Proyecto

-

Entrega Final Proyecto

Documentación completa

Machine Learning

Grupo 4

Ismael Alcaide

Jaime Revilla

Enrique Puente

INTRODUCCIÓN DEL PROYECTO

El objetivo principal de este proyecto es desarrollar un modelo de aprendizaje automático que prediga el tipo de delito más probable de ocurrencia de un crimen violento en un área específica o en un momento del día. Para ello, se **utilizará como variable objetivo la columna Primary Type,** que clasifica los delitos según su naturaleza. Basándose en datos históricos de crímenes reportados en Chicago, el modelo se entrenará con factores como la hora del día, la ubicación y el tipo de delito previo. Este análisis permitirá optimizar la asignación de recursos policiales y diseñar estrategias más efectivas de prevención del crimen.

CARGA DE LOS DATOS

El primer paso en nuestro análisis exploratorio fue la carga de los datos. Para ello, comenzamos importando las librerías necesarias para la manipulación y visualización del dataset. Utilizamos bibliotecas como pandas para el manejo de datos, numpy para cálculos numéricos y matplotlib y seaborn para las visualizaciones.

A continuación, cargamos el dataset en un DataFrame de pandas e imprimimos los primeros registros para obtener una visión general de su estructura y contenido. Esto nos permitió identificar las principales características del conjunto de datos y familiarizarnos con el formato de las variables incluidas.

Además, calculamos el tamaño total del dataset, identificando tanto el número de filas como de columnas, lo cual es crucial para planificar el análisis posterior. También procedimos a revisar la descripción de cada variable presente en el dataset, que incluye:

- **ID**: Identificador único de cada crimen.
- **Case Number**: Número de casos asignados por las autoridades para identificar el incidente.
- **Date**: Fecha y hora exacta en la que ocurrió el crimen.
- **Block**: Dirección aproximada donde tuvo lugar el crimen, presentada en forma de bloque.
- **IUCR**: Código estándar utilizado para clasificar el tipo de crimen, según las normativas policiales de Chicago.
- **Primary Type**: Tipo principal del crimen, categorizado en términos generales (por ejemplo, asalto, robo).
- **Description**: Descripción detallada del crimen, que proporciona información adicional sobre el incidente.
- **Location Description**: Lugar donde ocurrió el crimen (por ejemplo, calle, apartamento, residencia).
- **Arrest**: Indica si hubo un arresto relacionado con el crimen (True/False).
- **Domestic**: Indica si el incidente está relacionado con violencia doméstica (True/False).
- **Beat**: Código que identifica la patrulla policial asignada al área donde ocurrió el crimen.
- **District**: Distrito policial al que pertenece el área donde ocurrió el crimen.
- **Ward**: División política de la ciudad a la que pertenece el lugar del crimen.
- **Community Area**: Número de identificación que corresponde al área comunitaria donde ocurrió el crimen.

- **FBI Code**: Código asignado por el FBI para clasificar el tipo de crimen dentro de categorías generales.
- **X Coordinate** y **Y Coordinate**: Coordenadas en el sistema de referencia de Chicago, útiles para la ubicación geográfica.
- **Year**: Año en el que ocurrió el crimen.
- **Updated On**: Fecha y hora en la que se actualizó el registro del crimen en la base de datos.
- **Latitude** y **Longitude**: Coordenadas geográficas del lugar donde ocurrió el crimen.
- **Location**: Representación combinada de la latitud y longitud en formato de texto (por ejemplo, "(41.881, -87.623)").

La siguiente tabla muestra una vista previa del conjunto de datos con varias columnas relevantes para el análisis de crímenes.

	Unnamed: 0	ID	Case Number	Date	Block	IUCR	Primary Type	Description	Location Description	Arrest	...	Ward	Community Area	FBI Code	X Coordinate	Y Coordinate	Year	Updated On	Latitude	Longitude	Location
0	3	10508693	HZ250496	05/03/2016 11:40:00 PM	013XX S SAWYER AVE	0486	BATTERY	DOMESTIC BATTERY SIMPLE	APARTMENT	True	...	24.0	29.0	08B	1154907.0	1893681.0	2016	05/10/2016 03:56:50 PM	41.864073	-87.706819	(41.864073157, -87.706818608)
1	89	10508695	HZ250409	05/03/2016 09:40:00 PM	061XX S DREXEL AVE	0486	BATTERY	DOMESTIC BATTERY SIMPLE	RESIDENCE	False	...	20.0	42.0	08B	1183066.0	1864330.0	2016	05/10/2016 03:56:50 PM	41.782922	-87.604363	(41.782921527, -87.60436317)
2	197	10508697	HZ250503	05/03/2016 11:31:00 PM	053XX W CHICAGO AVE	0470	PUBLIC PEACE VIOLATION	RECKLESS CONDUCT	STREET	False	...	37.0	25.0	24	1140789.0	1904819.0	2016	05/10/2016 03:56:50 PM	41.894908	-87.758372	(41.894908283, -87.758371958)
3	673	10508698	HZ250424	05/03/2016 10:10:00 PM	049XX W FULTON ST	0460	BATTERY	SIMPLE	SIDEWALK	False	...	28.0	25.0	08B	1143223.0	1901475.0	2016	05/10/2016 03:56:50 PM	41.885687	-87.749516	(41.885686845, -87.749516983)
4	911	10508699	HZ250455	05/03/2016 10:00:00 PM	003XX N LOTUS AVE	0820	THEFT	$500 AND UNDER	RESIDENCE	False	...	28.0	25.0	06	1139890.0	1901675.0	2016	05/10/2016 03:56:50 PM	41.886297	-87.761751	(41.886297242, -87.761750709)

Imagen 1: Visualización de los datos cargados

El conjunto de datos contiene **1,456,714 registros** y **23 columnas**, lo que confirma la presencia de un volumen grande de datos. Este tamaño es adecuado para desarrollar modelos de predicción y análisis exploratorio.

```
(1456714, 23)
```

Imagen 2: Visualización del tamaños del dataset

Este análisis inicial nos permitió entender la estructura del dataset y las variables que potencialmente serán relevantes para predecir la probabilidad de ocurrencia de un crimen. También identificamos que algunas variables, como las coordenadas geográficas, podrían requerir transformación o limpieza adicional para ser usadas.

ANÁLISIS EXPLORATORIO DE DATOS (EDA)

El análisis exploratorio de datos (EDA) nos sirvió para comprender la distribución, las características y las relaciones entre las variables en nuestro dataset, antes de realizar cualquier preprocesamiento. A continuación, se detallan los pasos y conclusiones principales:

Resumen estadístico de las columnas numéricas

Realizamos un análisis estadístico básico de las variables numéricas del dataset, lo que nos permitió identificar posibles valores atípicos y patrones en los datos. Observamos los rangos significativos de las siguientes variables:

- **Coordenadas geográficas (X Coordinate, Y Coordinate, Latitude, Longitude):** Confirmaron que los datos están limitados a un área específica, dentro de Chicago.
- **Años (Year):** Los registros abarcan un período de tiempo consistente, lo que asegura datos históricos suficientes para nuestro modelo. Desde 2012 hasta 2017.
- **Áreas comunitarias (Community Area):** Identificamos un rango completo de áreas comunitarias, lo que indica una cobertura representativa de la ciudad.

Este análisis también nos ayudó a identificar posibles valores extremos que podrían requerir limpieza o tratamiento posterior.

Análisis de datos únicos y duplicados

Examinamos la calidad general del dataset mediante la detección de valores únicos y duplicados:

- **Datos únicos:** Se confirmó que variables clave como ID tienen valores únicos, lo que asegura la individualidad de cada registro.
- **Datos duplicados:** Detectamos registros duplicados asociados a ciertos números de caso (Case Number), lo que podría deberse a errores administrativos.

```
Unnamed: 0: 1456714 valores únicos
ID: 1456714 valores únicos
Case Number: 1456598 valores únicos
Date: 582146 valores únicos
Block: 32774 valores únicos
IUCR: 365 valores únicos
Primary Type: 33 valores únicos
Description: 342 valores únicos
Location Description: 142 valores únicos
Arrest: 2 valores únicos
Domestic: 2 valores únicos
Beat: 302 valores únicos
District: 24 valores únicos
Ward: 50 valores únicos
Community Area: 78 valores únicos
FBI Code: 26 valores únicos
X Coordinate: 67714 valores únicos
Y Coordinate: 111555 valores únicos
Year: 6 valores únicos
Updated On: 959 valores únicos
Latitude: 368076 valores únicos
Longitude: 367942 valores únicos
Location: 368286 valores únicos
```

Imagen 3: Visualización de las variables únicas de nuestro dataset

Distribución y frecuencia de crímenes

La distribución y frecuencia de los diferentes tipos de crímenes arrojaron interesantes hallazgos:

- **Crímenes más comunes:** El robo y el daño criminal son los delitos más frecuentes, mientras que el homicidio tiene una incidencia considerablemente menor.

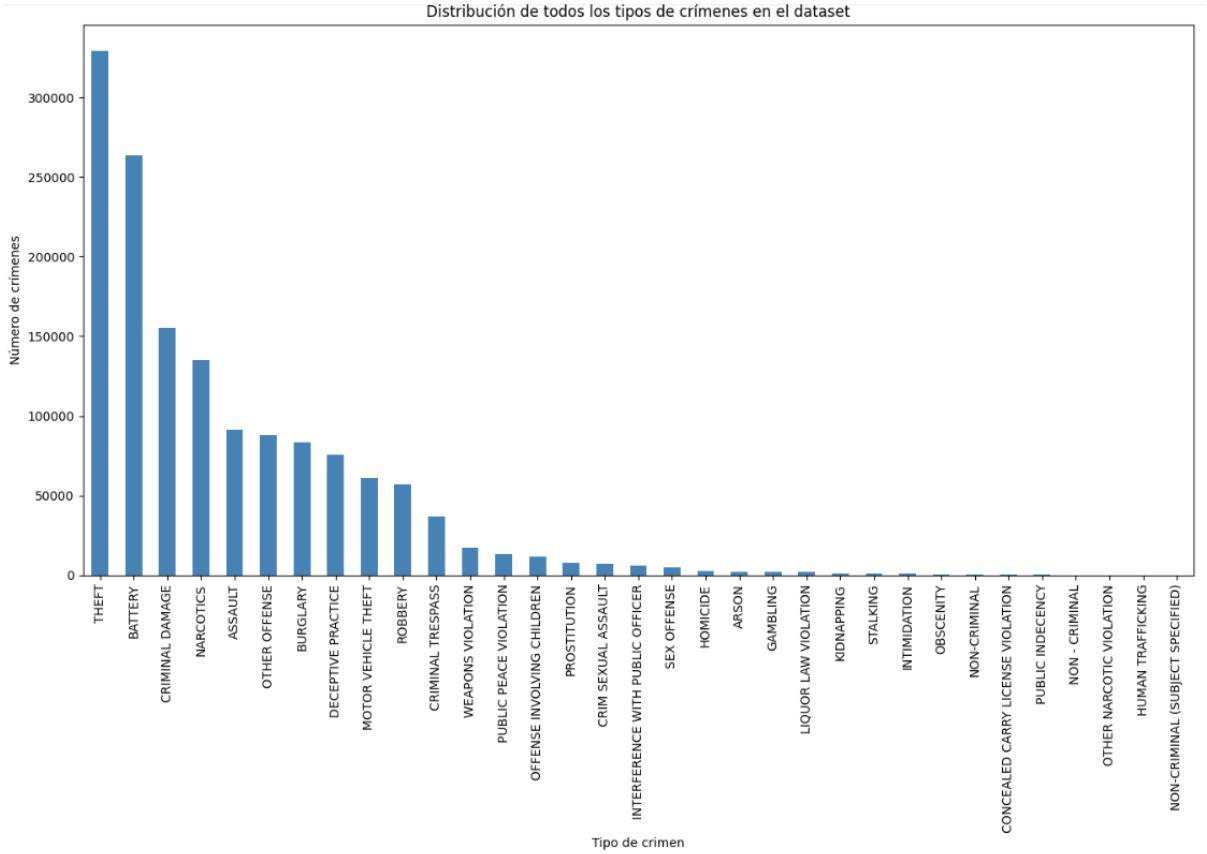

Imagen 4: Visualización de la distribución de todos los tipos de crímenes en el dataset

- **Frecuencia de arrestos:** Observamos que ciertos tipos de crímenes tienen tasas de arresto más altas, lo que podría ser un factor influyente en la probabilidad de arresto.

Imagen 5: Visualización de la frecuencia de arrestos por tipo de crimen

Análisis de variables categóricas

Se realizó un desglose por categorías para comprender mejor las características del dataset:

- **Crímenes domésticos vs. no domésticos:** Algunos tipos de crímenes están significativamente relacionados con ambientes domésticos, lo que podría ser relevante para el modelo predictivo.

Imagen 6: Visualización de la distribución de crímenes domésticos y no domésticos por tipo de crimen

- **Tipo de crimen:** Identificamos tendencias interesantes en los tipos de crímenes más comunes y su relación con otras variables.

Análisis temporal

Estudiamos las tendencias anuales y mensuales en los datos:

- **Tendencias anuales:** Identificamos fluctuaciones en la incidencia de crímenes a lo largo de los años, lo que podría estar relacionado con cambios en políticas de seguridad o factores sociales.

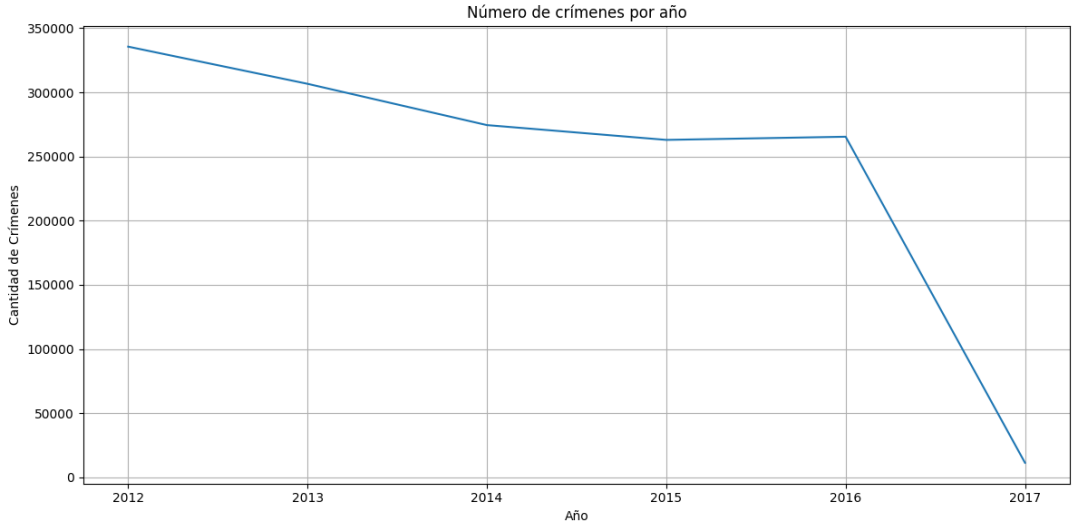

Imagen 7: Visualización del número de crímenes por año

- **Tendencias mensuales:** Algunos meses muestran un aumento en la actividad criminal, posiblemente debido a factores estacionales o eventos específicos.

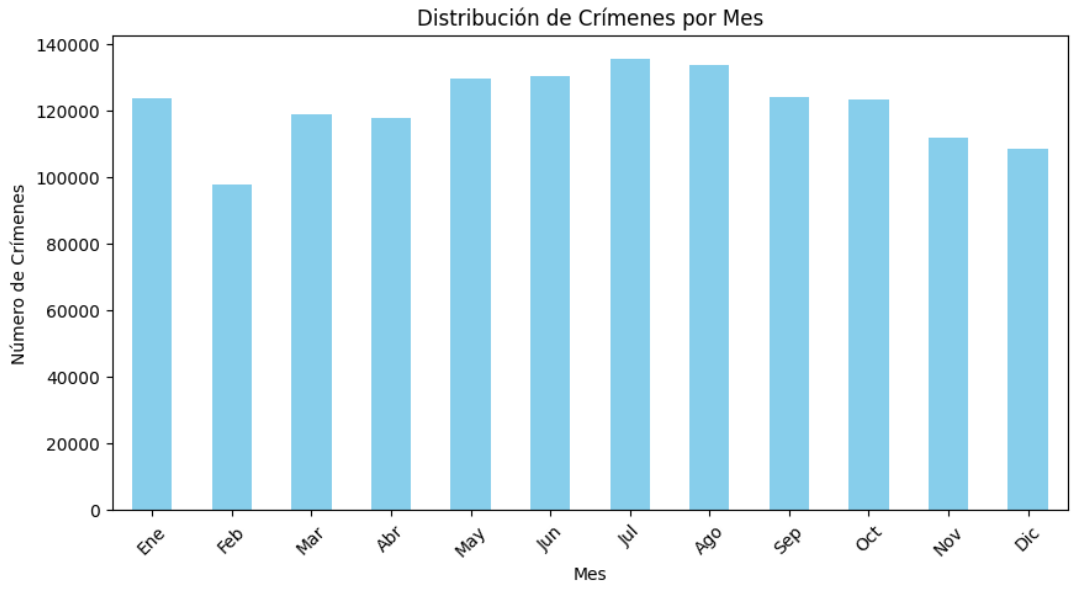

Imagen 8: Visualización de la distribución de crímenes por mes

Distribución geográfica

Analizamos las variables de latitud y longitud para entender la cobertura del dataset:

- **Patrones espaciales:** Confirmamos que los crímenes se distribuyen dentro de la ciudad de Chicago, lo que será esencial para modelar características espaciales y realizar predicciones geográficas.

