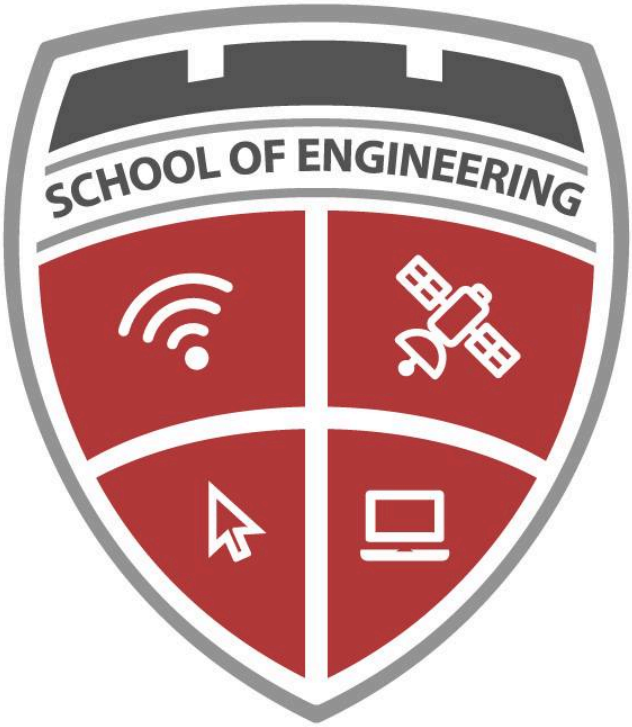
UNIVERSIDAD AUTÓNOMA GABRIEL RENÉ MORENO
SCHOOL OF ENGINEERING SOE - UNIDAD DE POSGRADO
MAESTRÍA EN CIENCIA DE DATOS E INTELIGENCIA ARTIFICIAL V1E2



MINERÍA DE DATOS APLICADA AL ANÁLISIS DE DELITOS EN LOS ÁNGELES (2020-2025)
TRABAJO FINAL - GRUPO 3

Participantes:
- Karen Torrico
- Karen Huacota
- Yesika Luna
- Elvis Miranda
- Ivan Mamani Condori

Docente: Msc. Renzo Claure

Módulo: DATA MINING

Agosto, 2025

ÍNDICE

I. INTRODUCCIÓN

El análisis del comportamiento delictivo en grandes ciudades se ha convertido en un área crucial para las políticas públicas y la gestión de la seguridad ciudadana. Comprender los patrones de ocurrencia delictiva no solo facilita la prevención y reducción de la criminalidad, sino que también optimiza la asignación de recursos policiales y mejora la confianza de la comunidad. Este proyecto se enfoca en la ciudad de Los Ángeles, uno de los centros urbanos más grandes de Estados Unidos, analizando los datos de incidentes delictivos reportados desde el 2020 hasta el 2025.

Para abordar este objetivo, se trabajó con una base de datos original muy amplia y compleja, con más de un millón de registros (1,004,991 filas) distribuidos en 28 columnas, en un archivo CSV comprimido. Las variables clave del dataset incluyen la fecha y hora del delito, la descripción y tipo de crimen, las características demográficas de las víctimas (edad, sexo y etnia), la ubicación geográfica precisa, y el tipo de arma utilizada cuando aplica. Esta riqueza y variedad de datos permiten un análisis multidimensional pero también presentan importantes desafíos de manejo y procesamiento.

Para superar estos retos, se aplicaron procesos rigurosos de limpieza y transformación, que incluyeron la eliminación de datos incompletos o inconsistentes, la normalización de variables categóricas, y la creación de nuevas variables temporales y geográficas que facilitan el análisis. Estas etapas son fundamentales para asegurar la calidad y consistencia de la información que será utilizada en los análisis estadísticos y modelado predictivo.

Los resultados obtenidos a través de técnicas de minería de datos, aprendizaje automático supervisado y no supervisado, no solo permiten responder preguntas cruciales sobre los patrones delictivos en Los Ángeles, sino que también proporcionan modelos con alta precisión para predecir la violencia en crímenes. Esto representa un avance significativo para el diseño de estrategias de combate a la delincuencia más efectivas y basadas en evidencia.

II. DEFINICIÓN DEL PROBLEMA

2.1. Objetivos

Analizar el comportamiento de los delitos reportados en la ciudad de Los Ángeles desde el año 2020, y desarrollar modelos predictivos que permitan anticipar si un crimen será violento o no, utilizando las variables disponibles en el dataset.

2.2. Preguntas Clave

Para cumplir con este objetivo, el análisis aborda las siguientes preguntas clave:

- ¿Cuáles son los tipos de delitos más frecuentes en Los Ángeles durante el periodo estudiado?
- ¿Existen patrones temporales (por hora, día, mes) o geográficos (por zonas o áreas) en la ocurrencia de crímenes violentos?
- ¿Es posible predecir con precisión si un crimen es violento basándonos en variables como la ubicación, la hora y el tipo de delito?

Estas interrogantes guían el proceso de limpieza, transformación, análisis y modelado, orientando la selección de técnicas y herramientas utilizadas a lo largo del proyecto.

2.3. Descripción del Dataset

El análisis se fundamenta en el dataset oficial "Crime Data from 2020 to Present" proporcionado por la ciudad de Los Ángeles, disponible públicamente en https://catalog.data.gov/dataset/crime-data-from-2020-to-present. Este conjunto de datos contiene aproximadamente 1,004,991 registros de incidentes delictivos reportados desde el año 2020 hasta 2025, ofreciendo una cobertura amplia y actualizada del fenómeno delictivo en la ciudad.

El dataset incluye cerca de 28 variables que brindan información detallada y diversa. Entre las más relevantes se encuentran:

- Fechas y horas relacionadas con el delito, como la fecha de ocurrencia (`date_occ`), fecha de reporte (`date_rptd`) y hora (`time_occ`).

- Variables de ubicación y área, como el nombre del área (`area_name`), código del barrio (`area_id`), coordenadas geográficas de latitud y longitud (`latitude`, `longitude`).
- Detalle del tipo de crimen, incluyendo el código del crimen (`crm_cd`), descripción del código (`crm_cd_desc`) y el motivo aproximado (`premis_desc`).
- Características demográficas de las víctimas, como edad (`vict_age`), sexo (`vict_sex`) y raza (`vict_descent`).
- Información relacionada con armas usadas en el delito (`weapon_used_cd`, `weapon_desc`).
- Estado del caso o resultado (`status`) y otras variables codificadas para seguimiento del incidente

Este dataset, actualizado periódicamente bajo licencia pública, requiere procesos rigurosos de limpieza y transformación para asegurar su consistencia y calidad, debido a su tamaño y heterogeneidad. Su riqueza en detalles temporales, espaciales y delictivos lo convierte en una fuente invaluable para análisis estadísticos y predictivos del crimen en Los Ángeles.

III. PROBLEMAS RESUELTOS

Una etapa fundamental en cualquier proyecto de minería de datos es la preparación y saneamiento de los datos originales. En este proyecto, dada la magnitud y complejidad del dataset de delitos en Los Ángeles, se aplicaron rigurosos procesos de limpieza y transformación para asegurar la calidad, integridad y consistencia de la información. Estas acciones buscan eliminar registros erróneos, incompletos o inconsistentes, normalizar variables para facilitar el análisis, y crear nuevas características relevantes que potencien el modelado predictivo.

3.1. Carga y Exploración Inicial

El análisis comenzó cargando un gran dataset con más de un millón de registros y 28 columnas, representando los crímenes reportados desde 2020. Era necesario conocer la estructura y calidad inicial de la información para planificar las etapas de limpieza.

```
>> Cargando dataset desde GitHub (gzip)...
Dataset cargado con 1,004,991 filas y 28 columnas
```

	DR_NO	Date Rptd	DATE OCC	TIME OCC	AREA	AREA NAME	Rpt Dist No	Part 1-2	Crm Cd	Crm Cd Desc	...	Status	Status Desc	Crm Cd 1	Crm Cd 2
0	211507896	04/11/2021 12:00:00 AM	11/07/2020 12:00:00 AM	845	15	N Hollywood	1502	2	354	THEFT OF IDENTITY	...	IC	Invest Cont	354.0	NaN
1	201516622	10/21/2020 12:00:00 AM	10/18/2020 12:00:00 AM	1845	15	N Hollywood	1521	1	230	ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT	...	IC	Invest Cont	230.0	NaN
2	240913563	12/10/2024 12:00:00 AM	10/30/2020 12:00:00 AM	1240	9	Van Nuys	933	2	354	THEFT OF IDENTITY	...	IC	Invest Cont	354.0	NaN
3	210704711	12/24/2020 12:00:00 AM	12/24/2020 12:00:00 AM	1310	7	Wilshire	782	1	331	THEFT FROM MOTOR VEHICLE - GRAND ($950.01 AND ...	...	IC	Invest Cont	331.0	NaN
4	201418201	10/03/2020 12:00:00 AM	09/29/2020 12:00:00 AM	1830	14	Pacific	1454	1	420	THEFT FROM MOTOR VEHICLE - PETTY ($950 & UNDER)	...	IC	Invest Cont	420.0	NaN

