaje supervisado se utiliza cuando se dispone de un conjunto de datos etiquetado, es decir, donde se conoce la variable objetivo que se desea predecir. Para este caso, un modelo de Random Forest es una excelente opción.

Justificación:

- Predicción de delitos: Ya que el objetivo del negocio es predecir o clasificar el tipo de delito, entonces Random Forest es una buena opción. Este algoritmo es robusto y puede manejar múltiples tipos de características y es menos propenso al sobreajuste en comparación con otros modelos más complejos.

- Manejo de desbalanceo de clases: Los conjuntos de datos de crímenes pueden tener clases desbalanceadas (por ejemplo, ciertos tipos de delitos ocurren con menos frecuencia). Random Forest puede manejar naturalmente este desequilibrio y generar predicciones precisas incluso en tales casos.

- Interpretabilidad: Aunque no es tan interpretable como algunos modelos lineales, Random Forest proporciona una idea de la importancia de cada característica en la predicción, lo que puede ser útil para comprender qué factores están relacionados con ciertos tipos de delitos.

    b. Aprendizaje No Supervisado: Análisis de Componentes Principales (PCA)

Justificación:

- Reducción de dimensionalidad: PCA se usa para reducir la dimensionalidad de los datos al transformar variables correlacionadas en un conjunto menor de variables no correlacionadas llamadas componentes principales. Al tener un gran número de características y variables en tus datos, PCA puede ayudar a reducir esta dimensionalidad, lo que puede simplificar el análisis y mejorar la eficiencia computacional.

- Visualización de datos: PCA también se utiliza comúnmente para la visualización de datos al proyectar los datos en un espacio de menor dimensión. Esto puede ayudar a detectar patrones, agrupamientos o tendencias en los datos que pueden no ser evidentes en su forma original de alta dimensionalidad. Identificación de características importantes: Además de la reducción de dimensionalidad, PCA puede ayudar a identificar las características más importantes o influyentes en el conjunto de datos. Esto puede ser útil para comprender qué características contribuyen más a la variabilidad en los datos y pueden ser importantes para predecir ciertos resultados, como los tipos de delitos o las tasas de criminalidad en diferentes áreas.

4. Preparación de los datos

    a. 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:

- Conjunto de datos – Pobreza en Nueva York
A continuación, se explica a detalle el proceso llevado a cabo para eliminar las variables fuertemente correlacionadas, con un índice de correlación superior al 80% en el

conjunto de datos de pobreza en Nueva York.

```
# Convertir a DataFrame de pandas para calcular la correlación
poverty_nonull_pd = poverty_nonull.toPandas()

# Calcular la matriz de correlación
correlation_matrix = poverty_nonull_pd.corr()

# Identificar características fuertemente correlacionadas
threshold = 0.8
upper_triangle = correlation_matrix.where(np.triu(np.ones(correlation_matrix.shape), k=1).astype(bool))
to_drop = [column for column in upper_triangle.columns if any(upper_triangle[column] > threshold)]

# Eliminar características fuertemente correlacionadas
poverty_nonull_pd.drop(columns=to_drop, inplace=True)

# Convertir de nuevo a DataFrame de Spark
data = spark.createDataFrame(poverty_nonull_pd)


data.printSchema()
```

Grafica 11 – Pobreza en Nueva York - Eliminación de correlaciones

En la Grafica 11 – Pobreza en Nueva York - Eliminación de correlaciones se observa cómo se identificaron las características fuertemente correlacionadas, a la vez que se establece un umbral de correlación (threshold) en 0.8. Luego, se crea una matriz triangular superior (upper_triangle) a partir de la matriz de correlación, donde se establecen como verdaderos los valores que están por encima de la diagonal principal (es decir, solo los valores de correlación entre diferentes características, no con ellas mismas). Luego, se identifican las características que tienen una correlación superior al umbral establecido para posteriormente eliminarlas.