Relaciones entre variables

Generamos una matriz de correlación para examinar la relación entre las variables numéricas:

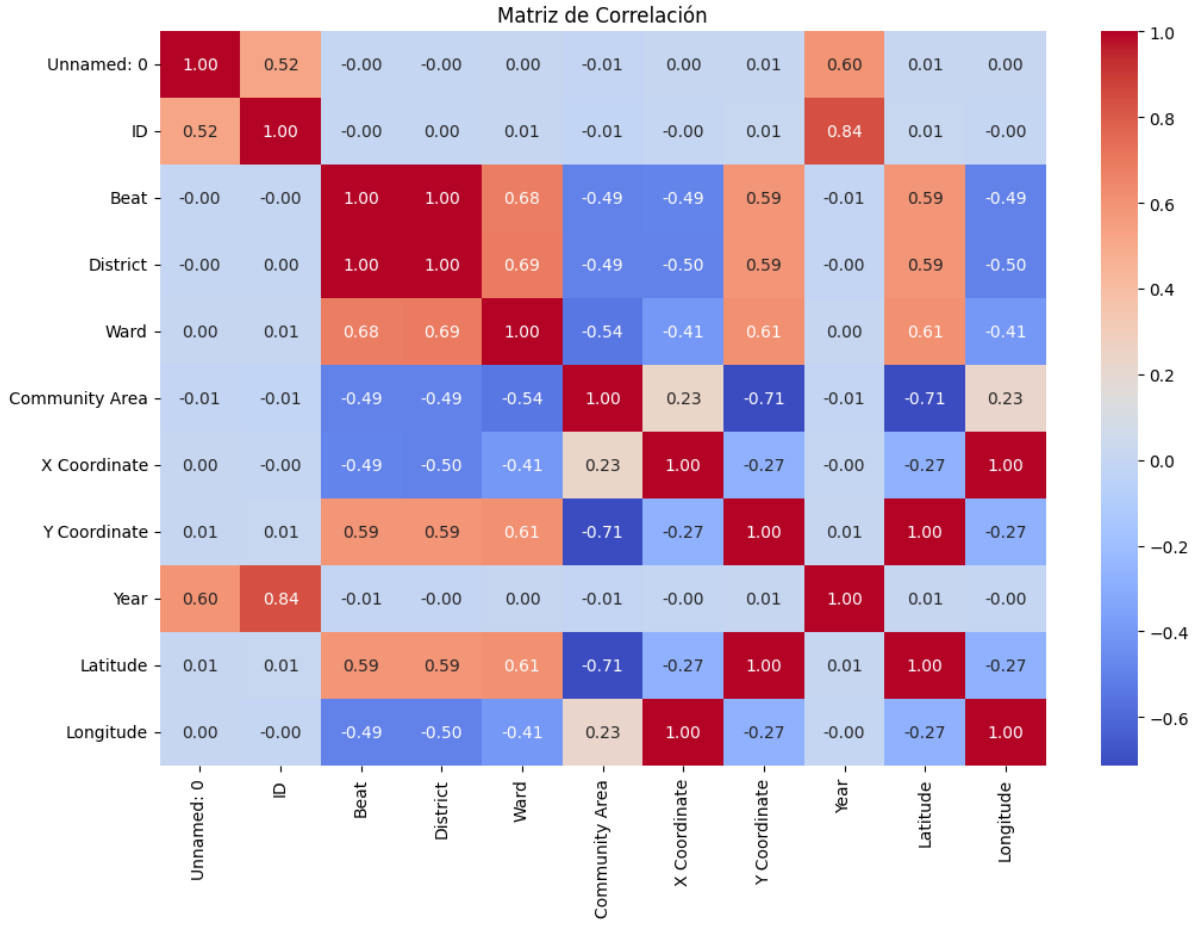

Imagen 9: Visualización de la matriz de correlación sin pre procesar los datos

- **Principales hallazgos:** Identificamos una relación entre las coordenadas geográficas y las áreas comunitarias, lo cual es consistente con la estructura espacial del dataset.
- **Variables no correlacionadas:** No se observaron correlaciones muy altas entre otras variables numéricas, lo que sugiere que la predicción dependerá de una combinación compleja de factores en lugar de una relación lineal sencilla.

Estos hallazgos del EDA nos proporcionan una base sólida para el preprocesamiento de los datos y el desarrollo de un modelo predictivo que capture patrones espaciales, temporales y categóricos en los datos.

Conclusiones del Análisis Exploratorio de Datos (EDA)

- **Resumen estadístico y análisis de las variables numéricas**, el análisis estadístico básico nos ha permitido identificar posibles valores atípicos y patrones en los datos, como rangos significativos en las coordenadas geográficas, años y áreas comunitarias.
- **Datos únicos y duplicados**, el análisis de valores únicos y duplicados confirma la calidad general del dataset, aunque detectamos registros duplicados asociados a ciertos números de caso, probablemente debido a errores administrativos.
- **Distribución y frecuencia de crímenes:**
    - **La distribución** de los tipos de crímenes nos mostrua que delitos como el robo y el daño criminal son los más comunes, mientras que otros como el homicidio tienen una incidencia menor.
    - **La frecuencia** de arrestos varía significativamente según el tipo de crimen, lo cual podría influir en la predicción de la probabilidad de arresto en futuros análisis.
- **Variables categóricas**, desglose por categorías (crímenes domésticos/no domésticos, tipo de crimen, etc.) nos ha permitido entender mejor la composición del dataset, destacando que ciertos delitos están más relacionados con ambientes domésticos, lo que podría ser relevante para el modelo predictivo.
- **Análisis temporal**, identificamos las tendencias anuales y mensuales en los datos. Algunos meses presentan una mayor incidencia de crímenes, lo que podría estar relacionado con factores estacionales o eventos específicos.
- **Distribución geográfica,** las variables de latitud y longitud confirmamos la cobertura del dataset en un área específica. Esta información es esencial para modelar las características espaciales de los crímenes.
- **Relaciones entre las variables,** la matriz de correlación destacamos la relación entre variables como las coordenadas geográficas y las áreas comunitarias. Sin embargo, no observamos correlaciones muy altas entre las variables numéricas, lo que sugiere que la predicción podría depender de una combinación compleja de factores.

PREPROCESAMIENTO

El preprocesamiento fue una etapa crucial para garantizar que los datos estuvieran listos para el análisis y modelado. Este proceso incluyó varias tareas enfocadas en tratar valores nulos, optimizar tipos de datos, eliminar redundancias y transformar las variables para un uso más eficiente. A continuación, se detallan los pasos realizados:

Tratamiento de valores nulos

```
Valores nulos por columna antes del procesamiento:
Date                        0
Block                       0
Primary Type                0
Description                 0
Location Description     1658
Arrest                      0
Domestic                    0
District                    1
Community Area             40
Latitude                37083
Longitude               37083
Location                37083
dtype: int64
```

Imagen 10: Visualización de los valores nulos

Se abordaron los valores nulos presentes en el dataset utilizando diferentes métodos según el tipo de variable:

- **Variables categóricas:** Los valores nulos fueron rellenados con la **moda** de cada columna, dado que representaba la categoría más frecuente y era menos propensa a introducir sesgos.
- **Variables numéricas:** Se optó por rellenar los valores nulos utilizando la **mediana**, ya que es más robusta frente a valores extremos. Por ejemplo:
    - Las variables **Latitude** y **Longitude** fueron completadas con su mediana para preservar la coherencia geográfica y evitar distorsiones en los análisis espaciales.

Evitar el uso de la media en las variables numéricas fue una decisión clave para mantener la precisión en las ubicaciones y otros atributos sensibles a la variabilidad.

Conversión de tipos de datos

Para optimizar el manejo de las variables:

- Las columnas numéricas se convirtieron a tipos **int** o **float**, asegurando un procesamiento adecuado en las etapas posteriores.
- Las variables categóricas se transformaron al tipo **category**, lo que redujo el uso de memoria y facilitó la aplicación de codificaciones específicas.

```
Tipos de datos después de la conversión:
Date                      datetime64[ns]
Block                           category
Primary Type                    category
Description                     category
Location Description            category
District                           int64
Community Area                     int64
Latitude                           int64
Longitude                          int64
Location                        category
Arrest_0                           int64
Arrest_1                           int64
Domestic_0                         int64
Domestic_1                         int64
dtype: object
```

Imagen 11: Visualización de los tipos de datos

Codificación de variables categóricas

Se aplicó **one-hot encoding** (creación de dummies) a las variables categóricas con valores booleanos:

- **Arrest** y **Domestic** fueron transformadas en columnas binarias (0 y 1), lo que permitió representar estas variables de forma compatible con los algoritmos de machine learning.

Eliminación de columnas irrelevantes

Se eliminaron columnas que no aportan valor significativo al objetivo del proyecto:

- **Unnamed: 0, ID, Case Number, IUCR, Beat, Ward, FBI Code, X Coordinates, Y Coordinates, Updated On y Year**.
- Estas columnas contenían identificadores únicos, códigos redundantes o información duplicada que no contribuyen a la predicción de crímenes violentos.

Conversión y estandarización de fechas

La columna **Date** fue convertida al formato de fecha y desglosada en componentes como:

- **Año, mes, día de la semana y hora**, lo que permitió realizar un análisis temporal más detallado e identificar tendencias relevantes.

```
Columnas convertidas a formato datetime:
                 Date
0 2016-05-03 23:40:00
1 2016-05-03 21:40:00
2 2016-05-03 23:31:00
3 2016-05-03 22:10:00
4 2016-05-03 22:00:00
```

Imagen 12: Visualización de la conversión de la columna Date

Normalización de columnas

Las variables seleccionadas fueron normalizadas para garantizar que todas tuvieran una escala consistente, mejorando así el rendimiento del modelo. Esto fue especialmente relevante para variables como:

- **Latitude, Longitude y Community Area**, donde las diferencias de escala podrían haber afectado negativamente el análisis.

Simplificación de la columna Primary Type

Simplificamos las variables de Primary Type para no tener tantos valores.

```
Primary Type Grouped
PROPERTY CRIME     668579
VIOLENT CRIME      440674
DRUG OFFENSE       135270
OTHER OFFENSE      119238
FRAUD               75495
WEAPONS OFFENSE     17323
NON-CRIMINAL           97
Name: count, dtype: int64
```

Imagen 13: Visualización de la simplificación de la columna Primary Type

Decisiones justificadas

Cada paso del preprocesamiento fue cuidadosamente diseñado en función de los hallazgos del análisis exploratorio de los datos. Las decisiones tomadas se enfocaron en:

1. Eliminar redundancias y asegurar la relevancia de las variables seleccionadas.
2. Optimizar el almacenamiento y el manejo de los datos.
3. Preparar un dataset limpio y estructurado, adecuado para los algoritmos de machine learning.

Este enfoque garantizó que los datos fueran consistentes, representativos y listos para ser usados en el desarrollo de un modelo predictivo eficiente.

EXTRACCIÓN DE KPIS

La extracción de KPIs (Indicadores Clave de Rendimiento) permitió identificar patrones relevantes y tendencias clave en los datos tras el preprocesamiento. A continuación, se detallan los principales hallazgos:

Análisis y visualización de variables

- **Frecuencia de crímenes en julio de 2012:** Los delitos más comunes fueron:
    - **Property Crime**, **Violent Crime** con una alta concentración en calles y aceras.

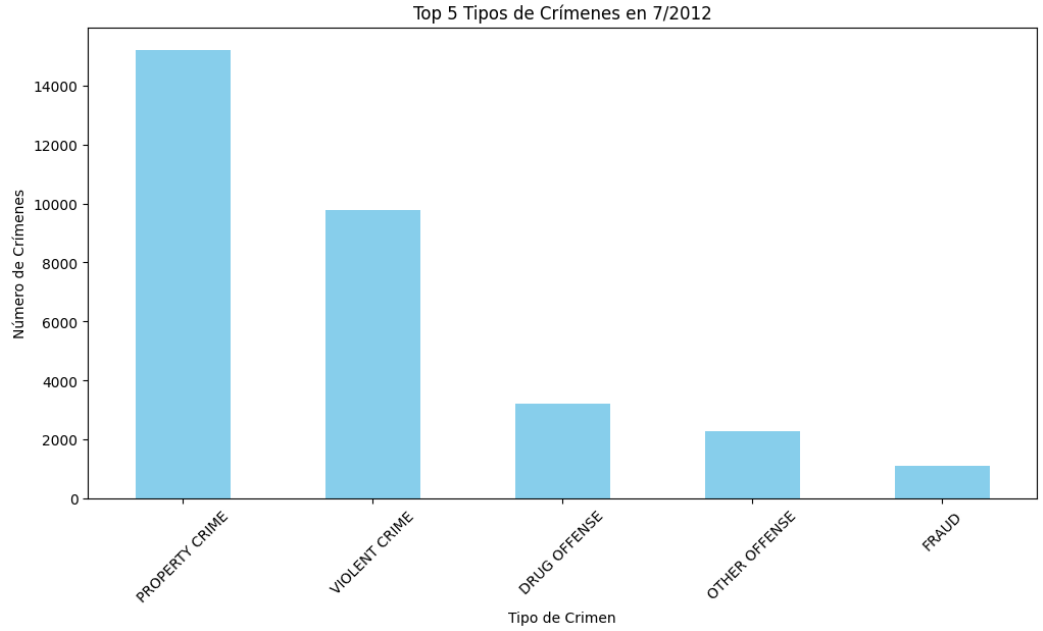

Imagen 14: Visualización de los crímenes más frecuentes en julio de 2012

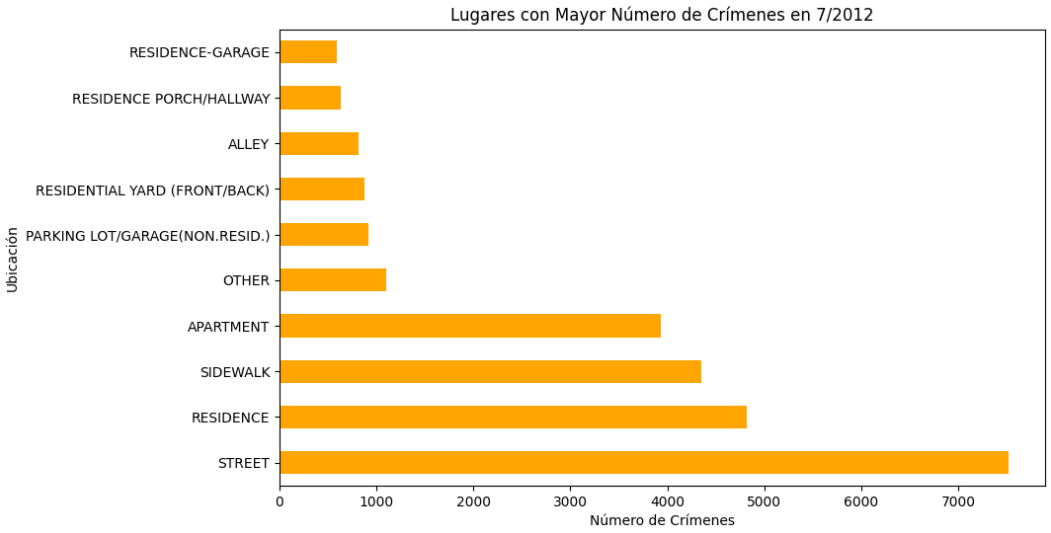

Imagen 15: Visualización de los lugares con mayor número de crimenes en julio de 2012

- Los **domestic crimes** representaron una minoría frente a los **non-domestic crimes**, mostrando diferencias claras en los patrones de ocurrencia.

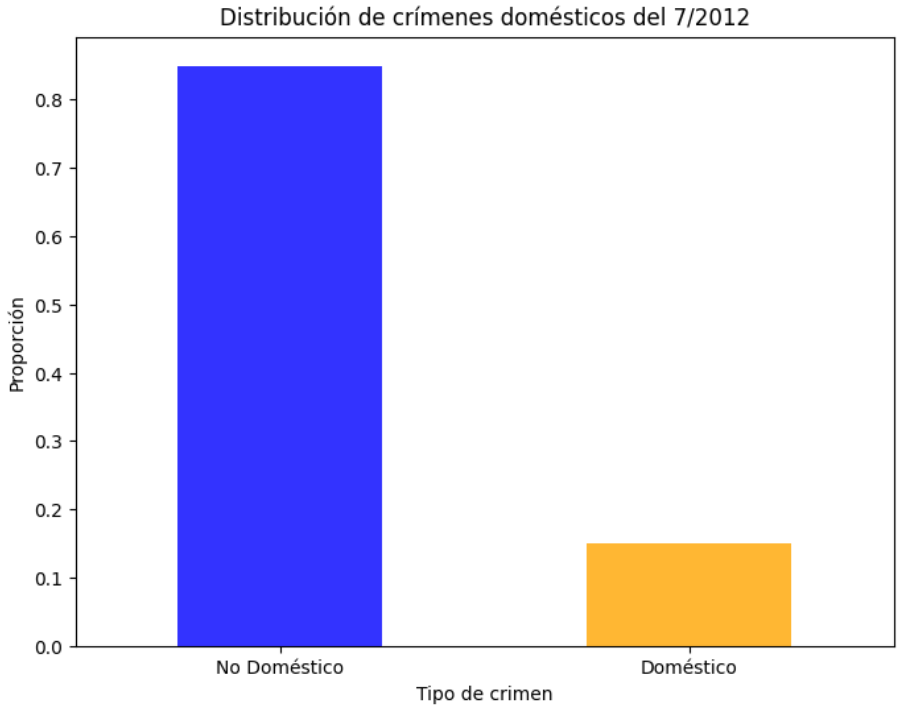

Imagen 16: Visualización de los crímenes domésticos en julio de 2012

- Un mapa de calor confirmó las "zonas calientes" con mayor incidencia de delitos.