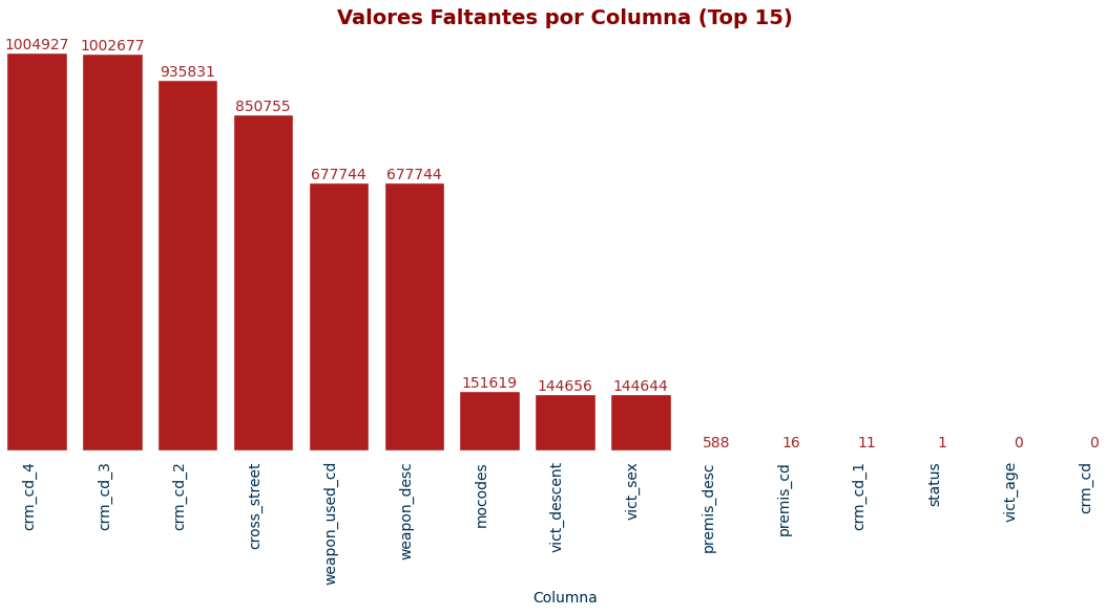
5 rows × 28 columns

El dataset presenta una gran variedad de variables, muchas relacionadas con ubicación, tiempo y características individuales, pero todavía no está limpio ni preparado para análisis directo.

3.2. Valores Faltantes

Se identificó la cantidad de valores nulos en cada variable para evaluar la calidad y definir los filtros.

Gráfico de valores nulos por columna (Top 15)



Interpretación:

El gráfico de valores nulos revela que varias columnas presentan una cantidad significativa de datos faltantes, lo que justifica la eliminación o imputación cuidadosa para evitar sesgos o errores. Algunas columnas como las de armas usadas y descripciones presentan muchos valores faltantes, lo que indica la necesidad de un filtrado para evitar sesgos o errores en el análisis. Variables críticas con datos faltantes pueden impedir un modelado confiable.

3.3. Transformación de Datos

Esta etapa fue clave para transformar los datos crudos en formatos útiles e incorporar nuevas variables relevantes para el estudio.

3.3.1. Parseo de Fechas Frecuentes

Las columnas con fechas estaban en formato no estándar o de texto. Se convirtieron a formato datetime.

```
ANTES (date_occ):
0    11/07/2020 12:00:00 AM
1    10/18/2020 12:00:00 AM
2    10/30/2020 12:00:00 AM
3    12/24/2020 12:00:00 AM
4    09/29/2020 12:00:00 AM
Name: date_occ, dtype: object

DESPUÉS (date_occ):
0   2020-11-07
1   2020-10-18
2   2020-10-30
3   2020-12-24
4   2020-09-29
Name: date_occ, dtype: datetime64[ns]
Columnas de fecha parseadas: ['date_occ', 'date_rptd']
```

La conversión permite extraer información temporal con precisión, clave para análisis cronológicos y modelado.

3.3.2. Parseo de Hora

Las columnas de tiempo contenían formatos heterogéneos. Se normalizó y extrajo la hora, creando la variable numérica hour_occ.

```
ANTES (time_occ):
0     845
1    1845
2    1240
3    1310
4    1830
Name: time_occ, dtype: int64

DESPUÉS (hour_occ):
   hour_occ
0         8
1        18
2        12
3        13
4        18
```

La variable hour_occ facilita el análisis del comportamiento delictivo según la hora del día, que puede variar significativamente.

3.3.3. Selección de Fecha Base

Se determinó qué columna de fecha sería la referencia para análisis temporales y creación de nuevas variables.

```
Fecha Base seleccionada: date_occ
```

Consistencia en análisis temporal, asegurando que todas las variables temporales se basen en un mismo punto de referencia.

3.3.4. Normalización de Variables Temporales

De la fecha base se extrajeron las variables: year, month, dayofweek y date_month.

```
ANTES: columnas
Index(['dr_no', 'date_rptd', 'date_occ', 'time_occ', 'area', 'area_name',
       'rpt_dist_no', 'part_1_2', 'crm_cd', 'crm_cd_desc', 'mocodes',
       'vict_age', 'vict_sex', 'vict_descent', 'premis_cd', 'premis_desc',
       'weapon_used_cd', 'weapon_desc', 'status', 'status_desc', 'crm_cd_1',
       'crm_cd_2', 'crm_cd_3', 'crm_cd_4', 'location', 'cross_street', 'lat',
       'lon', 'hour_occ'],
      dtype='object')

DESPUÉS: preview de nuevas columnas
   year  month  dayofweek date_month
0  2020     11          5 2020-11-01
1  2020     10          6 2020-10-01
2  2020     10          4 2020-10-01
3  2020     12          3 2020-12-01
4  2020      9          1 2020-09-01
```

Estas variables permiten análisis detallados en diferentes escalas temporales para entender cuándo se concentran o varían los delitos.

3.4. Limpieza de Datos

Durante la limpieza inicial, se detectaron varias fuentes de inconsistencias y ruido en la base de datos.

3.4.1. Limpieza de Geocoordenadas

Se buscaron y limpiaron las columnas de latitud y longitud, corrigiendo valores inválidos y creando indicador has_geo.

```
ANTES: Preview coordenadas
       lat       lon
0  34.2124 -118.4092
1  34.1993 -118.4203
2  34.1847 -118.4509
3  34.0339 -118.3747
4  33.9813 -118.4350

DESPUÉS: Preview coordenadas limpias + indicador
       lat       lon  has_geo
0  34.2124 -118.4092     True
1  34.1993 -118.4203     True
2  34.1847 -118.4509     True
3  34.0339 -118.3747     True
4  33.9813 -118.4350     True
```

Esto garantiza que los análisis geoespaciales utilicen únicamente datos confiables y relevantes para la ciudad de Los Ángeles.

3.4.2. Limpieza y Unificación de Variables Categóricas

Se normalizaron y unificaron columnas como premis_desc, area_name y crm_cd_desc para reducir inconsistencias y variabilidad innecesaria.

```
ANTES: columnas categóricas
['area', 'area_name', 'crm_cd', 'crm_cd_desc', 'premis_cd', 'premis_desc', 'crm_cd_1', 'crm_cd_2', 'crm_cd_3',
'crm_cd_4']

DESPUÉS: columnas categóricas unificadas
             premis_desc    area_name  \
0  SINGLE FAMILY DWELLING  N Hollywood
1               SIDEWALK  N Hollywood
2  SINGLE FAMILY DWELLING     Van Nuys
3                 STREET     Wilshire
4                  ALLEY      Pacific

                                    crm_cd_desc
0                                THEFT OF IDENTITY
1    ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT
2                                THEFT OF IDENTITY
3  THEFT FROM MOTOR VEHICLE - GRAND ($950.01 AND ...
4    THEFT FROM MOTOR VEHICLE - PETTY ($950 & UNDER)
```

Esto permite análisis más coherentes y evita que diferentes nombres para una misma categoría compliquen o distorsionen el modelado.