Conjunto de datos – Arrestos en Nueva York:

Análogamente al conjunto de datos de pobreza en Nueva York, se presenta el procedimiento para la eliminación de variables o columnas fuertemente correlacionadas en el conjunto de datos de arrestos en Nueva York, como se observa a continuación:

```
#Analogamente con el segundo dataframe



# Convertir a DataFrame de pandas para calcular la correlación
df_arrestos_ny_mode_pd = df_arrestos_ny_mode.toPandas()

# Calcular la matriz de correlación
correlation_matrix = df_arrestos_ny_mode_pd.corr()

# Identificar características fuertemente correlacionadas
threshold = 0.8
upper_triangle = correlation_matrix.where(np.triu(np.ones(correlation_matrix.shape), k=1).astype(bool))
to_drop = [column for column in upper_triangle.columns if any(upper_triangle[column] > threshold)]

# Eliminar características fuertemente correlacionadas
df_arrestos_ny_mode_pd.drop(columns=to_drop, inplace=True)

# Convertir de nuevo a DataFrame de Spark
data2 = spark.createDataFrame(df_arrestos_ny_mode_pd)


data2.printSchema()
```

Grafica 12 - Arrestos en Nueva York - Eliminación de correlaciones

En la Grafica 12 - Arrestos en Nueva York - Eliminación de correlaciones se muestra el código necesario para implementar la eliminación de variables o columnas fuertemente correlacionadas cuyo índice supera el 80%, con el fin de preparar el conjunto de datos para un procesamiento efectivo en las etapas posteriores del proyecto.

b. Normalización de variables numéricas:

En este apartado se espera que se realice una normalización de datos si es el caso:

Dado que el conjunto de datos de pobreza en Nueva York ya contiene totalidad de columnas en formato numérico, solo tenemos que normalizar el conjunto de datos de

arrestos y eliminar columnas que no se utilizaran.

```
from pyspark.sql.functions import col, when, lit
from pyspark.sql import SparkSession
from itertools import chain


data2 = data2.drop("ARREST_DATE" , 'New Georeferenced Column')

categorical_cols = ['PD_DESC', 'OFNS_DESC', 'LAW_CODE', 'LAW_CAT_CD', 'ARREST_BORO', 'AGE_GROUP', 'PERP_SEX', 'PERP_RACE']



for col_name in categorical_cols:
    if col_name == 'PERP_SEX':
        # Específico para la columna de género
        data2 = data2.withColumn(col_name, when(col(col_name) == 'M', 1).when(col(col_name) == 'F', 0).otherwise(None))
    else:
        # Obtener los valores únicos de la columna
        unique_values = data2.select(col_name).distinct().rdd.flatMap(lambda x: x).collect()

        # Crear un mapeo de los valores únicos a códigos numéricos
        mapping = {val: idx for idx, val in enumerate(unique_values)}

        # Crear una expresión de mapeo
        mapping_expr = create_map([lit(x) for x in chain(*mapping.items())])

        # Aplicar el mapeo al DataFrame y manejar valores nulos
        data2 = data2.withColumn(col_name, mapping_expr[col(col_name)].cast('int'))
```

Grafica 13 - Normalización del conjunto de datos - Arrestos en Nueva York

En la Grafica 13 - Normalización del conjunto de datos - Arrestos en Nueva York, se muestra el código requerido para normalizar el conjunto de datos de arrestos en Nueva York, lo cual convierte las variables categóricas presentes al interior de este, y las convierte en valores numéricos, los cuales pueden aportar efectivamente al procesamiento de los datos en etapas posteriores del desarrollo del proyecto.

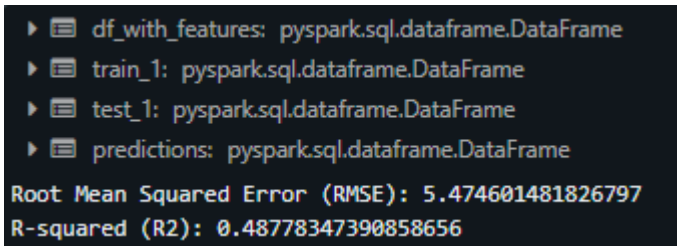
5. Aplicación de técnicas de ML
    a. Random Forest

El uso de Random Forest en análisis predictivo es importante por su capacidad para manejar conjuntos de datos grandes y complejos, así como por su capacidad para manejar características categóricas y numéricas. Al construir múltiples árboles de decisión y promediar sus predicciones, el modelo puede reducir el sobreajuste y mejorar la precisión de las predicciones. Además, Random Forest proporciona información sobre la importancia relativa de cada característica en la predicción, lo que permite identificar qué variables son más influyentes en el resultado.

```
# Crear la columna "features"
assembler = VectorAssembler(inputCols=["AGE_GROUP", "ARREST_BORO", "PERP_RACE", "PERP_SEX", "OFNS_DESC_onehot"], outputCol="features")
df_with_features = assembler.transform(df_encoded)
label = "OFNS_DESC_indexed"
```