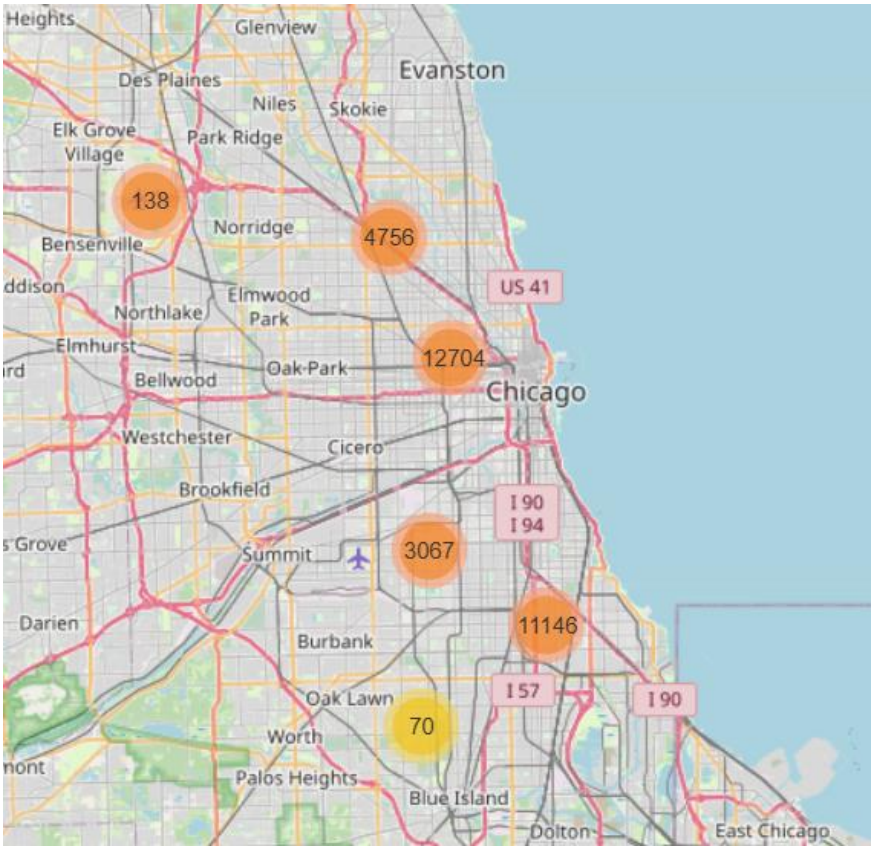

Imagen 17: Visualización del mapa mostrando las "zonas calientes"

- **Horarios con mayor incidencia de crímenes:**

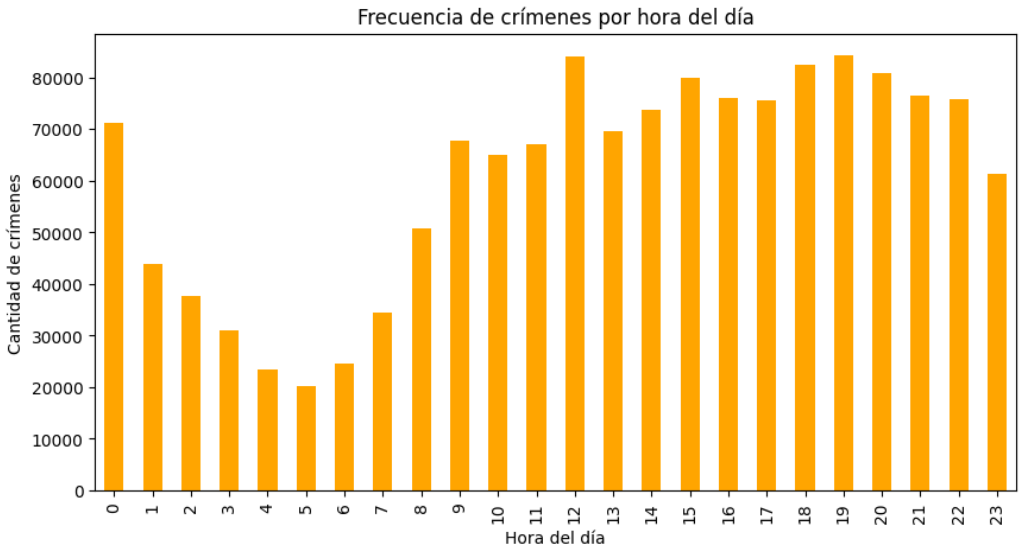

Imagen 18: Visualización de la incidencia de crímenes por hora

Las horas entre **15:00 y 21:00** concentraron la mayoría de los crímenes, con un pico significativo en las tardes y primeras horas de la noche.

- **Tipos de crímenes más comunes en áreas residenciales:**

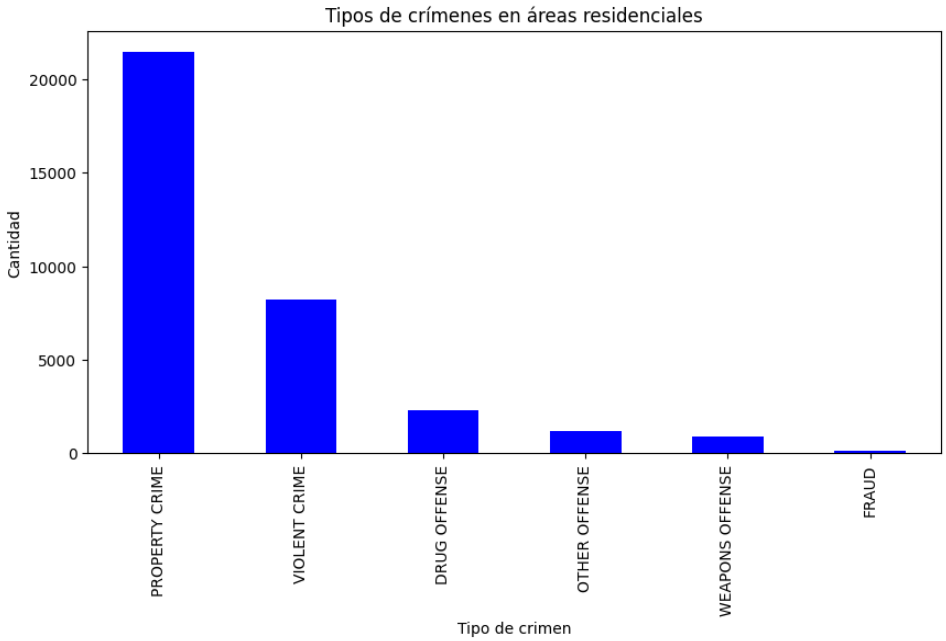

Imagen 19: Visualización de los crímenes más comunes en áreas residenciales

En áreas residenciales predominaron **property crime y violent crime** destacando la importancia de estas zonas como focos de delitos específicos.

- **Relación entre arrestos y tipos de crímenes:**

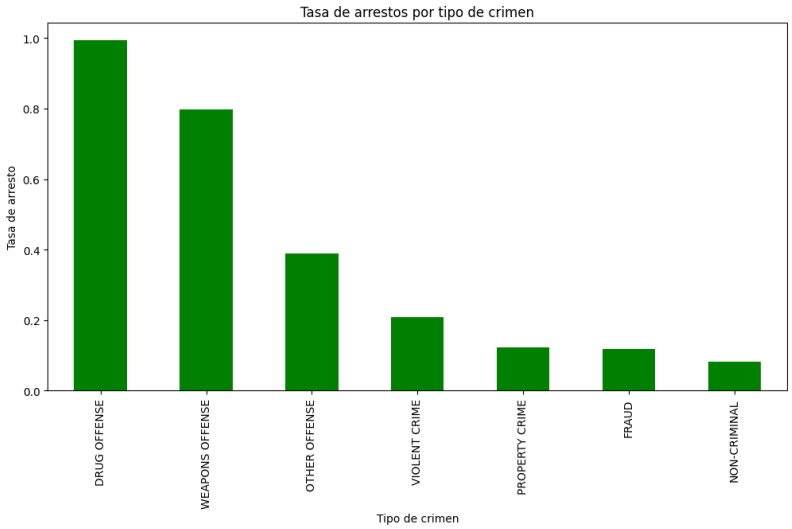

Imagen 20: Visualización entre los arrestos y los tipos de crímenes

Delitos como **drug offense y weapons offense** presentaron las tasas de arresto más altas. En contraste, los crímenes como **fraud y property crime** tuvieron tasas significativamente más bajas de arresto.

- **Incidencia de crímenes domésticos:**
  - Los **domestic crimes** representaron una proporción menor que los **non-domestic crimes**. Solo un **19.40%** de los crímenes domésticos terminaron en arresto, resaltando una tasa de resolución relativamente baja.

Análisis de crímenes en días festivos

- **Thanksgiving:**

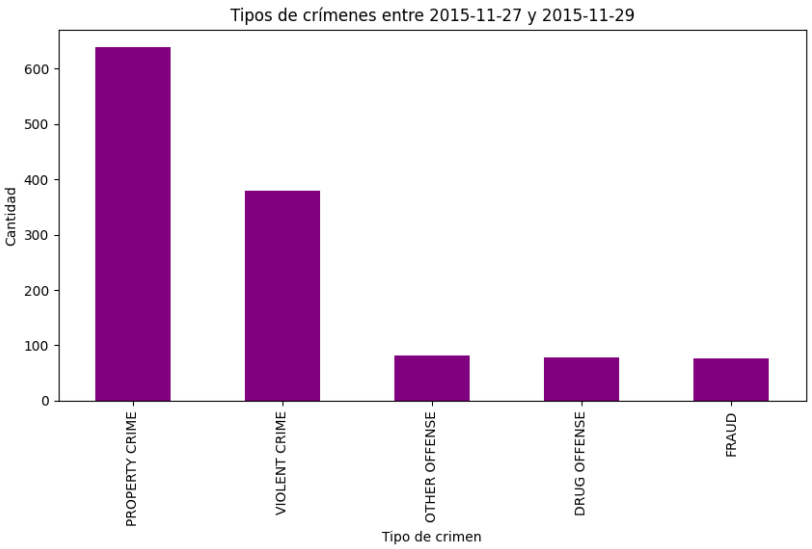

Imagen 21: Visualización de los crímenes en Thanksgiving

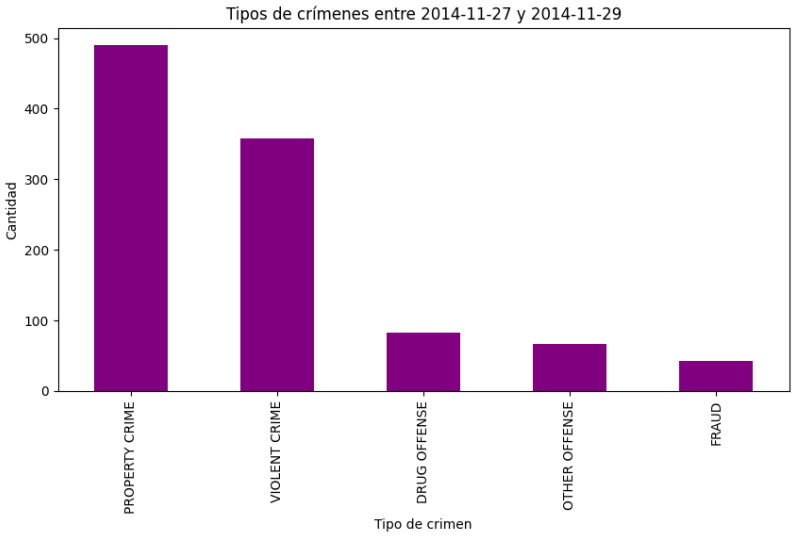

Imagen 22: Visualización de los crímenes en Thanksgiving

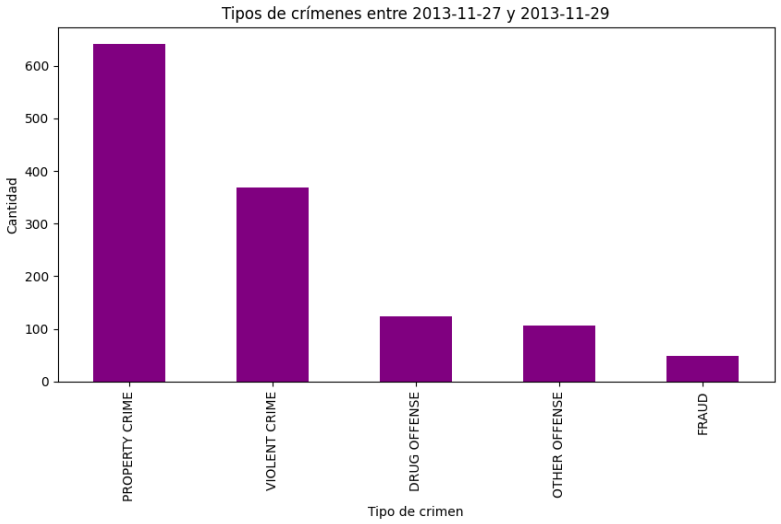

Imagen 23: Visualización de los crímenes en Thanksgiving

Los delitos más frecuentes fueron **property crime y violent crime**, con un leve aumento en incidentes, especialmente en espacios públicos.

- **Independence Day:**

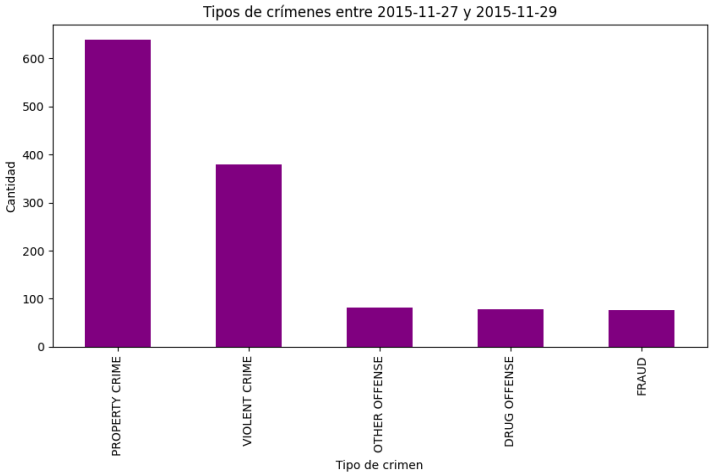

Imagen 24: Visualización de los crímenes en Independence day

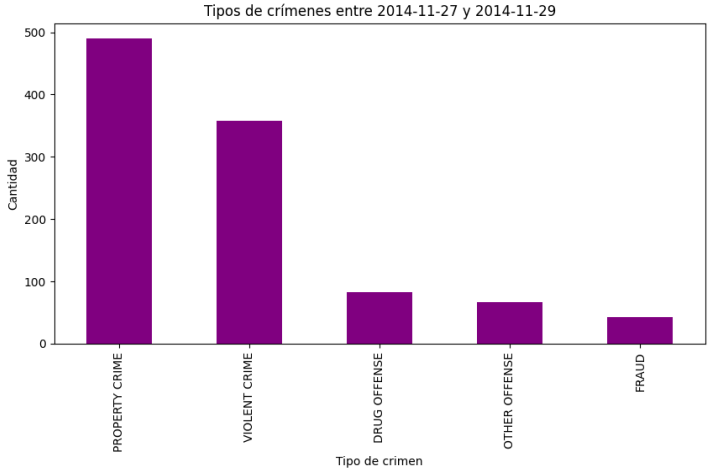

Imagen 25: Visualización de los crímenes en Independence day

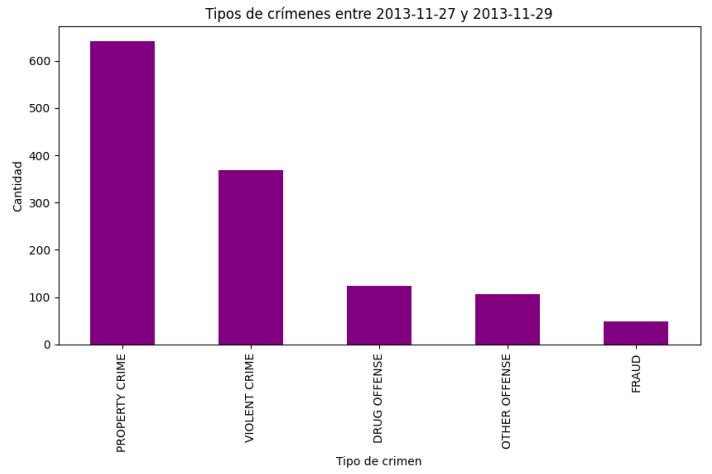

Imagen 26: Visualización de los crímenes en Independence day

Predominaron **property crime y violent crime**, con un aumento notable en delitos relacionados con **drug offense** durante las celebraciones.