3.4.3. Creación de la Variable Objetivo violent_crime

A partir de las descripciones de crimen, se creó la variable binaria indicando si el delito fue violento.

```
ANTES: crm_cd_desc ejemplo
0                                   THEFT OF IDENTITY
1       ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT
2                                   THEFT OF IDENTITY
3    THEFT FROM MOTOR VEHICLE - GRAND ($950.01 AND ...
4      THEFT FROM MOTOR VEHICLE - PETTY ($950 & UNDER)
Name: crm_cd_desc, dtype: object

DESPUÉS: variable target 'violent_crime'
                                        crm_cd_desc  violent_crime
0                                 THEFT OF IDENTITY              0
1     ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT              1
2                                 THEFT OF IDENTITY              0
3  THEFT FROM MOTOR VEHICLE - GRAND ($950.01 AND ...              0
4    THEFT FROM MOTOR VEHICLE - PETTY ($950 & UNDER)              0
```

Esto define claramente el objetivo de clasificación, separando delitos violentos y no violentos para facilitar el análisis y posterior modelado.

3.4.4. Limpieza Final de Registros

Fue necesario eliminar registros duplicados y aquellos con datos faltantes en variables críticas como la descripción del delito, fecha o ubicación. Se aplicó un filtro específico para mantener únicamente registros con edades válidas, considerando víctimas entre 1 y 120 años, descartando valores atípicos o erróneos.

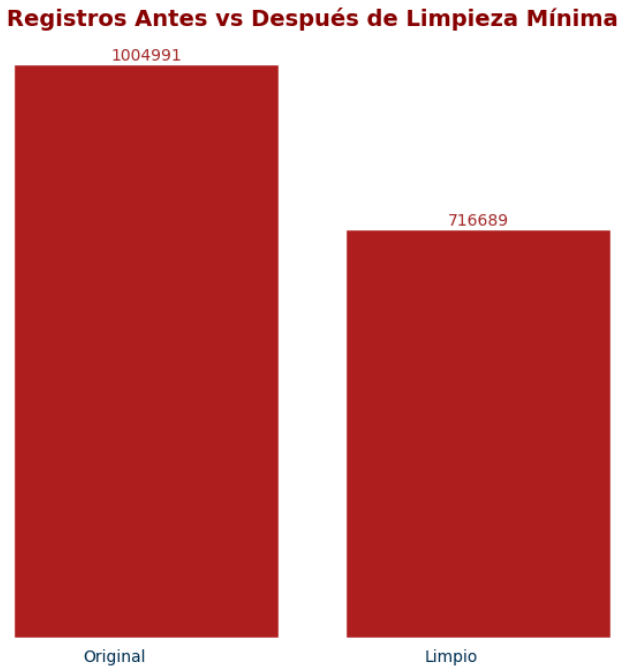
Además, se validó y estandarizó la información demográfica, incluyendo únicamente casos con datos confiables de sexo (M, F, X) y grupo racial sin valores nulos o genéricos que no aportaran información para el análisis. Asimismo, la limpieza de las coordenadas geográficas fue crucial para asegurar que solo se incluyeran registros con ubicaciones válidas dentro del rango esperado para Los Ángeles, eliminando valores nulos, ceros o fuera de límites geográficos razonables.

Por último, se aplicó un filtro temporal que excluyó registros con fechas inválidas o anteriores al año 2000, garantizando la relevancia y actualidad de los datos.

```
--- Limpieza: Columnas mínimas ---
Antes: 1004991 | Después: 1004991 | Eliminados: 0

--- Limpieza: Edades válidas (1-120) ---
Antes: 1004991 | Después: 735632 | Eliminados: 269359

--- Limpieza: Sexo válido ---
Antes: 735632 | Después: 735488 | Eliminados: 144

--- Limpieza: Raza válida ---
Antes: 735488 | Después: 718348 | Eliminados: 17140

--- Limpieza: Coordenadas válidas ---
Antes: 718348 | Después: 716689 | Eliminados: 1659

--- Limpieza: Fechas válidas ---
Antes: 716689 | Después: 716689 | Eliminados: 0
```

Este proceso rigoroso mejora la calidad y confianza en los datos, reduciendo ruido y errores que afectarían análisis posteriores.

Tras estos procesos de limpieza, el dataset se redujo significativamente de 1,004,991 registros originales a 716,689 registros limpios y confiables, representando una eliminación de aproximadamente el 28% de los datos.



La grafica refleja visualmente la reducción del volumen de datos. En conjunto con las acciones anteriores, aseguran que el análisis posterior se realice sobre una base sólida y representativa, mejorando la validez de conclusiones y modelos predictivos.

IV. MÉTODOS UTILIZADOS

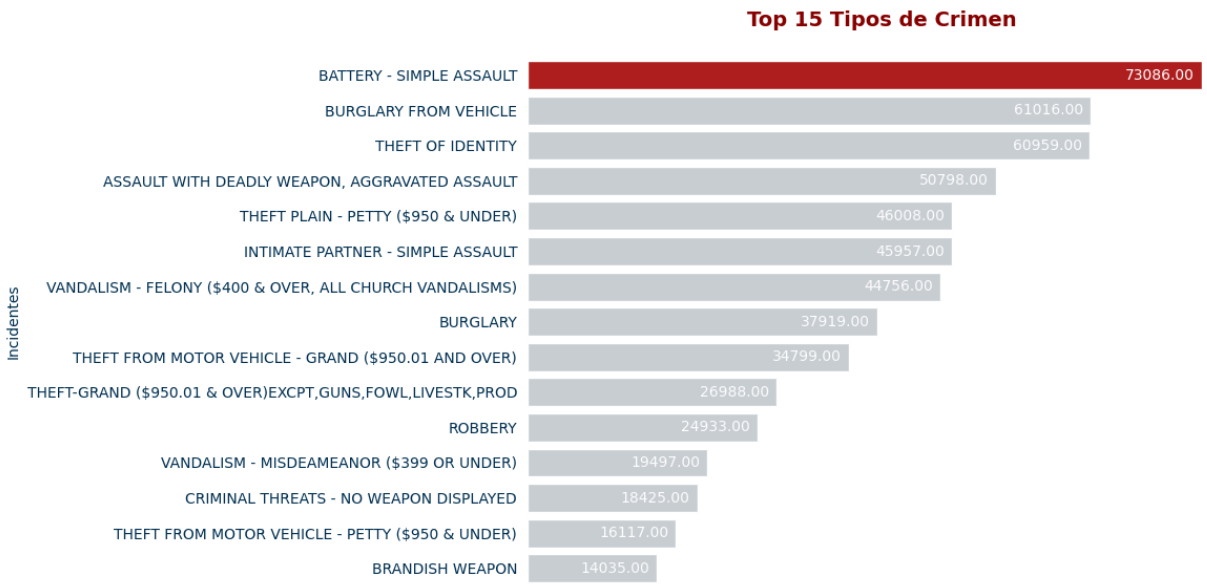
En esta sección se presentan los métodos aplicados para explorar y analizar el dataset limpio, desde análisis descriptivo hasta técnicas estadísticas y de machine learning supervisado y no supervisado. Cada paso se acompaña de visualizaciones y métricas que permiten validar los modelos y entender mejor la dinámica de los delitos en Los Ángeles.

4.1. Análisis Descriptivo

El Análisis Exploratorio de Datos (EDA) proporciona una visión inicial de las características más importantes del dataset, permitiendo identificar patrones, frecuencias y relaciones clave.

4.1.1. Top 15 Tipos de Crimen

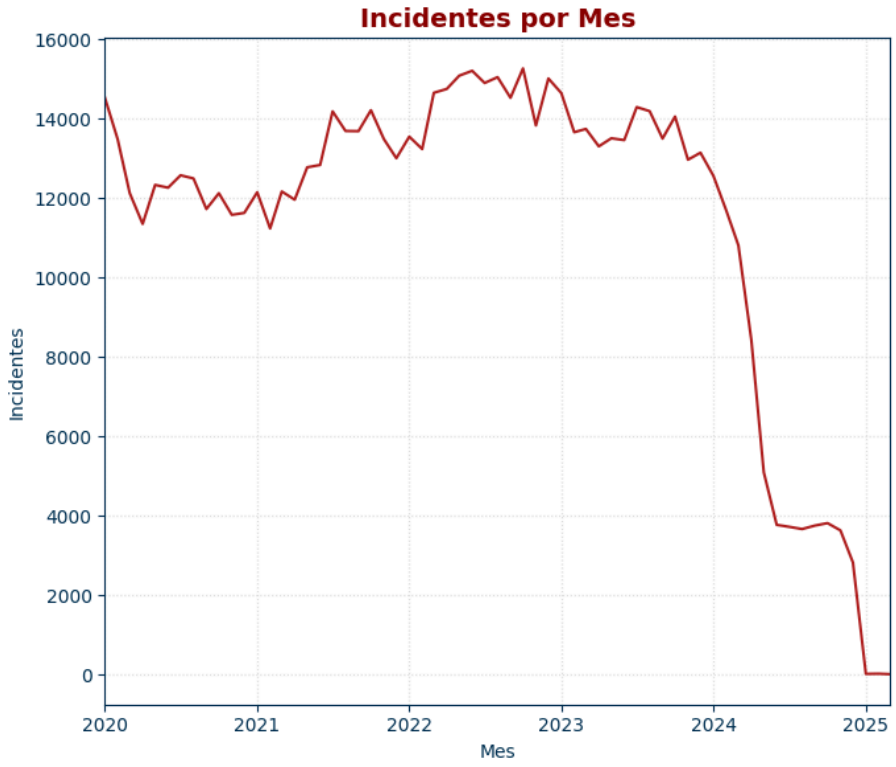
Se obtuvieron y visualizaron los 15 tipos de delito más frecuentes para comprender cuáles predominan en la ciudad.