Grafica 14 - Variables de Random Forest

Dentro de las características utilizadas en el modelo, la edad (`AGE_GROUP`) y el género del perpetrador (`PERP_SEX`) surgen como dos variables de gran importancia en el análisis delictivo. La edad proporciona información sobre los diferentes grupos demográficos y sus comportamientos delictivos, mientras que el género puede influir en los tipos de delitos cometidos y en las tasas de participación delictiva. Comprender cómo estas variables afectan la predicción del delito es fundamental para desarrollar políticas de seguridad pública efectivas.



Grafica 15 - Métricas finales Random Forest

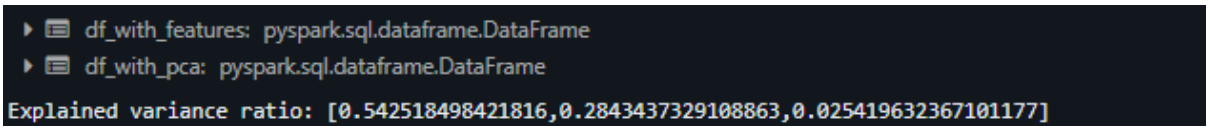
Los resultados del modelo presentes en la Grafica 15 - Métricas finales Random Forest muestran un rendimiento moderado, con un RMSE de alrededor de 5.47 y un coeficiente r^2 de aproximadamente 0.49. Aunque estos resultados indican una capacidad decente para predecir la variable objetivo, hay margen para mejoras. Un modelo como este, sería valioso para identificar tendencias delictivas, asignar recursos de manera efectiva y desarrollar estrategias de prevención del delito más eficaces para la alcaldía de Nueva York.

b. PCA

El Análisis de Componentes Principales (PCA) es una técnica poderosa en el análisis de datos que se utiliza para reducir la dimensionalidad de conjuntos de datos complejos mientras se conserva la mayor cantidad posible de información. En el contexto del análisis delictivo, donde los conjuntos de datos pueden ser grandes y multidimensionales, PCA puede ser especialmente útil para identificar patrones subyacentes y estructuras relevantes.

En cuanto a las variables utilizadas en el PCA, la raza del perpetrador (PERP_RACE) y el distrito de arresto (ARREST_BORO) son dos características fundamentales en el análisis de datos delictivos. La raza del perpetrador puede proporcionar información crucial sobre posibles disparidades en la aplicación de la ley y patrones delictivos dentro de diferentes grupos demográficos. Por otro lado, el distrito de arresto puede revelar

variaciones geográficas en la distribución de delitos y la efectividad de las estrategias de aplicación de la ley en diferentes áreas de la ciudad.



Grafica 16 - Métricas finales PCA

Los resultados del PCA muestran la importancia relativa de cada componente principal, donde el primer componente principal explica aproximadamente el 54.25% de la varianza en los datos, seguido por el segundo componente principal con alrededor del 28.43%, y el tercer componente principal con aproximadamente el 2.54%. Estas proporciones de varianza indican cuánta información se conserva al proyectar los datos en un espacio de menor dimensión.