- **New Year's Eve:**

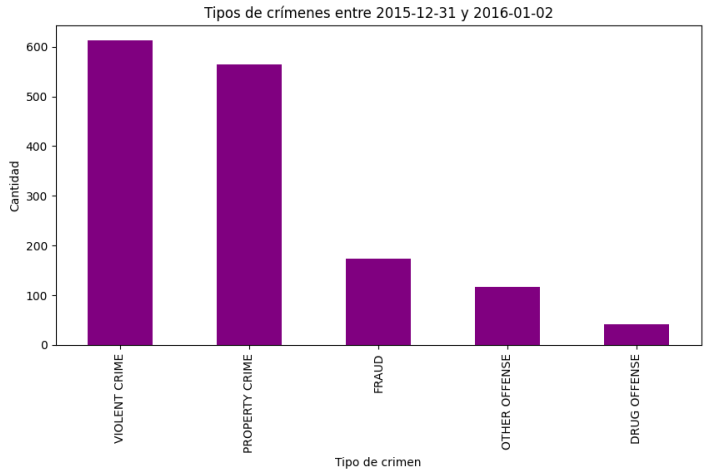

Imagen 27: Visualización de los crímenes en New Year´s Eve

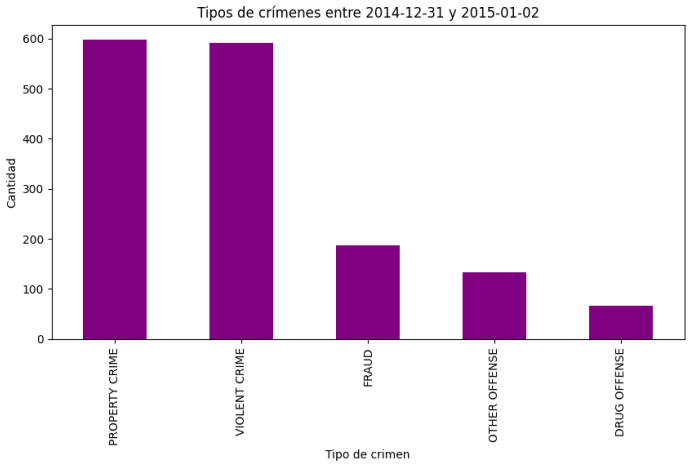

Imagen 28: Visualización de los crímenes en New Year´s Eve

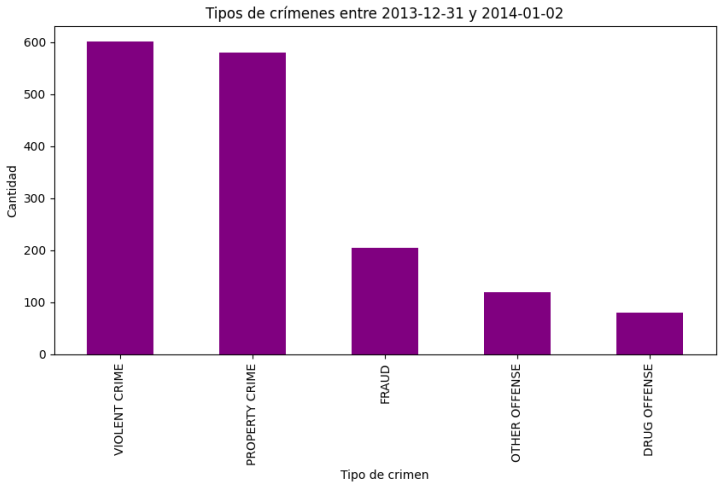

Imagen 29: Visualización de los crímenes en New Year´s Eve

Los crímenes más comunes fueron **property crime y violent crime**, con un incremento en delitos relacionados con fraudes y engaños, probablemente asociado al consumo elevado y la interacción social.

Patrones de crímenes según el tiempo

- **Durante la noche:**

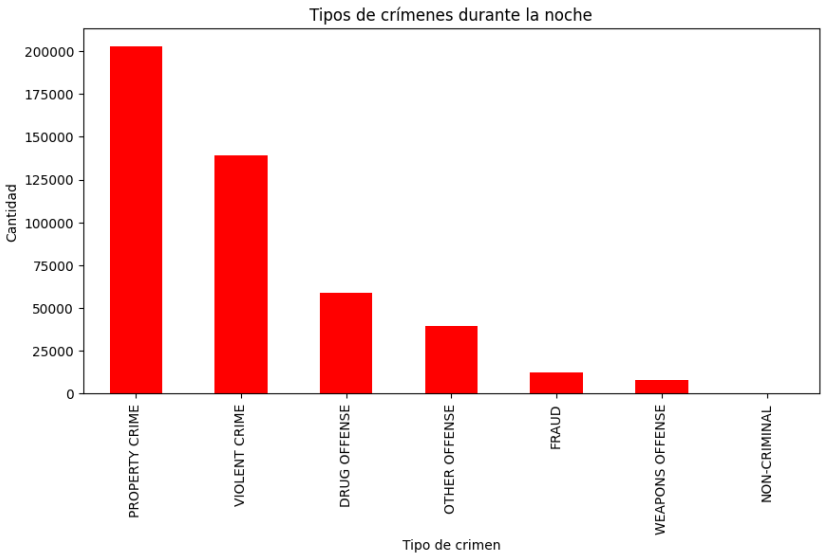

Imagen 30: Visualización de los crímenes durante la noche

Los delitos más comunes fueron **property crime y violent crime**.

- **Durante el día:**

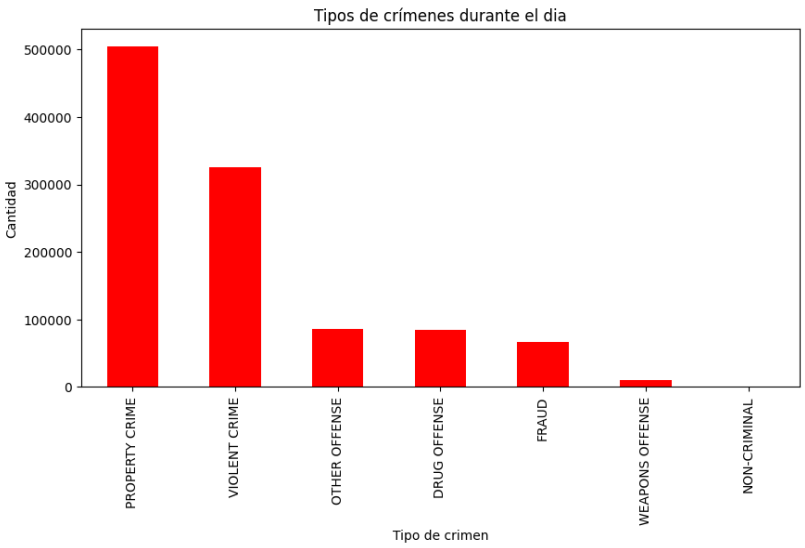

Imagen 31: Visualización de los crímenes durante el dia

Predominaron **property crime y violent crime,** reflejando diferencias significativas en los patrones según el horario y aumentando el número de estos crímenes.

Matrices de correlación y densidad

- **Matriz de correlación:**

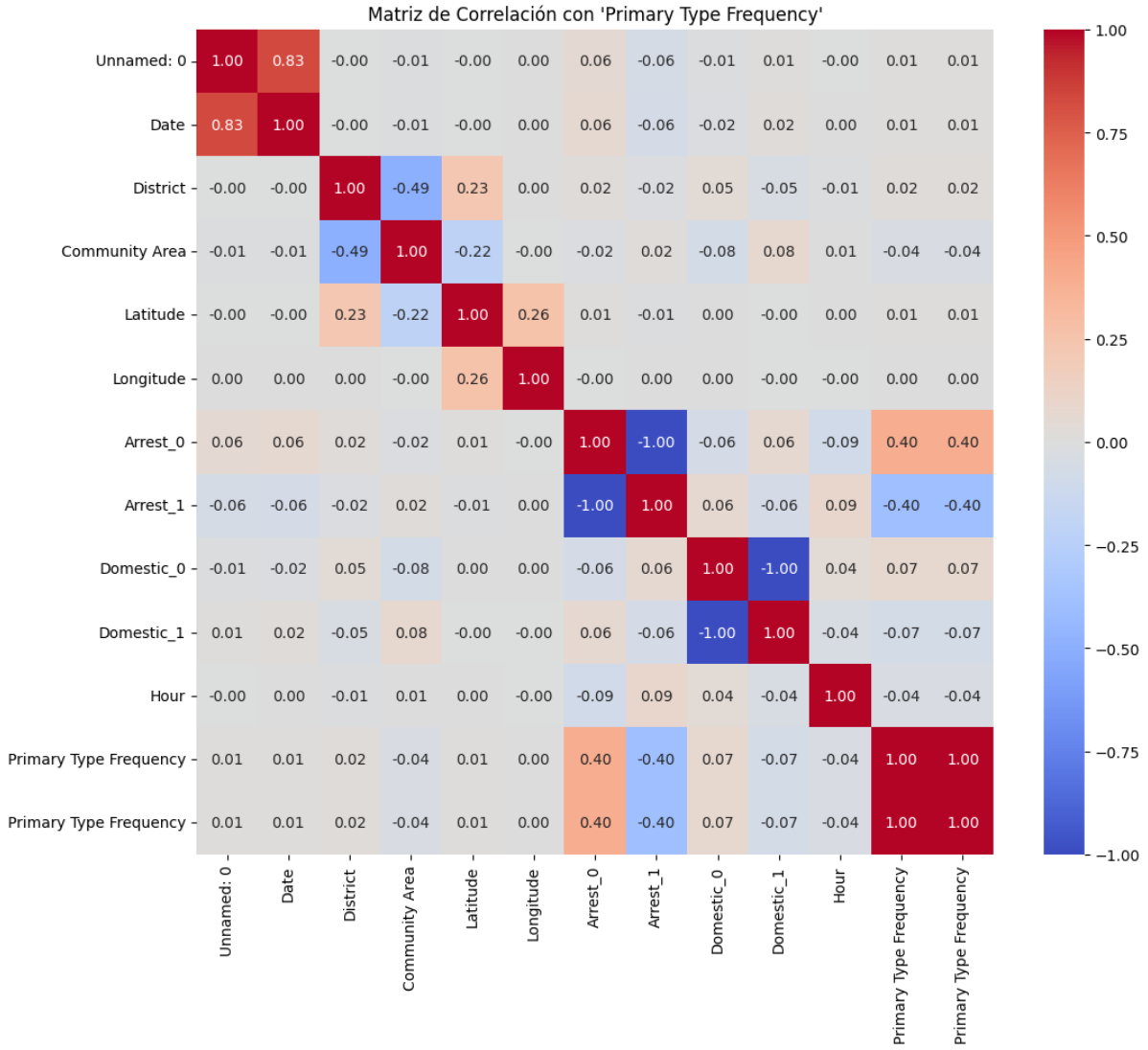

Imagen 32: Visualización de la matriz de correlación

Mostró relaciones débiles pero significativas entre variables como:

- **Community Area, Primary Type y Hour**, sugiriendo que estas variables interactúan de manera relevante para la predicción.

- **Matrices de densidad:**
    - Destacaron patrones temporales y espaciales específicos:

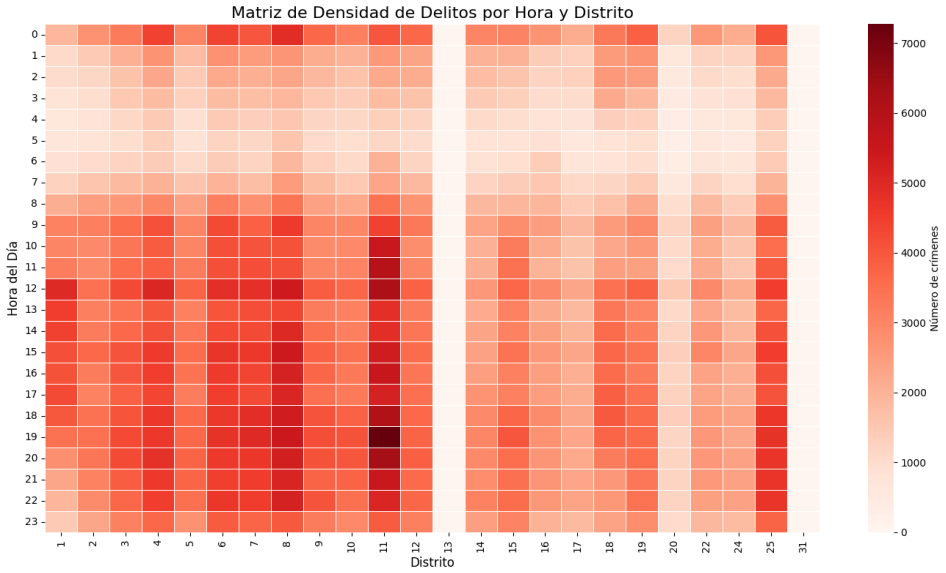

Imagen 33: Visualización de la matriz de densidad por hora y distrito

Las horas entre **17:00 y 22:00** mostraron una alta concentración de crímenes, especialmente en ciertos distritos.

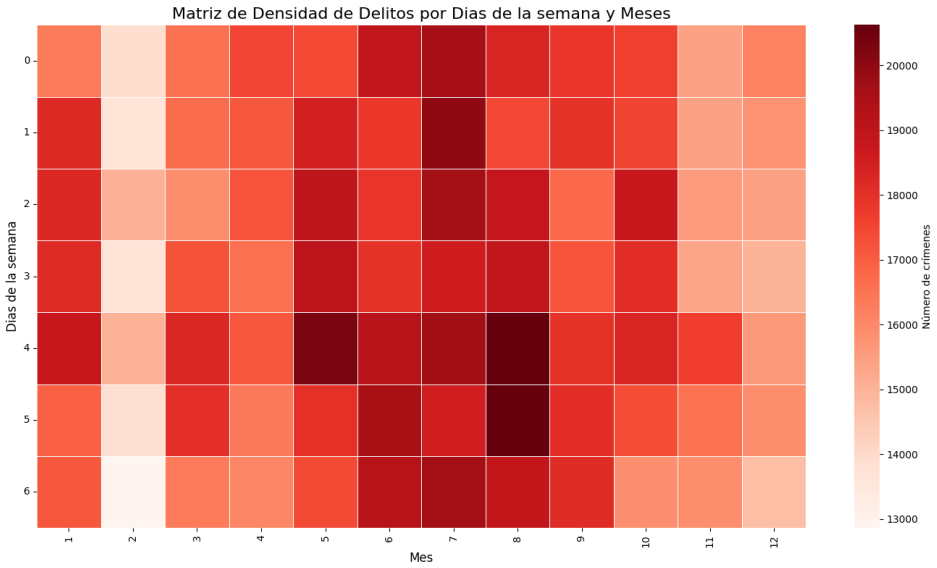

Imagen 34: Visualización de la matriz de densidad por días de la semana y meses

Los **viernes y sábados** presentaron mayor incidencia de crímenes, con picos durante los meses de verano como julio, indicando un claro patrón estacional.

Estos hallazgos proporcionaron información clave para el diseño de estrategias predictivas y la optimización de recursos policiales en función de patrones temporales, espaciales y categóricos.

RELACIÓN ENTRE LA VARIABLE OBJETIVO

El análisis de la relación entre nuestra variable objetivo, **Primary Type** (tipos de crímenes), y otras variables del dataset permitió identificar patrones significativos que serán clave para el desarrollo del modelo predictivo. A continuación, se detallan las principales relaciones encontradas:

Relación entre Primary Type y Date

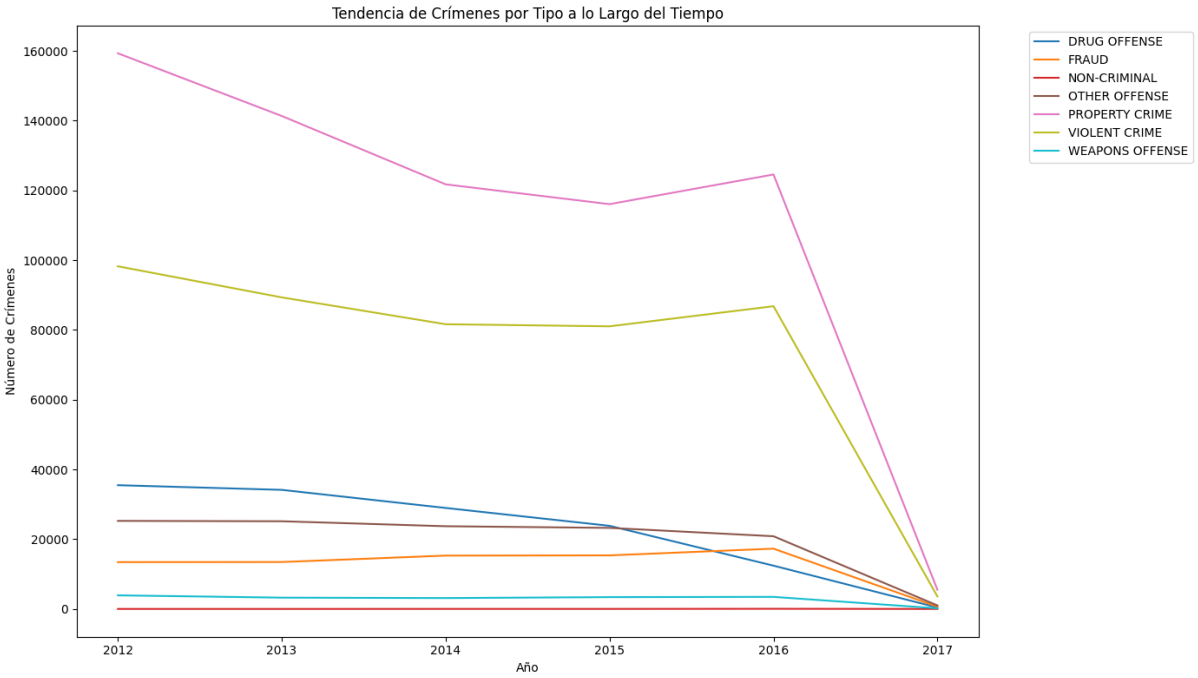

Imagen 35: Visualización de la relación entre Primary Type y Date

- Los datos revelaron una **tendencia general a la baja** en la cantidad de crímenes reportados entre **2012 y 2017**, abarcando la mayoría de las categorías delictivas.
- Sin embargo, los delitos más frecuentes, como **property crime y violent crime**, se mantuvieron persistentes a lo largo del tiempo, representando un problema constante en la actividad criminal.