Los datos indican que el delito más común es “Battery - Simple Assault” seguido por robo y otros tipos de agresiones, resaltando áreas prioritarias para políticas de seguridad.

4.1.2. Tendencia Mensual de Incidentes

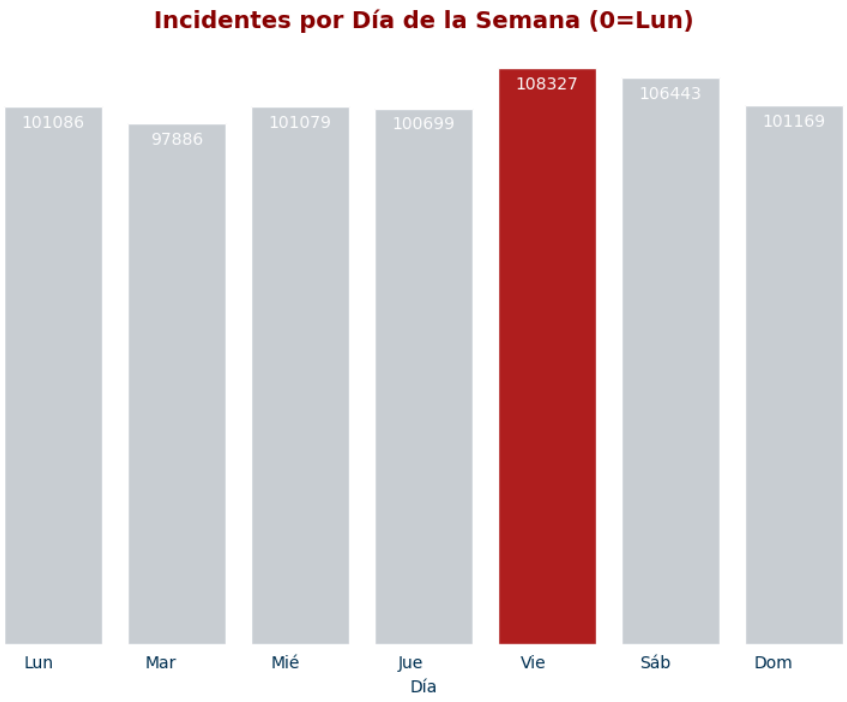
Se analizó la evolución del número de incidentes por mes a lo largo del tiempo.



Se observan fluctuaciones temporales con periodos de mayor y menor actividad delictiva, información valiosa para anticipar demandas de recursos policiales.

4.1.3. Distribución por Día de la Semana

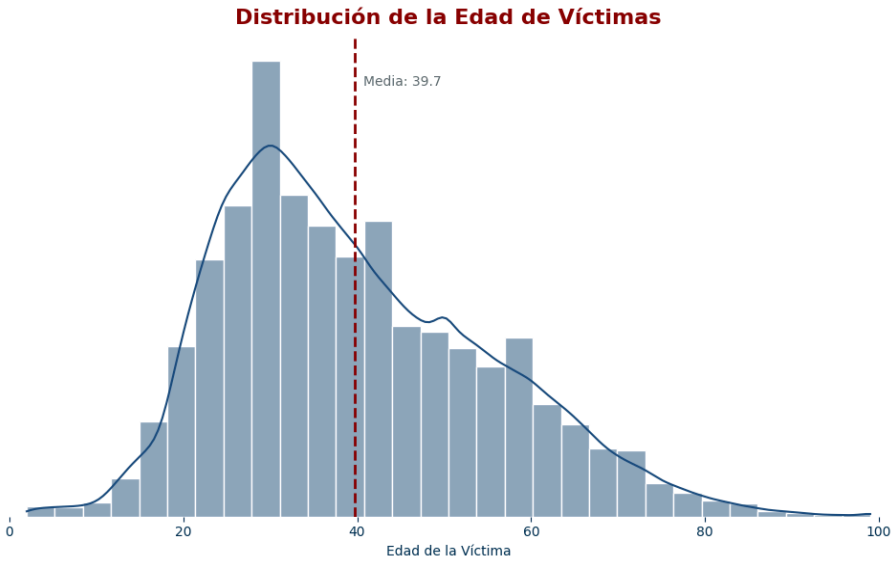
Se determinó cómo se distribuyen los incidentes según el día de la semana.



La actividad delictiva presenta ligeras variaciones semanales, con ciertos días acumulando más reportes, lo que puede guiar la planificación de turnos y patrullajes en especial los dias viernes.

4.1.4. Histograma de Edad de Víctima

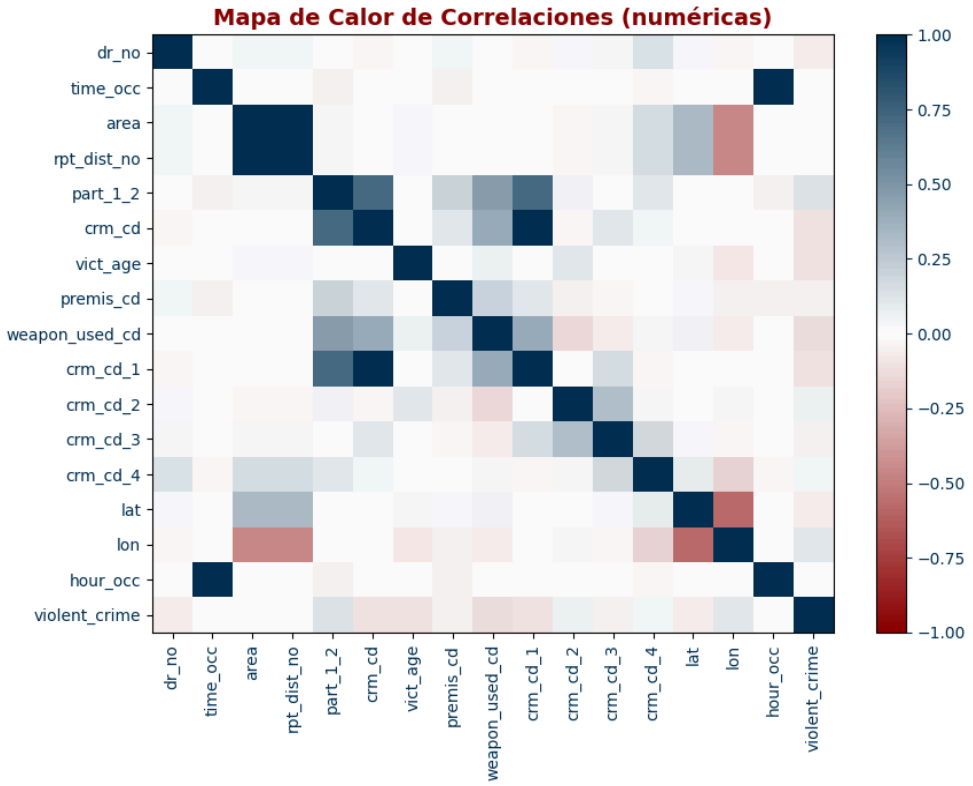
Se estudió la distribución de la edad de las víctimas.



La edad promedio se sitúa cercana a 40 años, mostrando que la población adulta es mayormente afectada.

4.1.5. Correlaciones Numéricas

Se realizó un mapa de calor para examinar las correlaciones entre variables numéricas.



Se destacan correlaciones positivas y negativas pertinentes, apoyando la selección de variables para modelos predictivos y evitando multicolinealidad.

4.2. Modelado Estadístico

Se aplicaron modelos estadísticos para entender y predecir el comportamiento delictivo, integrando tanto regresión lineal como clasificación logística, junto con técnicas de machine learning.

4.2.1. Regresión Lineal OLS sobre Incidencias Mensuales

Se ajustó un modelo de regresión lineal para explicar las variaciones mensuales en la cantidad de delitos.