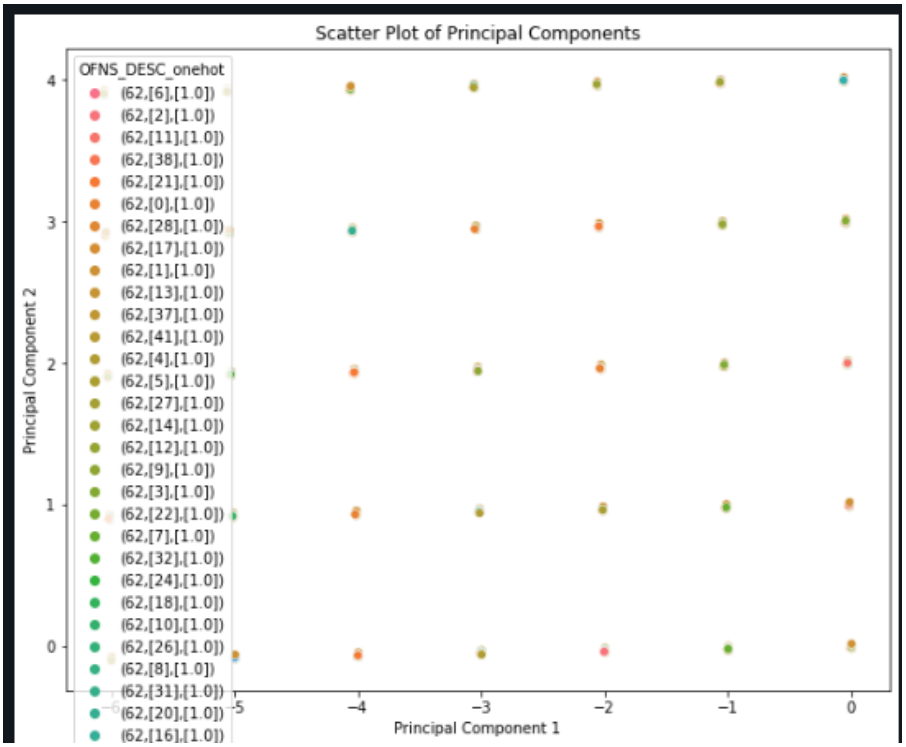
En el contexto del gobierno de Nueva York, el análisis de PCA puede ser valioso para comprender mejor la estructura y los patrones subyacentes en los datos delictivos. La capacidad de identificar relaciones y tendencias entre variables como la raza del perpetrador y el distrito de arresto puede proporcionar información útil para informar políticas de seguridad pública, asignación de recursos y estrategias de prevención del delito. Al comprender mejor estas dinámicas, las autoridades pueden tomar decisiones más informadas y efectivas para abordar las preocupaciones de seguridad y mejorar la calidad de vida en la ciudad.

6. Métricas

Para el Random Forest, obtuvimos un RMSE de alrededor de 5.47 y un coeficiente de determinación R^2 de aproximadamente 0.49. El RMSE nos dice cuánto difieren las predicciones del modelo de los valores observados, en promedio, en la misma escala que los datos originales. Un RMSE más bajo indica un mejor ajuste del modelo a los datos. Por otro lado, el coeficiente de determinación R^2 proporciona una medida de cuánta variabilidad en la variable objetivo es explicada por el modelo. Un valor de R^2 cercano a 1 indica un buen ajuste del modelo a los datos.

Ahora, en el caso del PCA, el "Explained Variance Ratio" nos muestra la proporción de varianza en los datos original que es explicada por cada uno de los componentes principales. En este caso, los tres primeros componentes principales explican el 85.25% de la varianza en los datos originales. Esto significa que al reducir la dimensionalidad de los datos a solo tres dimensiones, conservamos la mayor parte de la información contenida en los datos originales.

La elección de estas métricas es crucial en el contexto del crimen porque nos permiten evaluar la capacidad del modelo para predecir los patrones delictivos y comprender la estructura subyacente de los datos. Un RMSE bajo y un R^2 alto en el caso del Random Forest indican que el modelo puede predecir con precisión la variable objetivo, lo que es esencial para identificar áreas de alto riesgo y asignar recursos de manera efectiva para prevenir el crimen. Por otro lado, el "Explained Variance Ratio" en el PCA nos ayuda a comprender qué tan bien los componentes principales representan la variabilidad en los datos originales, lo que puede ser útil para identificar patrones y relaciones importantes entre las características del crimen.



Grafica 17 – Diagrama de dispersión PCA

El diagrama en la Grafica 17 – Diagrama de dispersión PCA corresponde a un diagrama de dispersión de componentes principales, resultado del PCA, es una herramienta valiosa

para analizar conjuntos de datos complejos, como los que se generan en la ciudad de Nueva York. Al reducir la dimensionalidad de los datos a través del análisis de componentes principales (PCA), es posible visualizar patrones y agrupaciones que de otro modo serían difíciles de detectar.

En una ciudad tan densa y diversa como Nueva York, comprender estas estructuras subyacentes en los datos es crucial para tomar decisiones informadas en áreas como la asignación de recursos, la planificación urbana, la aplicación de la ley y la prestación de servicios públicos.