Relación entre Primary Type y Block

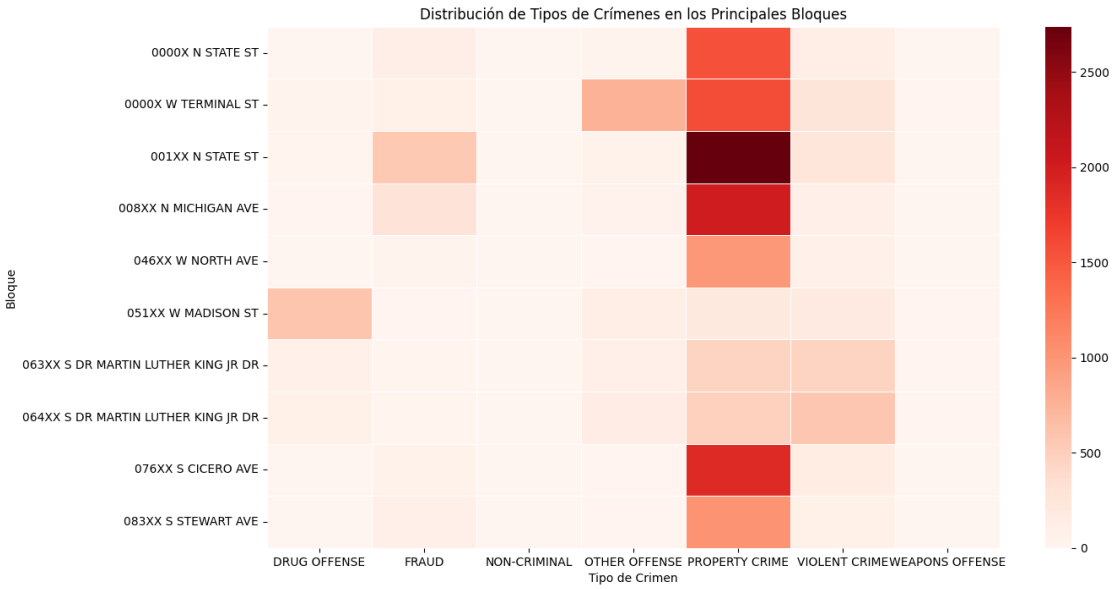

Imagen 36: Visualización de la relación entre Primary Type y Block

- Los bloques con mayor incidencia de crímenes incluyeron:
  - **000XX N STATE ST** y **008XX N MICHIGAN AVE**, que concentran delitos como **theft, battery y criminal damage**.
- Este patrón refleja una **alta actividad criminal** en áreas específicas del centro urbano, probablemente influenciada por la densidad poblacional y la actividad comercial en estas zonas.

Relación entre Primary Type y Location Description

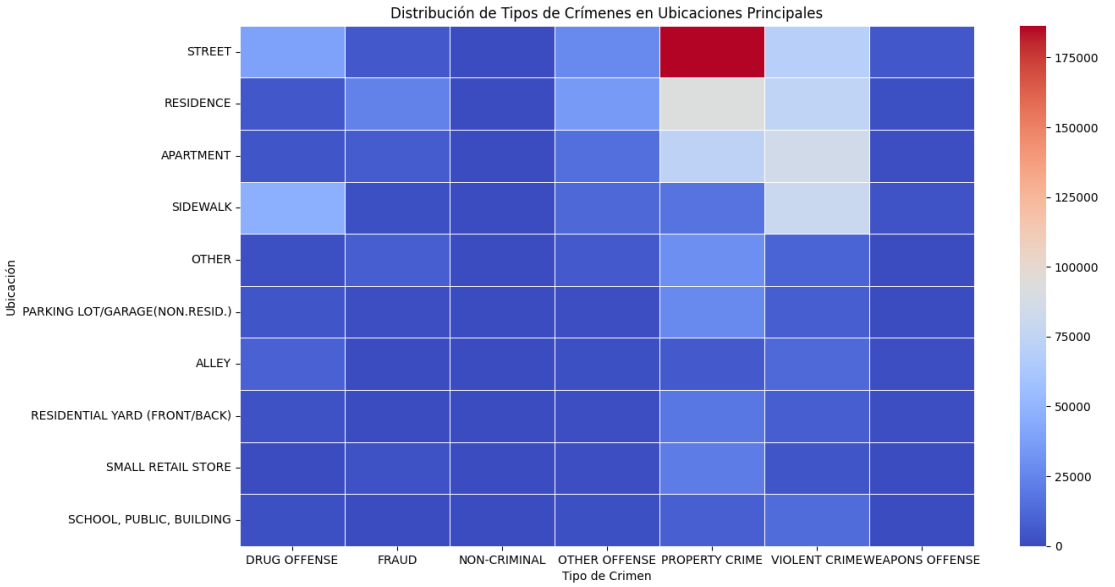

Imagen 37: Visualización de la relación entre Primary Type y Location Description

- Las ubicaciones más comunes para la ocurrencia de crímenes fueron:
    - **Street, residence y apartment**, donde predominan delitos como **property crime y violent crime**.
- Estos resultados destacan la importancia de estos espacios como **focos principales** de actividad criminal y sugieren la necesidad de medidas específicas de vigilancia en estos entornos.

Relación entre Primary Type y District

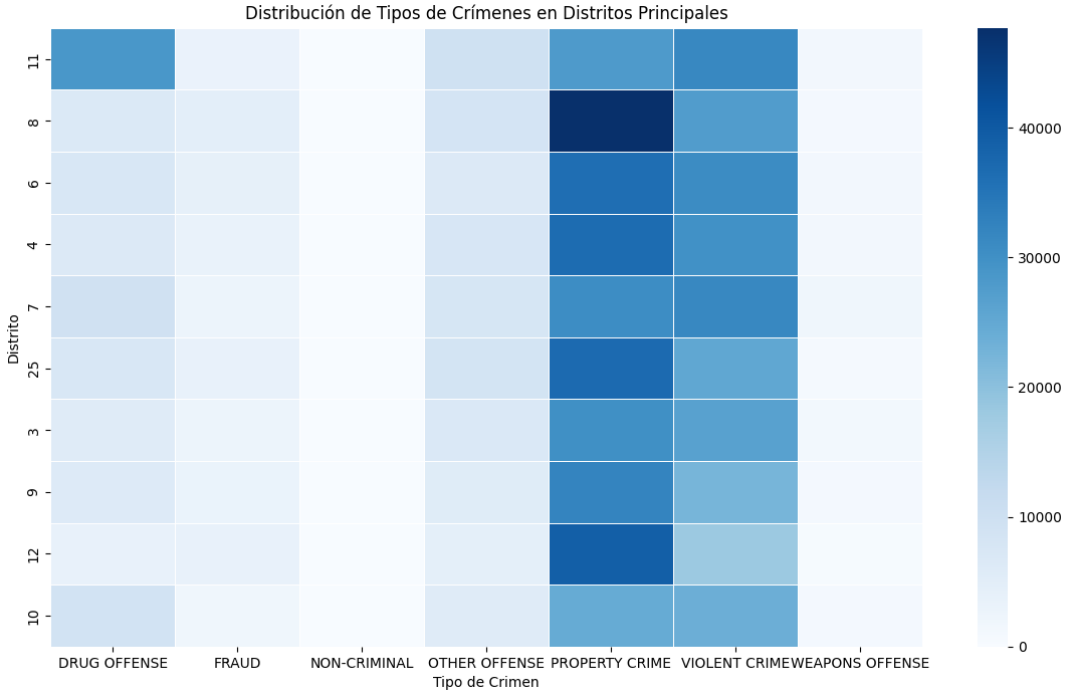

Imagen 38: Visualización de la relación entre Primary Type y District

- Los distritos con mayor cantidad de crímenes fueron:
    - **11, 8 y 6**, concentrando delitos como **primary type, violent crime y drug offense**.
- Esto indica la presencia de **focos específicos** de actividad delictiva que podrían requerir mayor atención por parte de las autoridades.

Relación entre Primary Type y Community Area

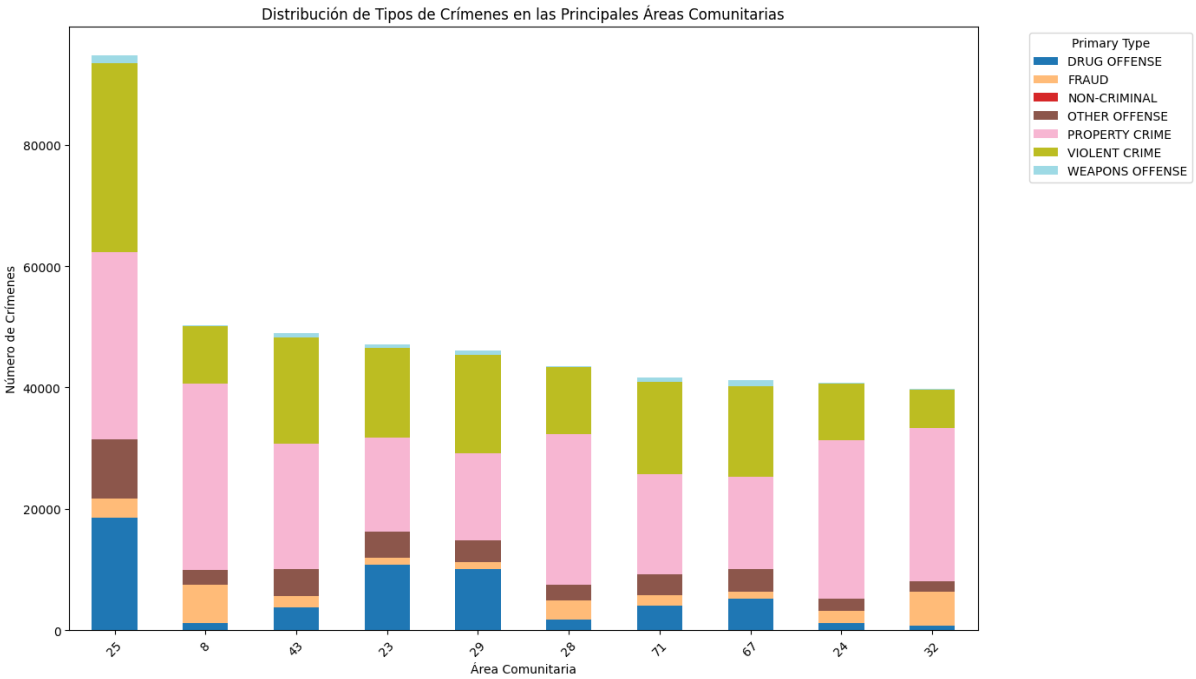

Imagen 39: Visualización de la relación entre Primary Type y Community Area

- Las áreas comunitarias con mayor número de crímenes fueron:
    - **25, 8 y 53**, donde predominan los delitos de **property crime y violent crime.**
- Este patrón refleja una persistencia en la actividad criminal en ciertas zonas comunitarias, proporcionando insights importantes para el análisis espacial y la distribución de recursos.

Conclusión

El análisis de la relación entre **Primary Type** y otras variables del dataset permitió identificar:

1. **Tendencias temporales significativas**, como la disminución general de crímenes reportados, con la persistencia de ciertas categorías delictivas.
2. **Focos espaciales clave**, como bloques, ubicaciones específicas, distritos y áreas comunitarias donde los delitos son más frecuentes.
3. **Categorías predominantes de crímenes**, como **property crime y violent crime**, que se repiten a lo largo de múltiples dimensiones.

Estos hallazgos destacan la importancia de integrar factores temporales, espaciales y categóricos en el modelo predictivo para capturar de manera efectiva los patrones de criminalidad.

ENTRENAMIENTO DE MODELOS

1. PREPARACIÓN DE LOS DATOS PARA LOS MODELOS

Separación de Variables

En primer lugar, el conjunto de datos fue dividido en variables independientes (X) y dependientes (y). Las variables independientes incluyen factores relevantes para el análisis, mientras que se descartaron columnas irrelevantes o redundantes como identificadores únicos, fechas, descripciones, entre otras. La variable dependiente seleccionada fue el "Primary Type", que representa el tipo de crimen.

Separar variables independientes (X) y dependiente (y)

```
X = dataL.drop(columns=['Primary Type', 'Date', 'Block', 'Description', 'Location Description', 'Location', 'Unnamed: 0'])
y = dataL['Primary Type']
```

```
X.head()
```

	District	Community Area	Latitude	Longitude	Arrest_0	Arrest_1	Domestic_0	Domestic_1	Hour	Day	Month	Year
0	10	29	41	-87	0	1	0	1	23	1	5	2016
1	3	42	41	-87	1	0	0	1	21	1	5	2016
2	15	25	41	-87	1	0	1	0	23	1	5	2016
3	15	25	41	-87	1	0	1	0	22	1	5	2016
4	15	25	41	-87	1	0	0	1	22	1	5	2016

```
y.head()
```

	Primary Type
0	VIOLENT CRIME
1	VIOLENT CRIME
2	OTHER OFFENSE
3	VIOLENT CRIME
4	PROPERTY CRIME

dtype: object

Imagen 40: Visualización de las separación de las variables

Manejo de Valores Nulos

Dado que la variable dependiente contenía algunos valores nulos, estos se reemplazaron con la etiqueta "UNKNOWN" para garantizar que todos los datos sean procesables por los modelos de machine learning.

```
[6] # Reemplazar valores nulos en y con "UNKNOWN"
    y.fillna("UNKNOWN", inplace=True)
```

```
[7] y.unique()
```

```
array(['VIOLENT CRIME', 'OTHER OFFENSE', 'PROPERTY CRIME',
       'WEAPONS OFFENSE', 'FRAUD', 'DRUG OFFENSE', 'NON-CRIMINAL',
       'UNKNOWN'], dtype=object)
```

Imagen 41: Visualización del manejo de valores nulos

División en Conjuntos de Entrenamiento y Prueba

El conjunto de datos fue dividido en dos subconjuntos: un 80% para entrenamiento y un 20% para prueba. Este procedimiento asegura que los modelos sean entrenados y evaluados en datos diferentes, evitando sobre-ajustes y mejorando su capacidad de generalización. Además, se codificaron las etiquetas de la variable dependiente para convertirlas en un formato numérico comprensible por los algoritmos.

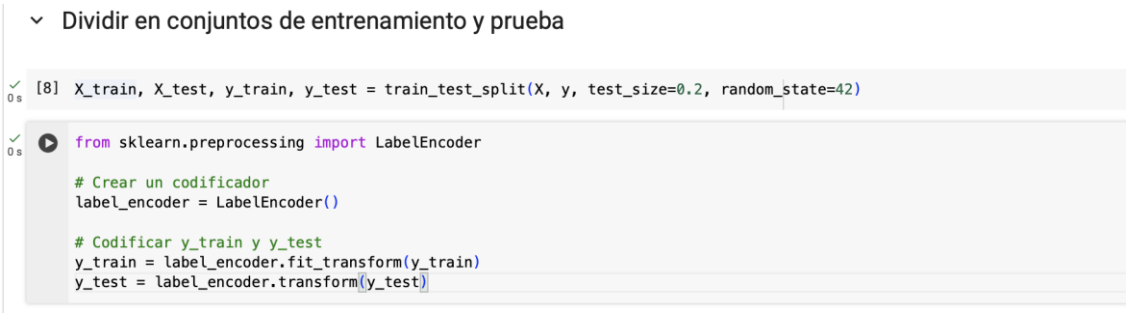

Imagen 42: Visualización de la división de entrenamiento y prueba

2. ENTRENAMIENTO DE MODELOS SUPERVISADOS IMPLEMENTADOS

1. Modelo de Árbol de Decisión (Criterio GINI)

El Árbol de Decisión utilizando el criterio de GINI fue uno de los primeros modelos entrenados. A continuación, se detallan los pasos realizados, junto con los resultados obtenidos:

Entrenamiento del Modelo

El modelo fue entrenado utilizando hiperparámetros ajustados para lograr una mejor interpretación visual del árbol de decisión:

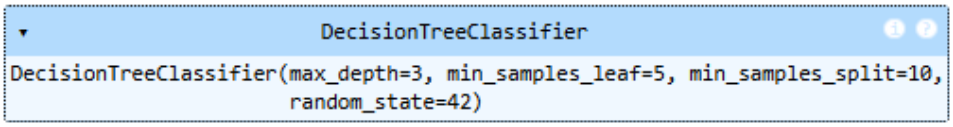

Imagen 43: Visualización del entrenamiento

- Criterio: gini
- max_depth: 3
- min_samples_split: 10
- min_samples_leaf: 5

La configuración de estos parámetros garantiza que el árbol no sea excesivamente complejo y mantenga un equilibrio entre generalización y precisión. En la imagen, se observa el código de entrenamiento y la confirmación de los parámetros aplicados.

Evaluación del Modelo

Para evaluar el modelo, se generaron predicciones sobre el conjunto de prueba y se calcularon las métricas de clasificación. Los resultados incluyen:

- Accuracy global: 59%.
- F1-score ponderado: 53%.

Como se aprecia en el Classification Report, las clases más representadas obtuvieron mejores métricas, mientras que las clases minoritarias presentaron dificultades en la predicción.

Matriz de Confusión

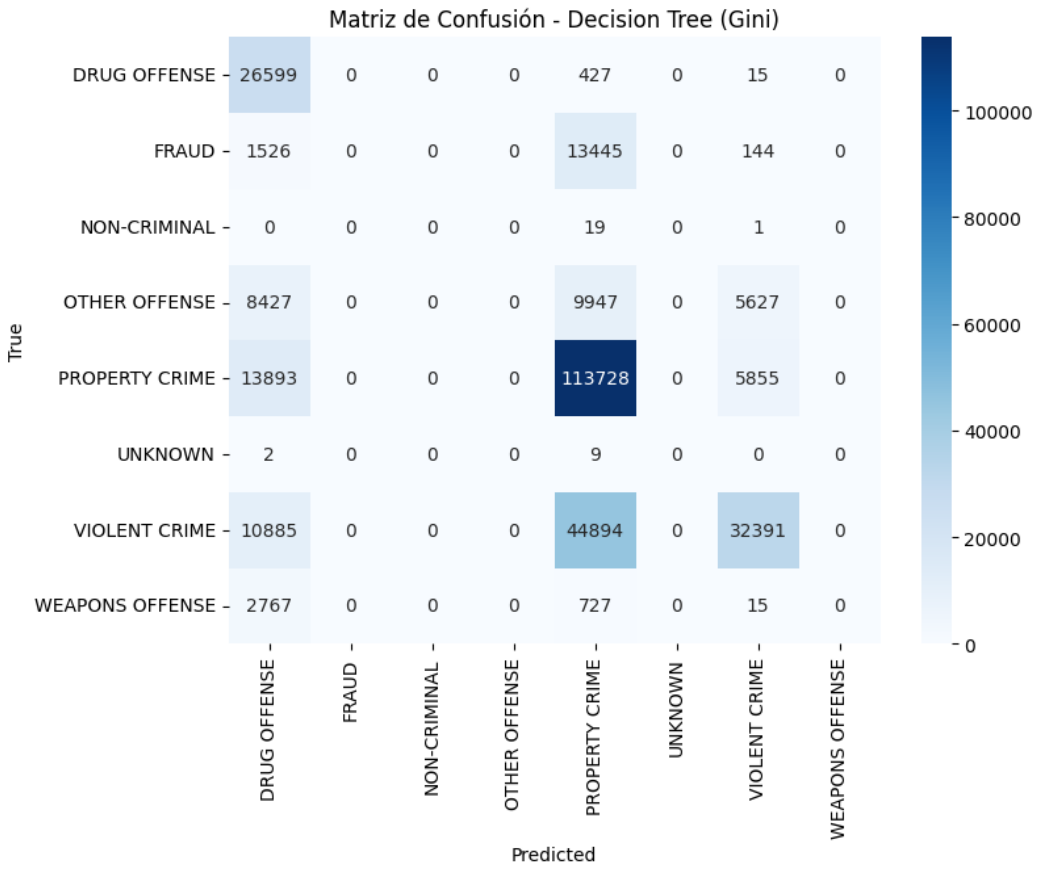

Imagen 44: Visualización de la matriz de confusión

La matriz de confusión revela cómo el modelo clasifica los datos de prueba entre las distintas categorías. Las categorías más frecuentes, como **Property Crime y Violent Crime**, presentan una mayor precisión en la clasificación. Sin embargo, algunas clases menos representadas, como **Weapons Offense**, presentan errores de clasificación significativos.