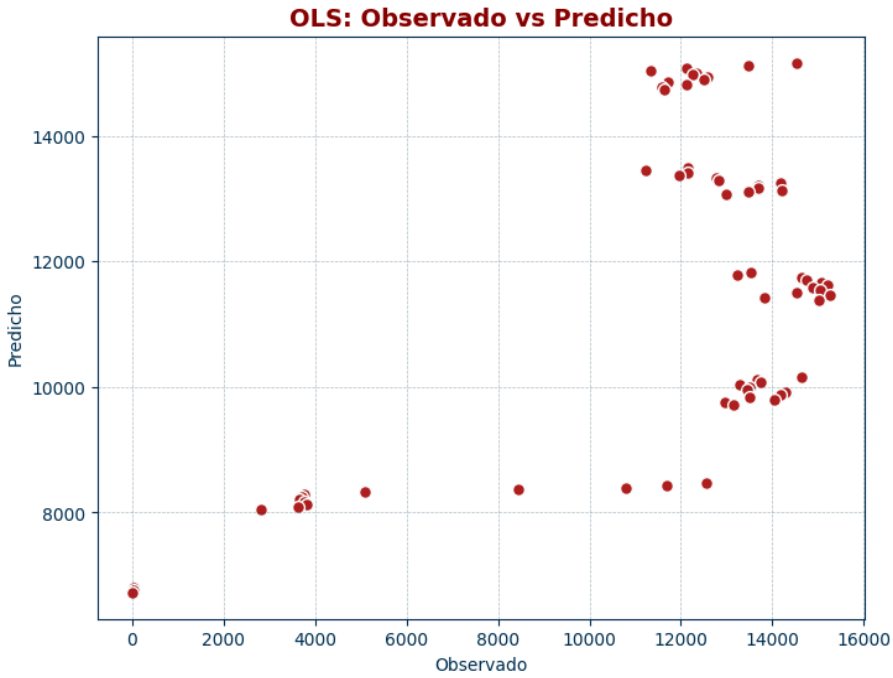
```
[OLS] Resumen:
                            OLS Regression Results
==============================================================================
Dep. Variable:              incidents   R-squared:                       0.361
Model:                            OLS   Adj. R-squared:                  0.340
Method:                 Least Squares   F-statistic:                     16.94
Date:                Fri, 22 Aug 2025   Prob (F-statistic):           1.47e-06
Time:                        03:02:23   Log-Likelihood:                -601.05
No. Observations:                  63   AIC:                             1208.
Df Residuals:                      60   BIC:                             1215.
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       3.391e+06   5.82e+05      5.829      0.000    2.23e+06    4.55e+06
year       -1671.1694    287.633     -5.810      0.000   -2246.520   -1095.818
month        -39.2029    124.732     -0.314      0.754    -288.704     210.299
==============================================================================
Omnibus:                       17.874   Durbin-Watson:                   0.072
Prob(Omnibus):                  0.000   Jarque-Bera (JB):                4.821
Skew:                          -0.316   Prob(JB):                       0.0898
Kurtosis:                       1.801   Cond. No.                     2.71e+06
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.71e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
```

El modelo de regresión lineal muestra que aproximadamente ***el 36% de la variabilidad en los incidentes (R^2 = 0.361) se explica por las variables año y mes***. El coeficiente para año (-1671.17, p < 0.001) es estadísticamente significativo, lo que indica que, ***en promedio, los incidentes disminuyen en 1671 casos por cada incremento de un año***, manteniendo constante el mes. En contraste, el efecto del mes (-39.20, p = 0.754) no resulta significativo, sugiriendo que ***las variaciones mensuales no influyen de manera relevante en el número de incidentes***. El modelo, aunque ***significativo en conjunto*** (F = 16.94, p < 0.001), presenta un valor bajo en la prueba Durbin-Watson (0.072), lo que ***sugiere la presencia de autocorrelación en los residuos***. Además, el alto número de condición indica posibles problemas de multicolinealidad o inestabilidad numérica. En síntesis,

los resultados evidencian una tendencia decreciente de los incidentes a lo largo de los años, pero no se observa un patrón claro por meses.

4.2.2. Ajustes del Modelo OLS

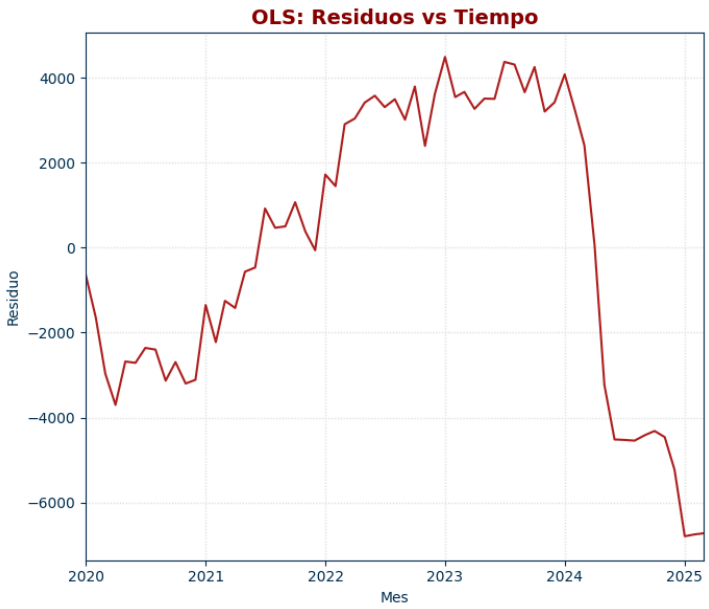
Se evaluó el ajuste del modelo con un gráfico de observado vs predicho.



La dispersión indica que el modelo predice razonablemente bien las tendencias generales, aunque ciertos puntos muestran desviaciones.

4.2.3. Residuos del Modelo OLS

Se graficaron los residuos a lo largo del tiempo.



Los residuos muestran patrones que sugieren la presencia de variabilidad no explicada, indicando limitaciones del modelo lineal simple.

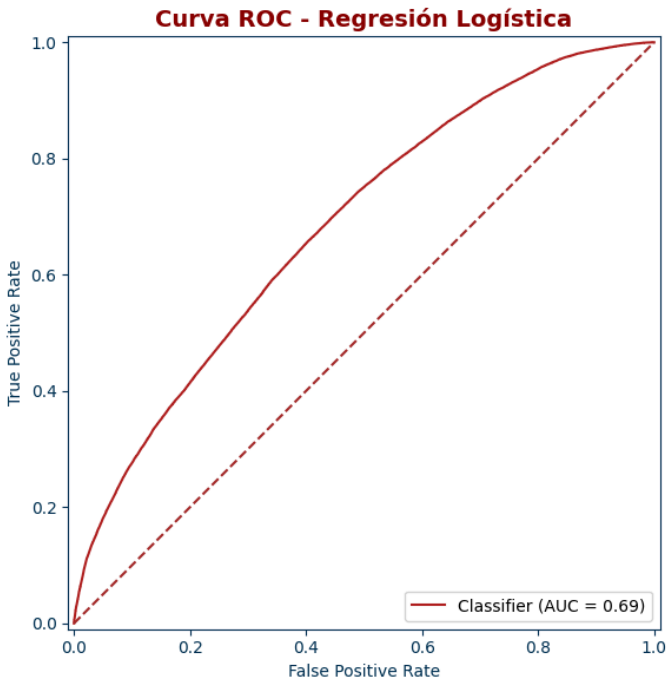
4.2.4. Regresión Logística para Odds Ratios

Se utilizó regresión logística para modelar la probabilidad de que un delito sea violento, con interpretación de coeficientes y odds ratios.

```
[Logit] Coeficientes y OR:
                                                    coef        OR  \
premis_desc_top_SIDEWALK                        0.816005  2.261447
const                                           0.402085  1.494938
area_name_top_Southeast                         0.103444  1.108984
dayofweek                                       0.037483  1.038194
hour_occ                                        0.007551  1.007579
month                                          -0.001302  0.998699
premis_desc_top_PARK/PLAYGROUND                -0.005294  0.994720
premis_desc_top_HOTEL                          -0.007476  0.992552
premis_desc_top_GAS STATION                    -0.015569  0.984551
area_name_top_Newton                           -0.168000  0.845354
premis_desc_top_RESTAURANT/FAST FOOD           -0.318516  0.727227
area_name_top_Rampart                          -0.336758  0.714082
area_name_top_Other                            -0.375281  0.687096
area_name_top_Southwest                        -0.436110  0.646547
area_name_top_Mission                          -0.515067  0.597461
premis_desc_top_MULTI-UNIT DWELLING (APARTMENT,... -0.521508  0.593625
area_name_top_Olympic                          -0.522139  0.593250
premis_desc_top_STREET                         -0.542824  0.581105
premis_desc_top_OTHER BUSINESS                 -0.628441  0.533423
premis_desc_top_Other                          -0.657661  0.518062
```

	pvalue
premis_desc_top_SIDEWALK	7.059446e-169
const	3.421455e-41
area_name_top_Southeast	2.080242e-12
dayofweek	2.428328e-173
hour_occ	4.200617e-76
month	8.932637e-02
premis_desc_top_PARK/PLAYGROUND	8.910349e-01
premis_desc_top_HOTEL	8.539403e-01
premis_desc_top_GAS STATION	7.201576e-01
area_name_top_Newton	4.337834e-28
premis_desc_top_RESTAURANT/FAST FOOD	1.506591e-17
area_name_top_Rampart	8.532860e-104
area_name_top_Other	2.425636e-195
area_name_top_Southwest	3.156892e-198
area_name_top_Mission	3.139760e-224
premis_desc_top_MULTI-UNIT DWELLING (APARTMENT,...	4.031200e-78
area_name_top_Olympic	2.981539e-254
premis_desc_top_STREET	2.608526e-85
premis_desc_top_OTHER BUSINESS	6.425413e-91
premis_desc_top_Other	2.525256e-120