Por ejemplo, este tipo de análisis podría ayudar a identificar patrones de delincuencia, segregación residencial o disparidades en el acceso a servicios, lo que permitiría a las autoridades abordar estos desafíos de manera más efectiva. En resumen, las métricas de reducción de dimensionalidad como el PCA son herramientas poderosas para extraer conocimientos accionables de los vastos conjuntos de datos que genera una metrópoli como Nueva York.

En conjunto, estas métricas nos proporcionan una evaluación integral del rendimiento del modelo y su relevancia para abordar los desafíos en el contexto del crimen.

7. Bono

Para el desarrollo del bono se pide en el enunciado desarrollar una red neuronal con base a los datos que se vienen trabajando a lo largo de todo el proyecto. En este sentido, debido a que todos los modelos de ML aplicados en el punto 5 están hechos en relación con el conjunto de datos de arrestos en Nueva York, es justo que este modelo de *deep learning* sea hecho relacionado a el conjunto de datos de pobreza, para así tener una cobertura más integral de todos los datos.

Teniendo esto en consideración, el primer paso fue escoger las bibliotecas a usar para desarrollar el modelo. Pyspark con su framework de ML no provee directamente una biblioteca para trabajar con redes neuronales, por lo cual para desarrollar la actividad con esta tecnología fue necesario usar otro modelo más pequeño que representa unidades más pequeñas del modelo de redes neuronales, estos son los perceptrones. Por

lo que para el desarrollo se utilizó perceptrones multicapa que en teoría son lo mismo que una red neuronal convencional, pero más simple.

Ahora, la pregunta a modelar es principalmente un problema de regresión, donde teniendo en consideración las evidencias en el conjunto de datos, información sobre edad, ingresos antes de impuestos, tipo de vivienda, etc; se pueda calcular de forma fidedigna la categoría o tipo de educación que se recibió.

En este problema de regresión, la columna que representa el nivel de educación hablara de un grupo para centrarse en el valor continuo. Esto debido a que a nivel computacional categorizar entre 24 categorías es demasiado exigente para la máquina que entrega DataBricks. Posteriormente se hizo el preprocesamiento correspondiente, normalizar los datos con la biblioteca de Pyspark y ensamblar los vectores de características. Después se procedió a crear una arquitectura basada en la experiencia del grupo, en este caso la arquitectura planteada fue de 64 y 32 neuronas en la capa oculta, con un entrenamiento de 100 epochs y un optimizador Adam.

Se intento de diferentes formas ejecutar el modelo en el entorno de DataBricks sin éxito. Esto debido a que la versión community no tiene los recursos para poder soportar modelos de este nivel de exigencia, a pesar de que ya se hubiera optimizado el problema. Por lo que la solución mejor encontrada fue cambiar de proveedor de entorno y ejecutar el modelo con los siguientes resultados:

Métrica	Valor
Precisión	0.33
Recall	0.25
F1-Score	0.27
MSE	14.52
RMSE	3.81
R^2	0.61

Tabla 1 - Métricas de la red neuronal

Con base en estas métricas podemos observar que, si podemos calcular la educación con cierto nivel de confianza, obviamente con un error asociado, pero este es muy bajo (considerando las métricas de regresión y no las del modelo).

8. Referencias

[1] "Poverty Data", *Nyc.gov*. [En línea]. Disponible en: https://www.nyc.gov/site/opportunity/poverty-in-nyc/poverty-data.page. [Consultado: 26-may-2024].

[2] "City of New York - NYPD arrest data (year to date)", *Data.gov*. [En línea]. Disponible en: https://catalog.data.gov/dataset/nypd-arrest-data-year-to-date. [Consultado: 26-may-2024].

[3] Police Department (NYPD), "NYPD arrest data (year to date)". 05-jun-2018.

[4] A. R. Morales, "The state of poverty and disadvantage in New York City", *Columbia University Center on Poverty and Social Policy*, 21-feb-2024. [En línea]. Disponible en: https://www.povertycenter.columbia.edu/nyc-poverty-tracker/the-state-of-poverty-and-disadvantage-in-2022. [Consultado: 26-may-2024].

[5] S. Chen, "Poverty has soared in New York, with children bearing the brunt", *The New York times*, The New York Times, 21-feb-2024.

[6] *Nyc.gov*. [En línea]. Disponible en: https://www.nyc.gov/assets/opportunity/pdf/20_poverty_measure_report.pdf. [Consultado: 26-may-2024].