Visualización del Árbol de Decisión

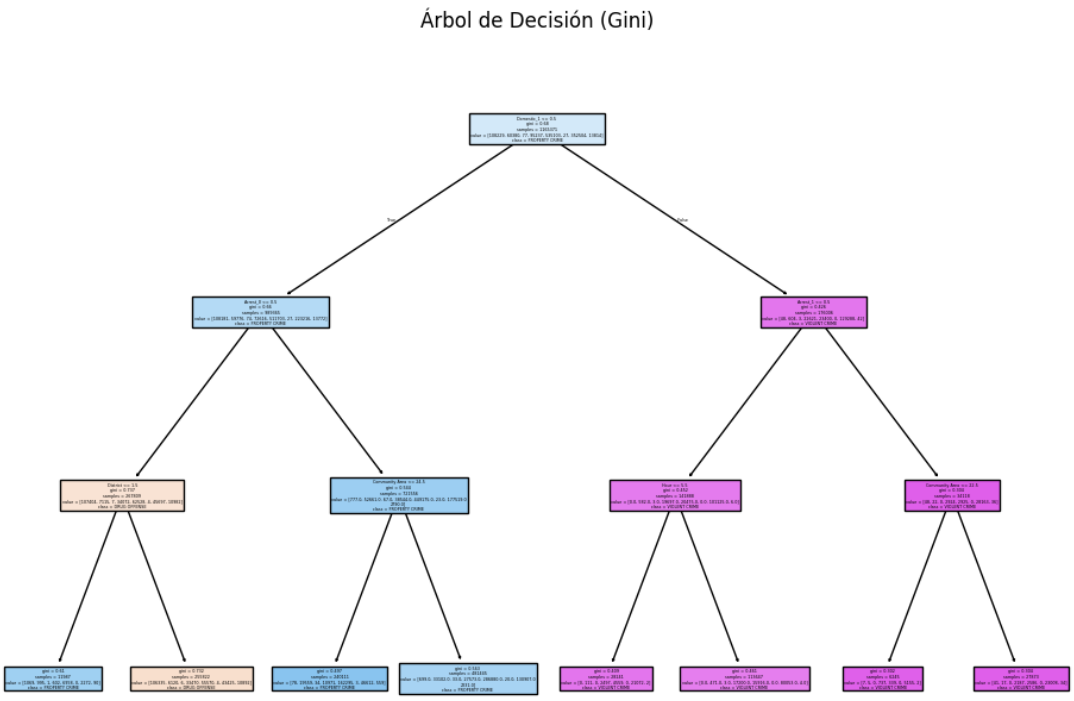

Imagen 45: Visualización del árbol de decisión

El árbol de decisión generado muestra las divisiones realizadas por el modelo en función de las características seleccionadas. Cada nodo incluye información sobre:

- La regla de decisión aplicada.
- La cantidad de muestras clasificadas en el nodo.
- La distribución de clases dentro de cada nodo.

Esta visualización permite interpretar las decisiones del modelo y verificar cómo cada característica contribuye a las predicciones

2. Modelo de Árbol de Decisión (Criterio ENTROPY)

Entrenamiento del Modelo

Se entrenó un Árbol de Decisión utilizando el criterio ENTROPY. Los hiper parámetros ajustados fueron:

```
DecisionTreeClassifier
DecisionTreeClassifier(criterion='entropy', max_depth=3, min_samples_leaf=5,
                       min_samples_split=10, random_state=42)
```

Imagen 46: Visualización del entrenamiento

- **min_samples_split**: 10

- **min_samples_leaf**: 5
- **max_depth**: 3

Evaluación del Modelo

El modelo con el criterio ENTROPY también alcanzó una precisión del 59%. Al igual que con GINI, la distribución desbalanceada de clases afectó el rendimiento en las clases minoritarias.

Matriz de Confusión

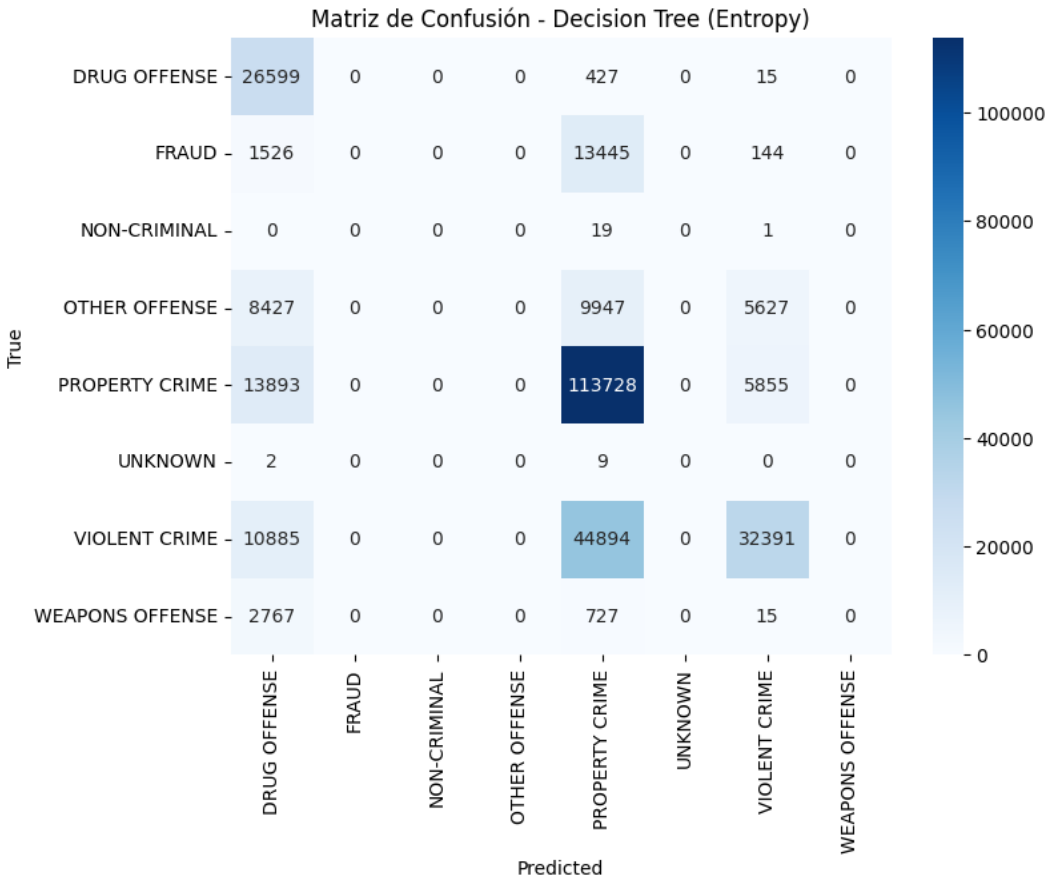

Imagen 47: Visualización de la matriz de confusión

La matriz de confusión refleja un patrón similar al modelo GINI, destacando una mejor predicción en clases mayoritarias y resultados débiles en las minoritarias.

Visualización del Árbol

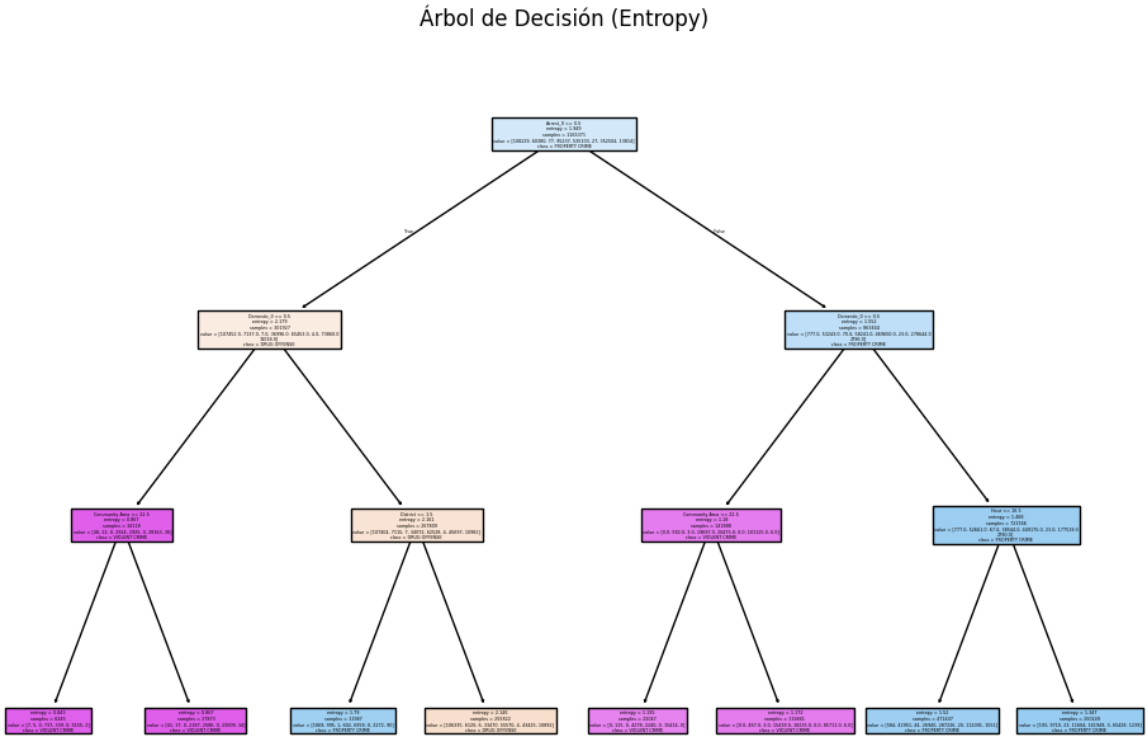

Imagen 48: Visualización del árbol de decisión

La visualización del árbol entrenado con ENTROPY permite observar las reglas generadas basadas en este criterio, mostrando diferencias leves en comparación con el árbol basado en GINI.

3. Modelo de Random Forest

Entrenamiento del Modelo

Se entrenó un modelo Random Forest utilizando 250 estimadores (árboles) con un estado aleatorio fijo para garantizar reproducibilidad. Este modelo se caracteriza por la combinación de múltiples árboles de decisión para mejorar la precisión y reducir el sobreajuste.

```
RandomForestClassifier
RandomForestClassifier(n_estimators=250, random_state=42)
```

Imagen 49: Visualización del entrenamiento

Evaluación del Modelo

El modelo Random Forest alcanzó una precisión general del 54%. Aunque mostró mejoras en la clasificación de algunas clases en comparación con los árboles de decisión individuales, las clases minoritarias aún enfrentaron desafíos significativos en términos de precisión y recall.

Matriz de Confusión

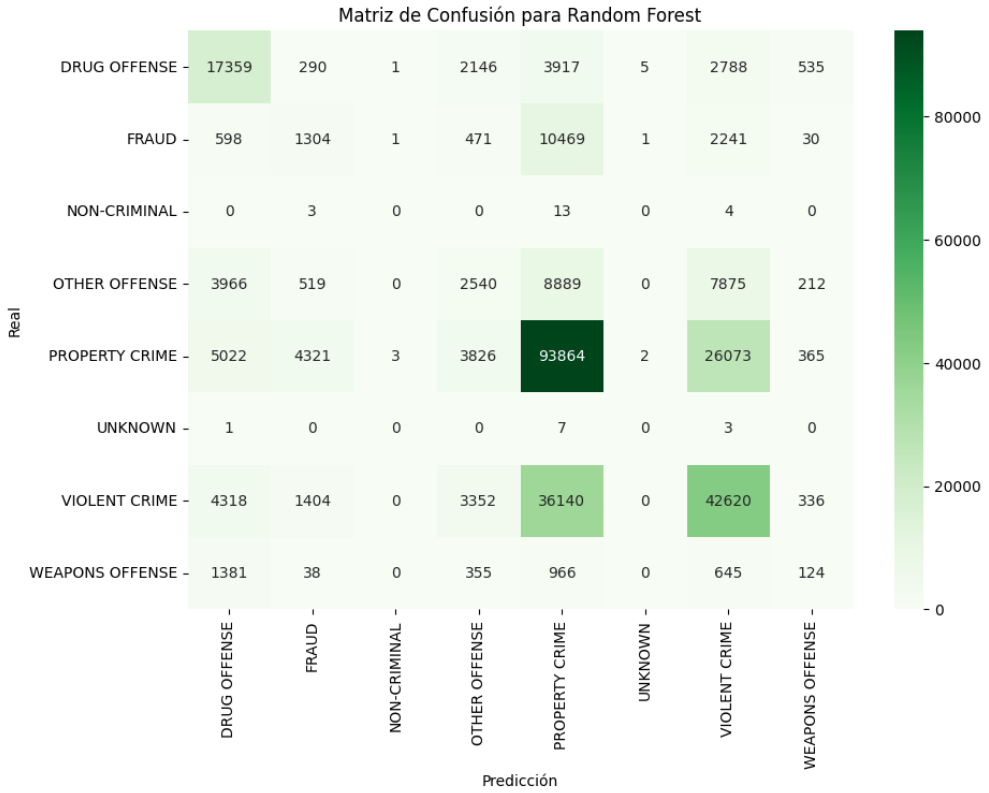

Imagen 50: Visualización de la matriz de confusión

La matriz de confusión muestra una mejor distribución de predicciones en comparación con los modelos de árbol individuales. Sin embargo, aún se observan dificultades en la clasificación de clases menos representadas.

Importancia de las Características

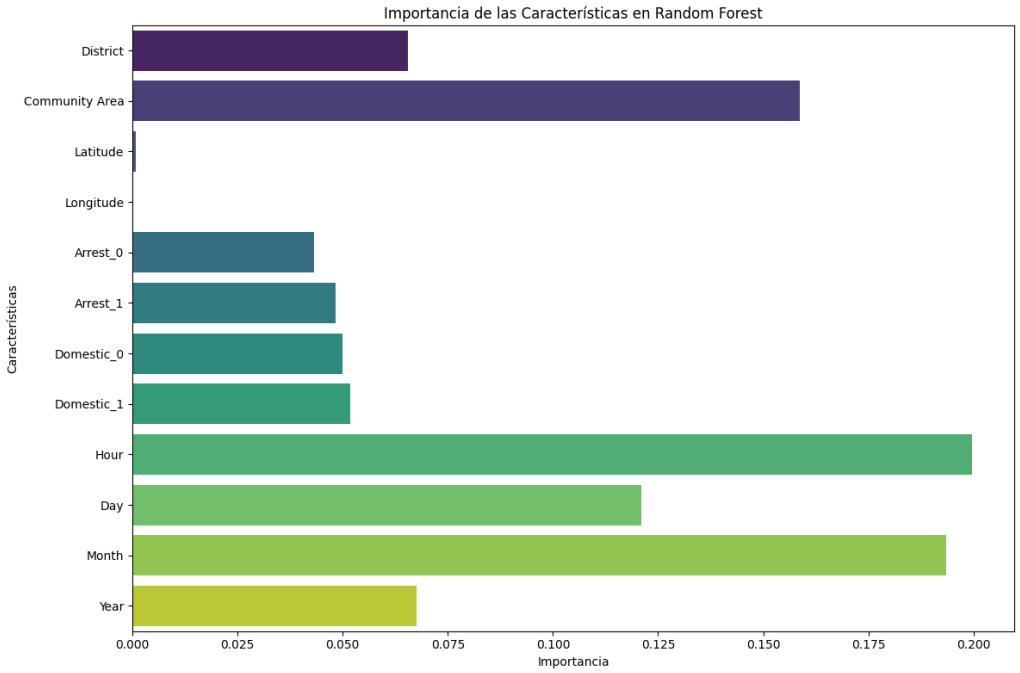

Imagen 51: Visualización de la importancia de las características

Se generó un gráfico que muestra la importancia de cada característica utilizada en el modelo. Factores como la hora del día, el mes y la ubicación (**Community Area)** se destacaron como los más influyentes en las predicciones del modelo.

4. Modelo de Gradient Boosting

Entrenamiento del Modelo

El modelo de **Gradient Boosting** fue entrenado utilizando la configuración predeterminada del clasificador **GradientBoostingClassifier** en Python. Este modelo combina múltiples árboles de decisión de manera secuencial para minimizar el error y mejorar la precisión. Se utilizó un conjunto de datos dividido en entrenamiento y prueba para entrenar el modelo con un estado aleatorio fijo de 42 para asegurar reproducibilidad.

```
GradientBoostingClassifier
GradientBoostingClassifier(random_state=42)
```

Imagen 52: Visualización del entrenamiento

Evaluación del Modelo

El reporte de clasificación muestra una precisión general del **60%**, con un mejor rendimiento en las clases mayoritarias, como "PROPERTY CRIME", y una menor precisión en las clases minoritarias, como "NON-CRIMINAL". Aunque el modelo logra un mejor balance en comparación con otros enfoques, aún enfrenta dificultades al clasificar correctamente las categorías menos representadas debido a la naturaleza desbalanceada del conjunto de datos.