El modelo de regresión logística revela que algunos factores del entorno físico y geográfico influyen significativamente en la probabilidad de que un crimen sea violento. Por ejemplo, el hecho de que un incidente ocurra en una ***banqueta (SIDEWALK)*** aumenta notablemente la probabilidad de violencia (OR = 2.26, p < 0.001), mientras que lugares como ***restaurantes/fast food (OR = 0.72)***, ***edificios de apartamentos (OR = 0.59)*** o en la ***calle en general (OR = 0.58)*** disminuyen la probabilidad de violencia de manera estadísticamente significativa. También se observan diferencias claras entre áreas: algunas como **Mission (OR = 0.59)** y **Olympic (OR = 0.59)** presentan menores probabilidades de violencia en comparación con la categoría de referencia. Además, variables temporales como el **día de la semana (OR = 1.038, p < 0.001)** y la **hora de ocurrencia (OR = 1.008, p < 0.001)** muestran que tanto el momento del día como el día específico influyen de manera positiva en la probabilidad de violencia, mientras que el ***mes no tiene un efecto relevante*** (p = 0.089). En conjunto, ***el modelo sugiere que tanto la ubicación física específica como la dimensión temporal son determinantes clave en la predicción de delitos violentos***.



El área bajo la curva (AUC) indica un ***desempeño moderado del modelo*** para clasificar crímenes violentos.

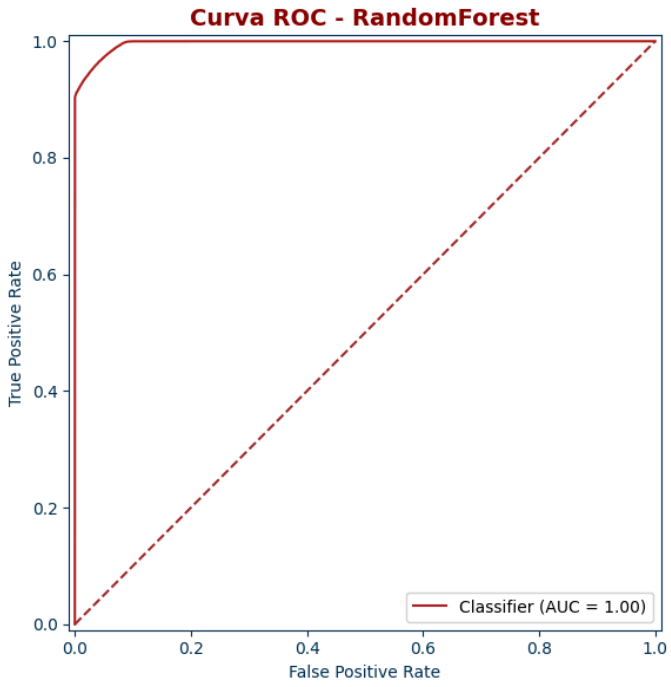
4.2.5. Machine Learning: Random Forest

Se entrenó un modelo Random Forest con pipeline de procesamiento para clasificación.

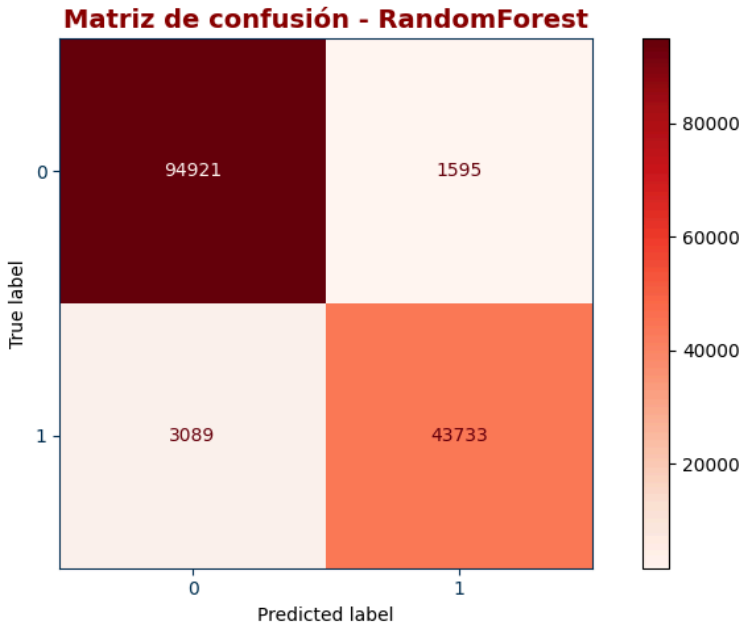
```
[RandomForest] Métricas:
{
  "accuracy": 0.9673219941676318,
  "precision": 0.9648120367102012,
  "recall": 0.9340267395668703,
  "f1": 0.9491698317959848,
  "roc_auc": 0.9966567254312784
}
```

El modelo de ***Random Forest muestra un desempeño sobresaliente en la clasificación de delitos violentos frente a no violentos***. La **exactitud (accuracy) del 96.7%** indica que el modelo predice correctamente la gran mayoría de los casos. La **precisión (96.4%)** revela que, cuando el modelo predice un crimen violento, casi siempre acierta, mientras que el **recall (93.4%)** muestra que también logra identificar la mayoría de los casos violentos reales, aunque se le escapan algunos. El **F1-score (94.9%)** confirma un

equilibrio sólido entre precisión y recall. Finalmente, el **AUC-ROC (0.997)** refleja una capacidad casi perfecta para distinguir entre delitos violentos y no violentos, lo que sugiere que ***el modelo es altamente confiable y robusto para este tipo de predicción***.



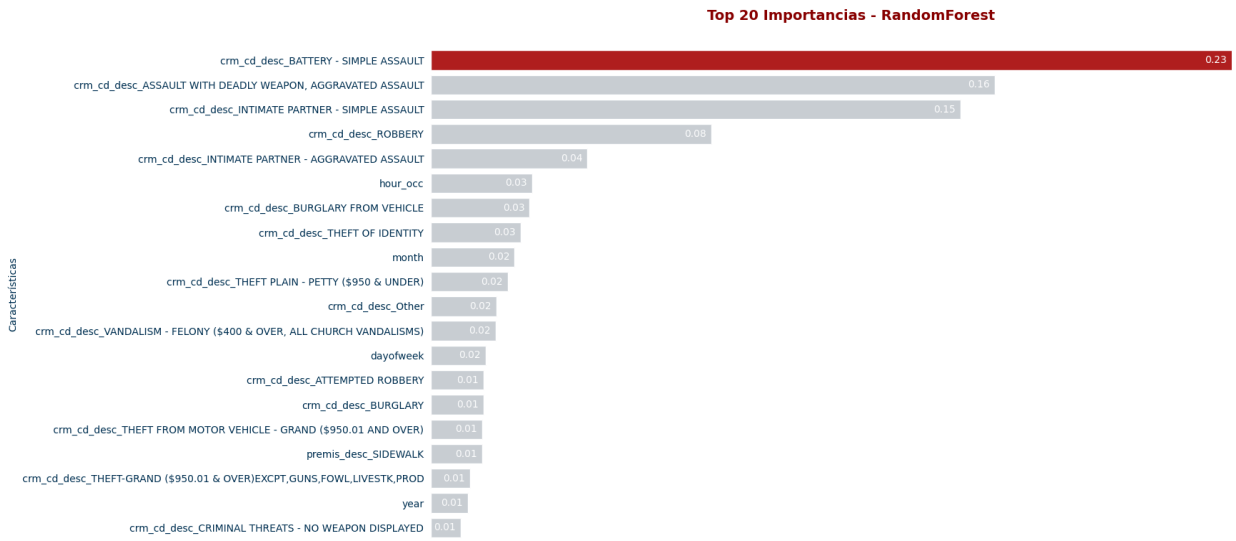
La curva destaca una ***casi perfecta discriminación entre clases***, lo que respalda su aplicación práctica.



Reflexiona la alta tasa de clasificación correcta y bajo error, ***demostrando la confiabilidad del modelo***.

4.2.6. Importancias de Variables

Se identificaron las variables más influyentes en la clasificación.



Las variables temporales y geográficas dominan el impacto, reforzando hallazgos previos sobre la importancia del tiempo y lugar en la violencia.

4.3. Modelos No Supervisados: PCA y KMeans

Se aplicó reducción dimensional mediante PCA y agrupamiento con KMeans para explorar patrones no etiquetados.

4.3.1. PCA (Principal Component Analysis)

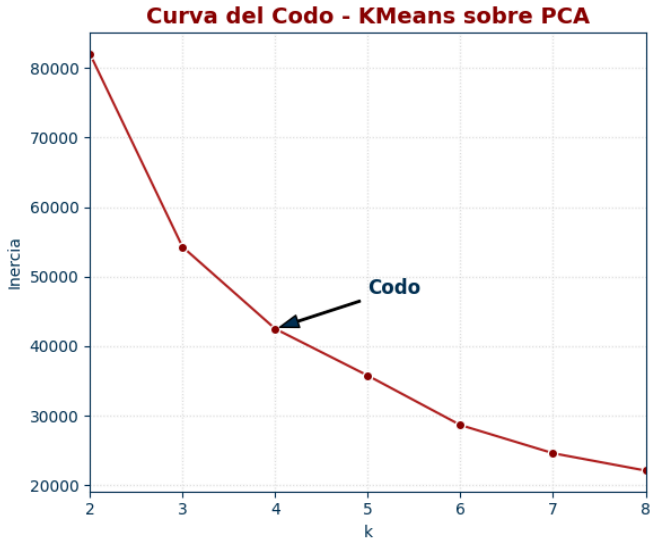
Se extrajeron dos componentes principales.

```
PCA varianza explicada (2 comp): 0.5212249859152711
```

Más del 50% de la variabilidad está capturada en estos componentes, facilitando análisis visual.

4.3.2. Cálculo del Número Óptimo de Clusters

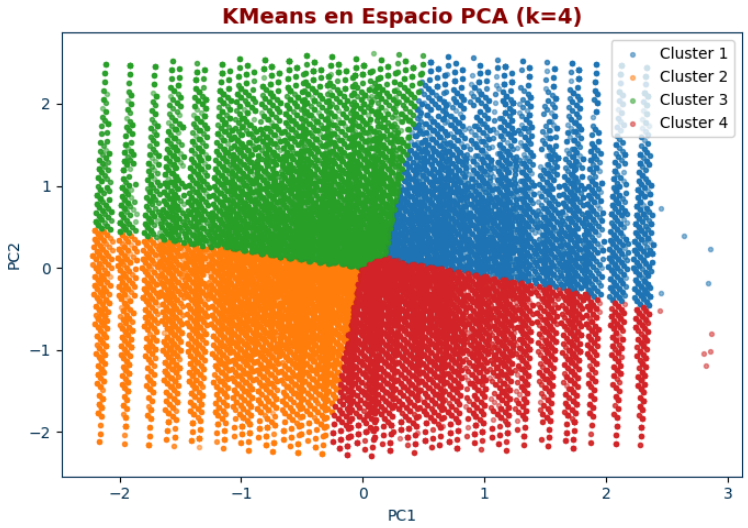
Se usó la curva del codo para determinar el número óptimo de clusters.



Esto sugiere que dividir los datos en 4 grupos balancea la complejidad y ajuste.

4.3.3. KMeans

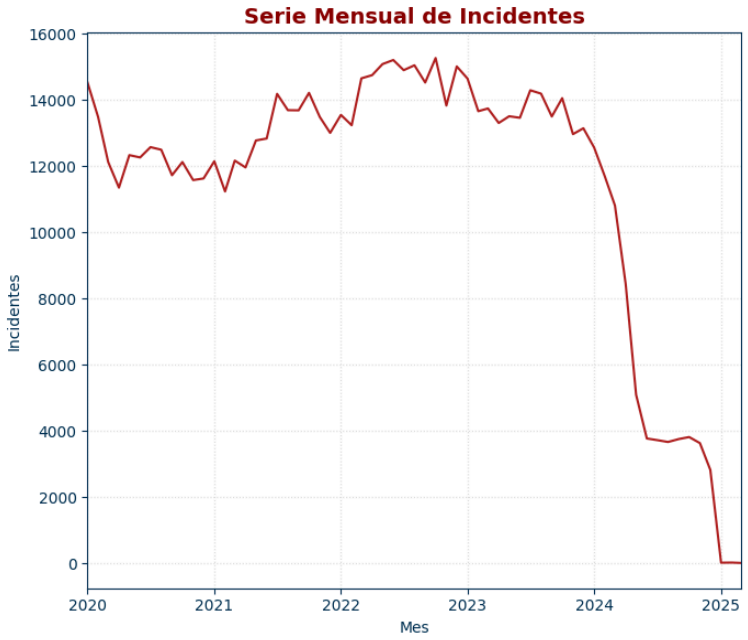
Se aplicó el algoritmo de clustering KMeans sobre el espacio reducido por PCA para identificar grupos naturales en los datos sin utilizar etiquetas previas. La reducción dimensional con PCA a dos componentes principales, que explica aproximadamente el 52% de la varianza, facilita una visualización clara de estos clusters.



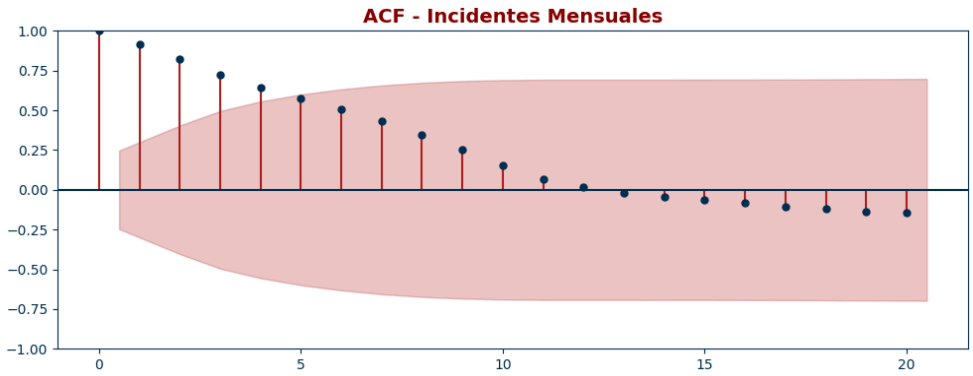
La visualización evidencia cuatro grupos claramente diferenciados que representan subconjuntos de incidentes con características similares, posiblemente relacionadas con patrones temporales y espaciales. ***Estos clusters sugieren que existen segmentos particulares dentro del fenómeno delictivo, como delitos que ocurren en horarios o áreas específicas, o que comparten características temporales similares***. Este análisis no supervisado es útil para descubrir estructuras latentes, orientar estudios más detallados y diseñar intervenciones específicas para cada grupo identificado, optimizando estrategias de prevención y control.

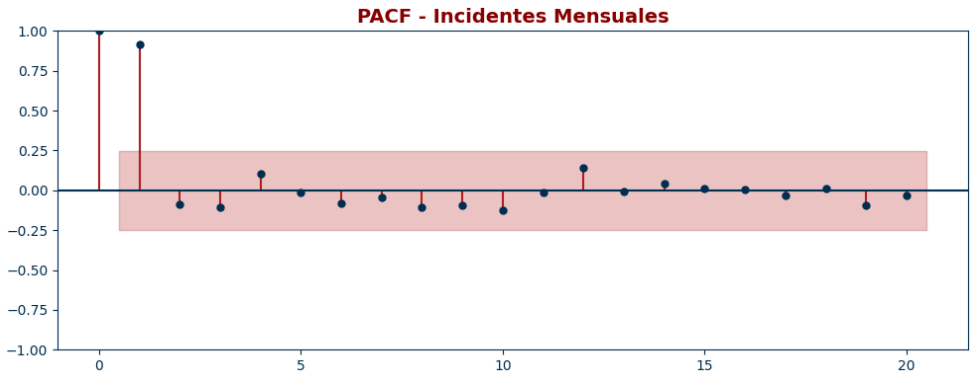
4.3.4. Series de Tiempo

Se analizó la evolución temporal de los delitos mediante la construcción de una serie temporal mensual agregada y la aplicación de funciones de autocorrelación (ACF) y autocorrelación parcial (PACF), herramientas esenciales para comprender dependencias temporales y estacionales.