Matriz de Confusión

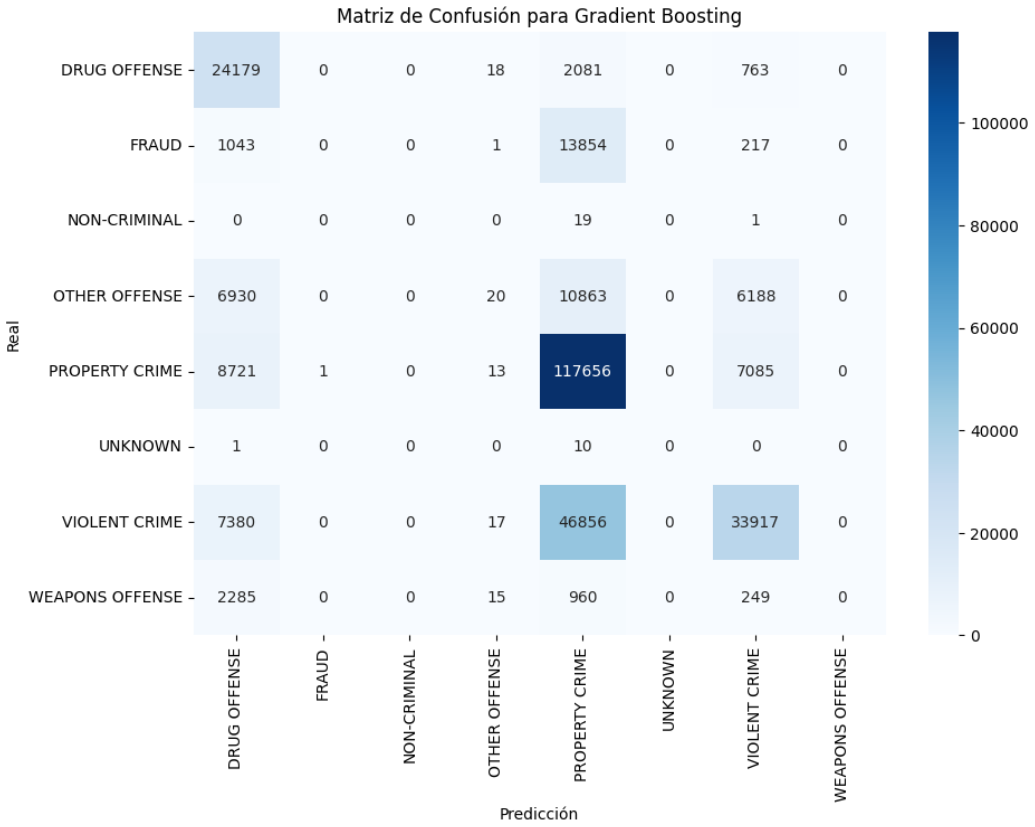

Imagen 53: Visualización de la matriz de confusión

La matriz de confusión generada para este modelo refleja una fuerte capacidad para identificar correctamente los delitos más comunes, como "PROPERTY CRIME" y "VIOLENT CRIME", pero errores significativos en las predicciones de clases con menos datos disponibles. Estas observaciones resaltan la necesidad de ajustar el modelo o implementar técnicas para manejar datos desbalanceados.

Visualización de un Árbol de Decisión en Gradient Boosting

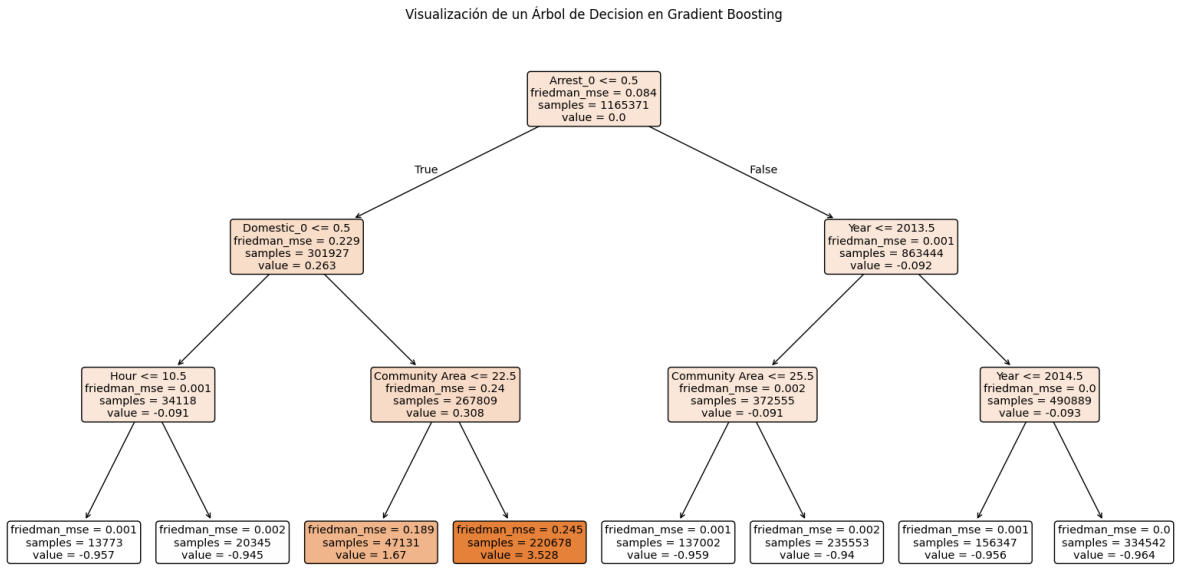

Imagen 54: Visualización del árbol

Una representación gráfica de uno de los árboles de decisión utilizados en el modelo muestra cómo las características clave, como el área comunitaria (**Community Area**), las horas (**Hour**) y el historial de arrestos (**Arrest_0**), influyen en las decisiones del modelo. Esta visualización es útil para interpretar el modelo y comprender cómo se establecen las reglas que determinan las predicciones.

5. Modelo K-Nearest Neighbors (KNN)

Entrenamiento del Modelo

El modelo KNN se entrenó utilizando un número de vecinos igual a 7 (n_neighbors=7). Este valor se seleccionó de manera arbitraria, con la posibilidad de ajustarlo posteriormente mediante optimización de hiperparámetros.

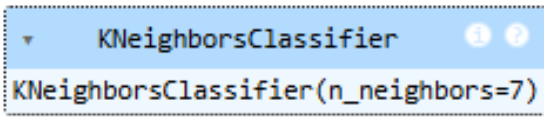

Imagen 55: Visualización del entrenamiento

Evaluación del Modelo

El desempeño del modelo se evaluó en el conjunto de prueba. El reporte de clasificación mostró los siguientes resultados:

- Precisión general: **56%**
- Pobre desempeño en clases minoritarias debido al desbalance de datos.

- El valor promedio ponderado de precisión, recall y f1-score fue bajo, lo que indica que el modelo podría no ser adecuado para datos con distribuciones desiguales.

Matriz de Confusión

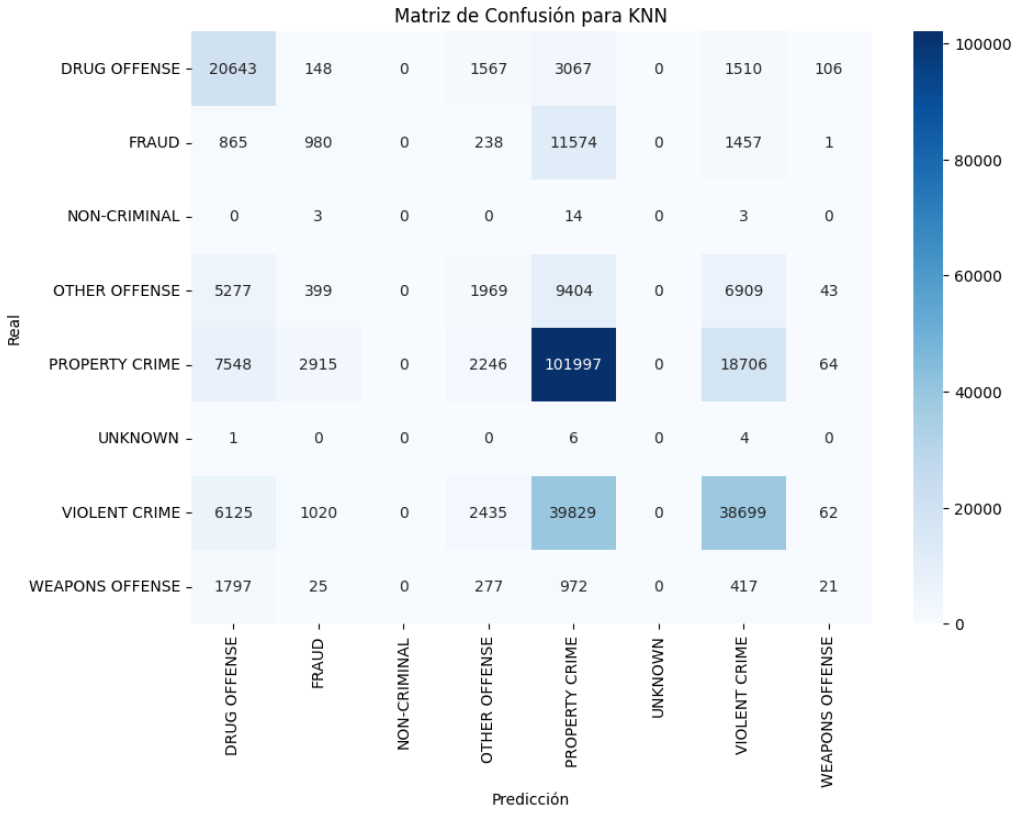

Imagen 56: Visualización de la matriz de confusión

La matriz de confusión reveló que:

- Las clases mayoritarias, como "PROPERTY CRIME" y "VIOLENT CRIME", fueron predichas con mayor precisión.
- Clases con menos representación, como "NON-CRIMINAL", presentaron un desempeño débil, reflejando el impacto del desbalance en los datos.

Visualización del Modelo

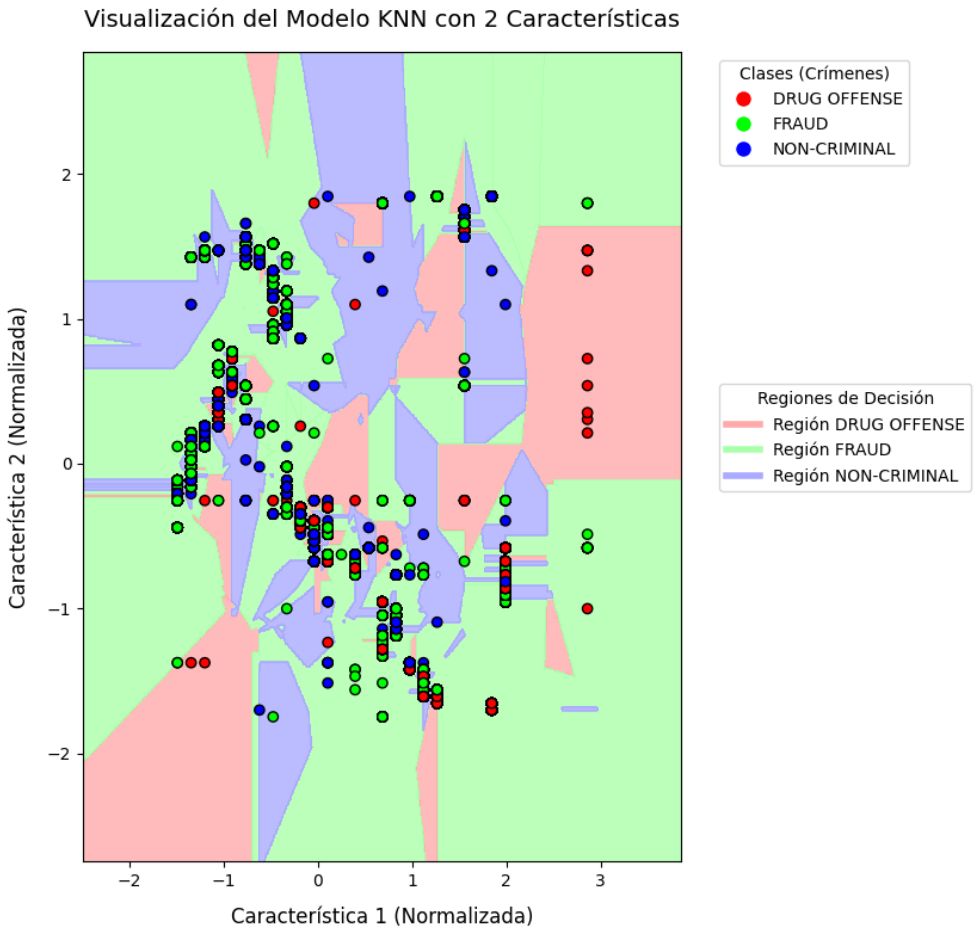

Imagen 57: Visualización del modelo

EJE X = '**District**' y EJE Y = '**Comunity Area**', los puntos que se pintan son los tipos de crímenes que más inciden entre esos dos puntos.

El modelo KNN (K-Nearest Neighbors):

- El modelo clasifica cada crimen observando los puntos más cercanos (sus "vecinos").Por ejemplo:
    - Si un punto nuevo cae en una región verde, el modelo lo calificaría como "FRAUD".
    - Si cae en una región roja, lo calificaría como "**DRUG OFFENSE"**.
    - Patrones en los datos:
        - Hay zonas claras donde ciertos tipos de crímenes son dominantes.
    - Ejemplo:
        - Muchas zonas verdes (FRAUD) están mezcladas con zonas azules (NON-CRIMINAL), lo que indica que estas clases podrían estar más relacionadas entre sí.
        - Las zonas rojas (DRUG OFFENSE) parecen más separadas, lo que sugiere que esta clase podría ser más distinta de las otras.

El modelo KNN demostró limitaciones en la clasificación de clases minoritarias y en conjuntos de datos desbalanceados. Esto indica que el algoritmo podría beneficiarse de técnicas de ajuste adicionales, como el balanceo de clases o la optimización de hiper parámetros.

6. Modelo Redes Neuronales

Preprocesamiento de Datos

Se normalizaron las características utilizando Standard Scaler, y la variable objetivo fue codificada en formato categórico (one-hot encoding) para ser compatible con el modelo de red neuronal.

Construcción del Modelo

Model: "sequential"

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 128)	1,664
dense_1 (Dense)	(None, 64)	8,256
dense_2 (Dense)	(None, 8)	520

Total params: 10,440 (40.78 KB)
Trainable params: 10,440 (40.78 KB)
Non-trainable params: 0 (0.00 B)

Imagen 58: Visualización de la construcción del modelo

La arquitectura de la red neuronal incluyó:

- Una capa de entrada con 128 neuronas y función de activación ReLU.
- Una capa oculta con 64 neuronas y función de activación ReLU.
- Una capa de salida con tantas neuronas como clases, utilizando una función de activación softmax.

Entrenamiento del Modelo

El modelo se entrenó durante 20 épocas con un tamaño de lote de 64. La precisión en el conjunto de validación alcanzó un valor máximo de aproximadamente 60%, mientras que la pérdida mostró una disminución constante.

Evaluación del Modelo

El modelo obtuvo una precisión global del 46% en el conjunto de prueba. Aunque su rendimiento fue inferior al de otros modelos, demostró cierta capacidad para identificar patrones en clases mayoritarias.

Matriz de Confusión

```
Matriz de Confusión:
[[     0      0      0      0  27041      0      0      0]
 [     0      0      0      0  15115      0      0      0]
 [     0      0      0      0     20      0      0      0]
 [     0      0      0      0  24001      0      0      0]
 [     0      0      0      0 133476      0      0      0]
 [     0      0      0      0     11      0      0      0]
 [     0      0      0      0  88170      0      0      0]
 [     0      0      0      0   3509      0      0      0]]
```

Imagen 59: Visualización de la matriz de confusión

La matriz de confusión muestra que el modelo predice adecuadamente la clase "**PROPERTY CRIME**", pero tiene dificultades significativas con otras clases.

Evolución de Métricas

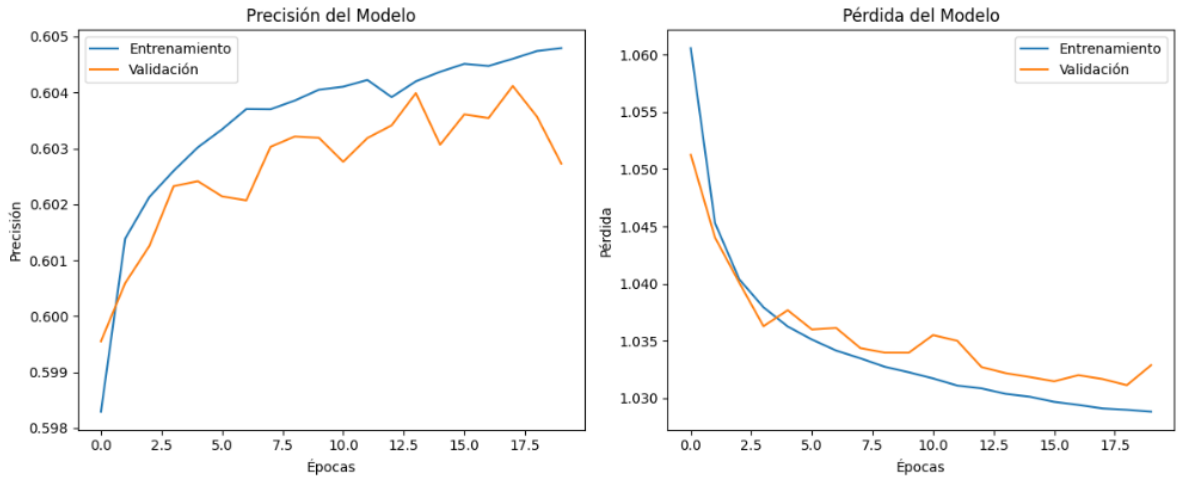

Imagen 60: Visualización de las métricas

Los gráficos de precisión y pérdida durante el entrenamiento y validación muestran una mejora constante a lo largo de las épocas, reflejando el aprendizaje del modelo.

3. ENTRENAMIENTO DE MODELOS NO SUPERVISADOS IMPLEMENTADOS

1. Modelo de K-Means

Preparación de los Datos

Para aplicar el modelo de K-Means, los datos fueron preparados eliminando columnas irrelevantes, como información categórica y variables redundantes. Posteriormente, se normalizaron las características utilizando Standard Scaler para garantizar que todas las variables contribuyeran de manera equitativa al cálculo de las distancias en el espacio de características.

Método del Codo

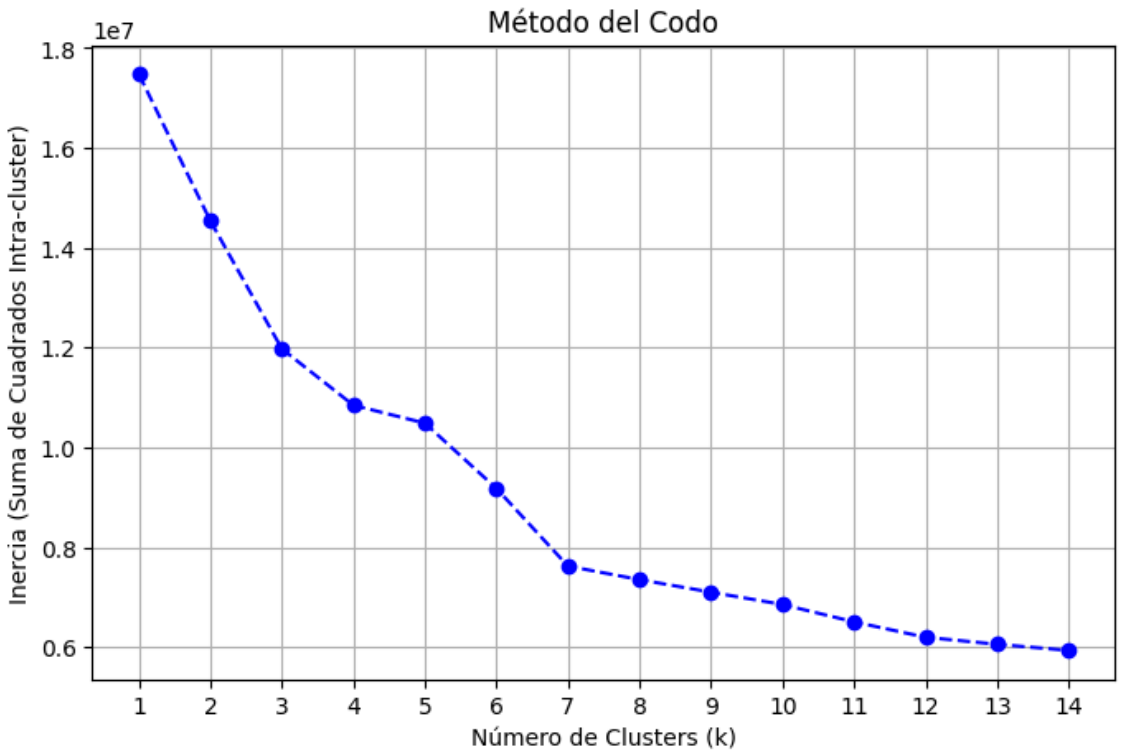

Imagen 61: Visualización del método del codo

Para determinar el número óptimo de clústeres, se aplicó el método del codo. Este método consiste en calcular la inercia (suma de las distancias intra-clúster) para diferentes valores de k, entre 1 y 15. La visualización generada muestra un cambio significativo en la pendiente alrededor de k=5, indicando que este es el número óptimo de clústeres para el modelo.

Ajuste del Modelo

El modelo de K-Means fue ajustado utilizando el valor óptimo de k=5. Cada punto en el espacio de características fue asignado a uno de los cinco clústeres, y se generó una representación visual que destaca las regiones de los clústeres, junto con sus centroides.

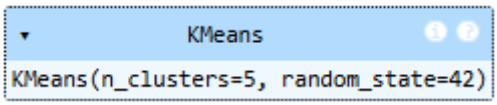

Imagen 62: Visualización del entrenamiento

Visualización

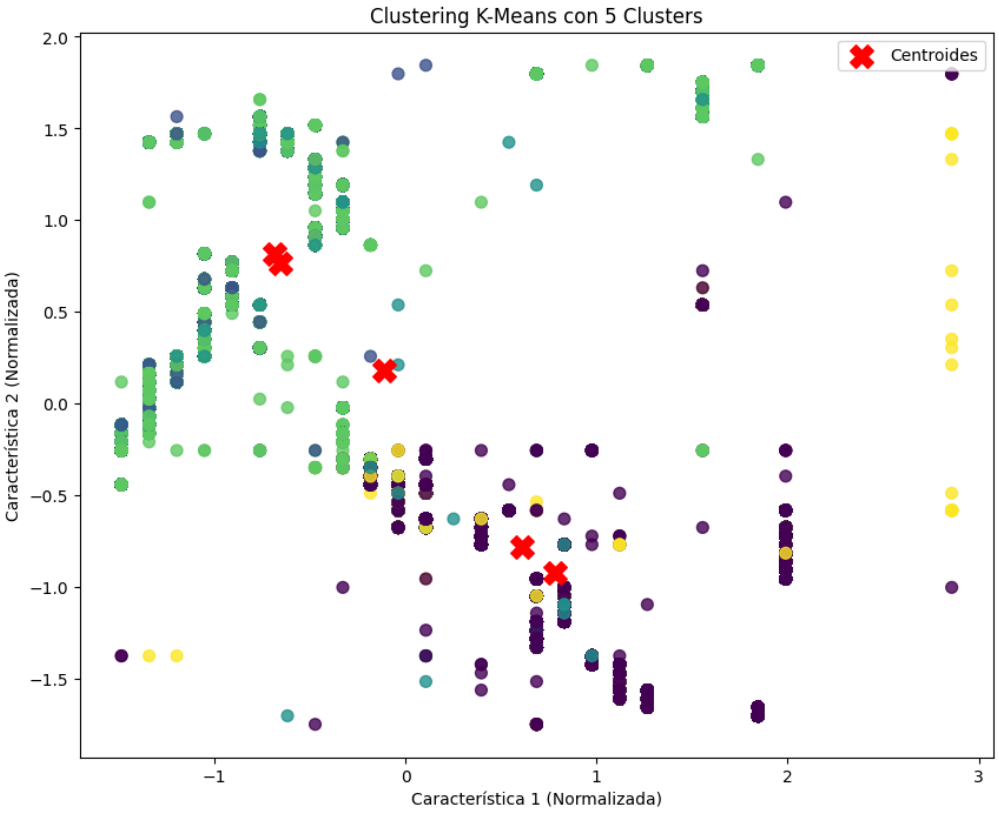

Imagen 63: Visualización de las métricas

La gráfica muestra cómo se distribuyen los datos en torno a los centroides de los clústeres, proporcionando una interpretación visual de la agrupación.

4. COMPARACIÓN DE MODELOS Y DECISIONES PARA MEJORAR LA PRECISIÓN

Comparación de Desempeño

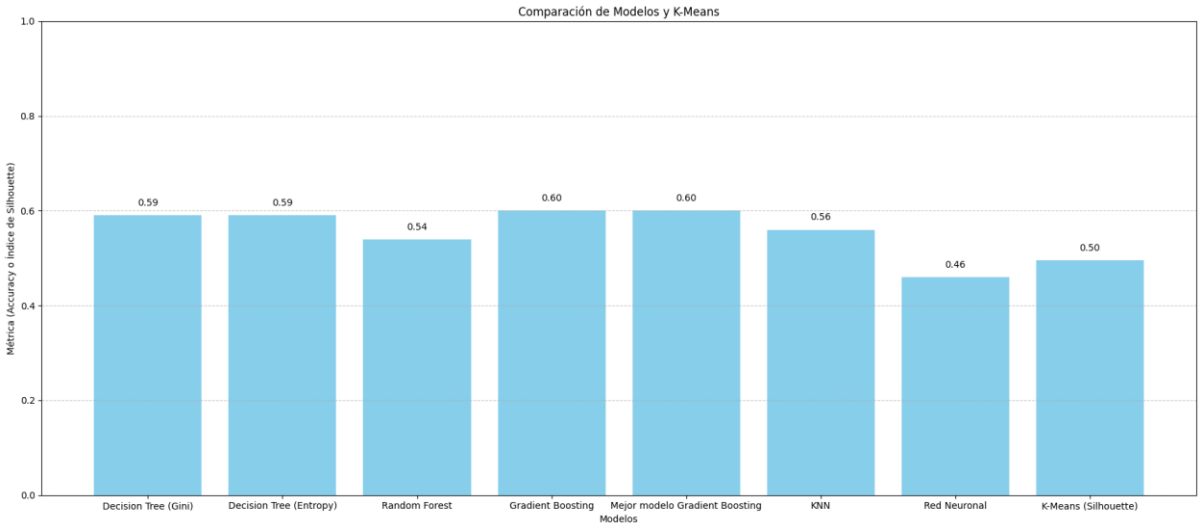

Imagen 64: Visualización de las comparaciones de los modeloa

La gráfica anterior muestra la comparación de los modelos en términos de su métrica principal, ya sea **accuracy** para los modelos supervisados o el **índice de Silhouette** para el modelo de K-Means. Los resultados clave son los siguientes:

1. **Gradient Boosting** alcanzó el mejor desempeño con una precisión de **0.60**, destacándose como el modelo más efectivo para clasificar los datos.
2. Los modelos de **Árbol de Decisión (Gini y Entropy)** lograron una precisión de **0.59**, demostrando una capacidad moderada para la clasificación.
3. El modelo de **Random Forest** obtuvo una precisión de **0.54**, lo que sugiere un desempeño consistente pero menos efectivo que Gradient Boosting.
4. **KNN** alcanzó una precisión de **0.56**, lo que muestra un desempeño aceptable en comparación con otros modelos supervisados.
5. La **Red Neuronal** registró una precisión de **0.46**, indicando que su arquitectura actual puede no ser óptima para este conjunto de datos.
6. El modelo de **K-Means**, evaluado mediante el índice de Silhouette, obtuvo un puntaje de **0.19**, lo que sugiere que los grupos encontrados no son bien definidos en el espacio de características.

5. MEJORA DE MODELOS

Debido a una falta de tiempo y recursos, decidimos realizar las mejoras a determinados modelos que nos parecieron más interesantes.

Con base en los resultados, se han identificado áreas de mejora para los siguientes modelos con el objetivo de aumentar su desempeño:

Modelo: Gradient Boosting

Se intentaron varias técnicas para mejorar el rendimiento del modelo:

1. **Ajuste de Hiperparámetros:**

```
GradientBoostingClassifier
GradientBoostingClassifier(max_depth=5, min_samples_leaf=5, n_estimators=200,
                           random_state=42)
```

Imagen 65: Visualización del entrenamiento

   - Se probaron diferentes valores para los parámetros n_estimators, learning_rate, y max_depth.
2. **Balanceo de Datos:**
   - Se aplicaron técnicas de sobremuestreo (SMOTE) para equilibrar las clases del conjunto de datos.
3. **Eliminación de Características Irrelevantes:**
   - Se realizó una selección de características utilizando SelectFromModel basada en su importancia en el modelo.
4. **Regularización con subsample:**
   - Se introdujo un valor menor a 1 en el parámetro subsample para reducir el sobreajuste.

Resultados y Conclusión

Tras aplicar estas técnicas, la precisión se mantuvo constante en el **60%**, indicando que el modelo probablemente ha alcanzado su límite de rendimiento con el conjunto de datos actual. Esto sugiere que el modelo es robusto pero podría beneficiarse de:

1. **Mayor cantidad de datos**.
2. **Atributos adicionales** que capturen información más relevante.

Classification Report:

	precision	recall	f1-score	support
0	0.48	0.89	0.62	27041
1	0.00	0.00	0.00	15115
2	0.00	0.00	0.00	20
3	0.24	0.00	0.00	24001
4	0.61	0.88	0.72	133476
5	0.00	0.00	0.00	11
6	0.70	0.38	0.50	88170
7	0.00	0.00	0.00	3509
accuracy			0.60	291343
macro avg	0.25	0.27	0.23	291343
weighted avg	0.56	0.60	0.54	291343

Imagen 66: Visualización de los resultados

Modelo: Redes Neuronales

Se implementaron las siguientes técnicas, lo que permitió incrementar la precisión al 60%:

1. **Optimización del Optimizador:**
    - Se reemplazó el optimizador Adam por SGD (Stochastic Gradient Descent), añadiendo momentum para mejorar la convergencia.python
2. **Regularización:**
    - Se utilizó Dropout para evitar el sobreajuste. Este método desactiva aleatoriamente neuronas durante el entrenamiento.
3. **Early Stopping:**
    - Se implementó EarlyStopping para detener el entrenamiento cuando la pérdida en validación dejara de mejorar.
4. **Ajuste del Número de Épocas:**
    - Se incrementó el número de épocas a 50 para permitir que el modelo aprenda patrones más complejos, manteniendo el monitoreo con EarlyStopping.
5. **Batch Normalization:**
    - Se añadió normalización de lotes para acelerar la convergencia y estabilizar el entrenamiento.
6. **Incremento del Tamaño del Modelo:**
    - Se añadieron más capas y neuronas para mejorar la capacidad del modelo de capturar patrones complejos.

Resultados

- Precisión Inicial: 40%
- Precisión Final (Mejorada): 60%

Conclusión

Las técnicas aplicadas llevaron a un aumento significativo en la precisión del modelo, haciéndolo comparable con Gradient Boosting. Sin embargo, el modelo aún puede beneficiarse de:

- Mayor cantidad de datos.
- Mayor afinación de hiperparámetros utilizando herramientas como Optuna.

MEJORAS FURRAS

- **Optimización Adicional:** Usar herramientas como *Optuna* o *Grid Search* para refinar hiperparámetros.
- **Mayor Recolección de Datos:** Incorporar atributos relevantes para capturar un panorama más amplio del comportamiento criminal.
- **Técnicas de Balanceo:** Implementar métodos como *SMOTE* para reducir el impacto del desbalance de clases.

CONCLUSIONES

Este proyecto tuvo como objetivo principal el desarrollo de un modelo de aprendizaje automático para predecir el tipo de crimen más probable en Chicago, utilizando un extenso conjunto de datos históricos. A continuación, se destacan las principales conclusiones en cada etapa del trabajo:

1. Carga y Análisis Exploratorio de Datos (EDA):

- **Tamaño y Estructura del Dataset:** El conjunto de datos incluyó más de 1.4 millones de registros con 23 columnas, lo que ofreció un panorama amplio de los delitos en Chicago.
- **Distribución de Crímenes:** Los delitos más comunes fueron *Property Crime* y *Violent Crime*, mientras que delitos como *Weapons Offense* y *Non-Criminal* fueron minoritarios.
- **Tendencias Temporales:** Se observaron picos de criminalidad durante ciertos meses (verano) y horarios (tardes y noches), lo que indica patrones estacionales y horarios significativos.
- **Patrones Geográficos:** Las "zonas calientes" de criminalidad se concentraron en distritos específicos como el 11, 8 y 6, lo que refleja la necesidad de estrategias focalizadas.

2. Preprocesamiento:

- **Limpieza de Datos:** Se trataron valores nulos y duplicados, y se eliminaron columnas irrelevantes (e.g., *ID*, *Case Number*).
- **Transformaciones:** Variables categóricas fueron codificadas con *One-Hot Encoding*, y las numéricas fueron normalizadas para mejorar el rendimiento de los modelos.
- **Simplificación de Clases:** La columna *Primary Type* se reorganizó para agrupar categorías similares, lo que permitió un enfoque más eficiente en la predicción.

3. Entrenamiento de Modelos:

- **Modelos Supervisados:**
    - **Gradient Boosting:** Obtuvo la mejor precisión (**60%**), destacando su capacidad para capturar patrones complejos en los datos.
    - **Redes Neuronales:** Mejoró significativamente con técnicas como *Dropout*, *Batch Normalization* y *Early Stopping*, pasando de una precisión inicial del **40% al 60%**.
    - **Árboles de Decisión y Random Forest:** Lograron precisiones moderadas (54%-59%), con limitaciones al clasificar clases minoritarias.
    - **KNN:** Mostró un desempeño adecuado (**56%**) en comparación con los árboles de decisión.
- **Modelos No Supervisados:**
    - **K-Means:** Identificó grupos basados en patrones espaciales y temporales, aunque el índice de Silhouette (**0.19**) indicó grupos poco definidos.

4. Desafíos y Limitaciones:

- **Desbalance de Clases:** Las categorías minoritarias presentaron dificultades en todos los modelos, debido al predominio de crímenes comunes como *Property Crime*.

- **Ausencia de Datos Complementarios:** Variables adicionales como datos socioeconómicos o índices de criminalidad recientes podrían haber enriquecido los resultados.
- **Escalabilidad:** Algunos modelos, como Random Forest y Redes Neuronales, requirieron altos recursos computacionales para entrenar.

5. Impacto del Proyecto:

Los resultados obtenidos proporcionan una base sólida para diseñar estrategias de prevención y optimización de recursos policiales en Chicago. Modelos como Gradient Boosting y Redes Neuronales demostraron ser útiles para identificar patrones en crímenes frecuentes. Sin embargo, futuras iteraciones deberán enfocarse en abordar las limitaciones actuales, particularmente en clases minoritarias y en la incorporación de atributos más representativos.

En resumen, este proyecto mostró el potencial de las técnicas de aprendizaje automático para enfrentar desafíos complejos en el análisis criminal, destacando la importancia de un enfoque integral en la preparación y análisis de datos.