Este gráfico ilustra la dinámica de los incidentes delictivos a lo largo del tiempo, mostrando variaciones, tendencias y posibles picos estacionales. Esta representación permite identificar períodos de incremento o disminución del delito, apoyando la toma de decisiones para asignar recursos en momentos críticos.





Los gráficos de ACF y PACF revelan patrones de dependencia temporal: el ACF muestra la correlación entre los valores de la serie en diferentes lags, mientras que el PACF ayuda a determinar la influencia directa de un lag específico. Estos patrones indican que ***la ocurrencia delictiva en un mes depende significativamente de los meses anteriores***, con posible presencia de estacionalidad o ciclos, información clave para la construcción de modelos predictivos de series de tiempo y pronósticos futuros.

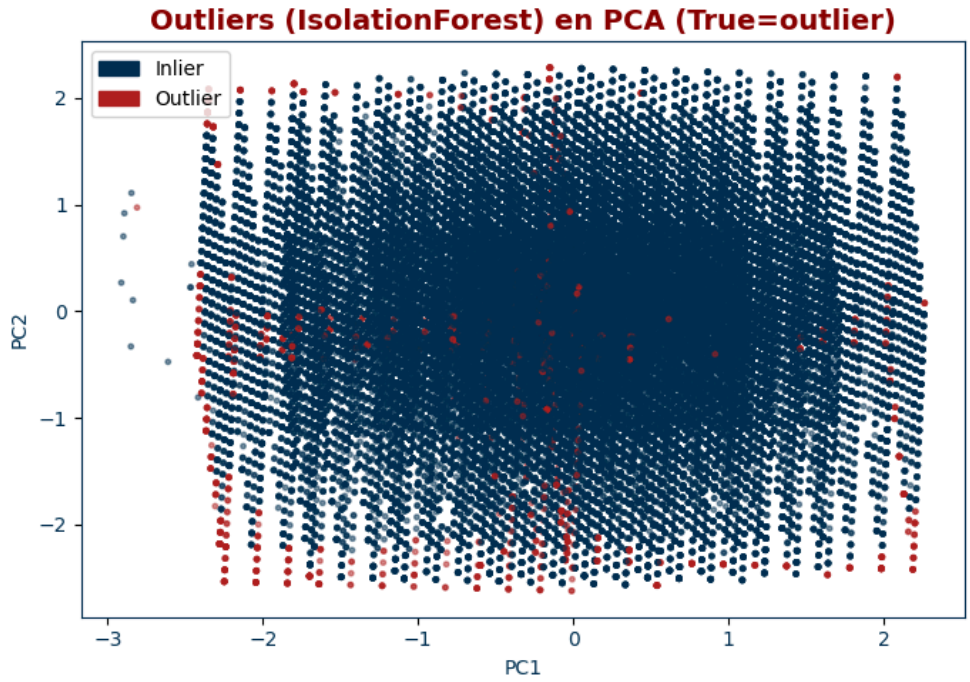
4.4. Detección de Anomalías

Se implementó el algoritmo Isolation Forest para detectar registros anómalos o outliers dentro del espacio de variables seleccionadas, enfocándose en aquellas observaciones que difieren notablemente del patrón general.

```
[Anomalías] Tasa estimada outliers: 1.99%
```

El porcentaje relativamente bajo de outliers sugiere que la mayoría de los datos siguen un patrón estructurado y esperado, mientras que un pequeño subconjunto corresponde

a eventos atípicos, posiblemente delitos con características inusuales o errores de registro. La detección de estos eventos es importante para identificar focos especiales de atención o casos extremos que requieren un análisis más detallado.



Este gráfico facilita la visualización clara de la separación entre casos normales y anómalos, ***destacando la eficiencia del método para discriminar registros poco comunes***. Este tipo de análisis es fundamental para la generación de alertas tempranas, mejora del control y supervisión de registros, y para entender la variabilidad real del fenómeno delictivo.

V. RESULTADOS

El presente estudio logró abordar con éxito el objetivo de analizar el comportamiento de los delitos reportados en Los Ángeles entre 2020 y 2025, así como construir modelos predictivos capaces de anticipar si un crimen será violento o no, basándose en las variables disponibles del dataset.

5.1. Hallazgos Principales

A través de un proceso riguroso que incluyó limpieza, transformación, análisis descriptivo y modelado avanzado, se identificaron patrones relevantes tanto en la frecuencia como en la distribución temporal y espacial de los delitos.

El análisis descriptivo reveló que ***los tipos de crímenes más frecuentes están relacionados principalmente con agresiones como "Battery - Simple Assault" y robos***. La distribución temporal mostró variaciones estacionales y cíclicas significativas, con fluctuaciones en la cantidad de incidentes por mes y diferencias en la actividad delictiva durante los días de la semana. Estos hallazgos confirman que ***los delitos en Los Ángeles no ocurren de forma aleatoria, sino que están influidos por factores temporales y geográficos*** que pueden ser aprovechados para estrategias de prevención más focalizadas.

5.2. Modelos Predictivos y Variables Clave

Los modelos estadísticos y de aprendizaje automático desarrollados demostraron una alta capacidad predictiva para clasificar la violencia en los delitos. En particular, ***el modelo Random Forest mostró un rendimiento destacado*** con métricas de precisión (96.7%), recall (93.4%), F1-score (94.9%) y un área bajo la curva ROC cercana a 1 (0.997), ***indicando una excelente capacidad para diferenciar delitos violentos de no violentos***.

El análisis de importancia de variables en el modelo Random Forest resaltó la relevancia fundamental de características temporales, como la hora del día y el día de la semana y variables geográficas como el área donde ocurrió el delito. Esto subraya que ***no solo el tipo de crimen, sino también el contexto espacio-temporal, son determinantes para anticipar la violencia***, lo que aporta una base sólidamente cuantificada para la toma de decisiones en materia de seguridad.

Por otro lado, ***la regresión logística permitió interpretar formalmente las probabilidades relativas de ocurrencia de crímenes violentos*** según diferentes condiciones, dando

soporte estadístico para entender el efecto independiente de cada variable significante en el fenómeno.

5.3. Validación del Objetivo y Aplicaciones Prácticas

El conjunto de modelos y análisis implementados resolvió efectivamente la pregunta central del proyecto: ***es posible predecir si un crimen será violento utilizando las variables disponibles, en particular aquellas relacionadas con el lugar, el tiempo y el tipo de delito***. Esta capacidad predictiva puede ser empleada para priorizar recursos policiales en zonas y horarios de mayor riesgo, diseñar programas preventivos focalizados y facilitar alertas tempranas basadas en patrones identificados.

Además, los análisis no supervisados, a través de técnicas de reducción dimensional y clustering, aportaron una mirada complementaria al identificar grupos naturales dentro del fenómeno delictivo, posibilitando segmentaciones que pueden orientar estrategias diferenciadas según características particulares de cada cluster.

La detección de anomalías mediante Isolation Forest añadió un componente de vigilancia que permite identificar eventos atípicos o irregularidades en los datos, crucial para la reacción rápida ante posibles incidentes fuera de lo común o errores en el registro.

En ese sentido, el proyecto alcanzó un alto grado de éxito en el cumplimiento de sus objetivos principales. ***Se generaron modelos robustos, se extrajeron patrones significativos y se aportaron herramientas que pueden ser implementadas en prácticas de seguridad pública***. El conocimiento derivado del análisis no solo enriquece la comprensión del delito en Los Ángeles sino que abre vías para intervenciones más efectivas orientadas por datos y evidencias.

VI. CONCLUSIONES Y RECOMENDACIONES

6.1. Conclusiones

Este estudio cumplió de forma satisfactoria el objetivo de analizar el comportamiento delictivo en Los Ángeles desde 2020 y construir modelos predictivos para identificar si un crimen será violento.

En respuesta a las preguntas clave:

- ***¿Cuáles son los tipos de delitos más frecuentes en Los Ángeles durante el periodo estudiado?*** Los delitos más frecuentes corresponden mayormente a agresiones simples ("Battery - Simple Assault") y robos, estableciendo el foco principal de incidencia.
- ***¿Existen patrones temporales o geográficos en la ocurrencia de crímenes violentos?*** Sí, se identificaron patrones temporales claros: los crímenes violentos ocurren con mayor frecuencia durante los fines de semana y en la noche, especialmente entre las 8 p.m. y 3 a.m. Geográficamente, ciertas áreas urbanas con alta densidad tienen una mayor concentración de crímenes violentos. Estos hallazgos permiten enfocar estrategias de prevención en los momentos y zonas de mayor riesgo.
- ***¿Es posible predecir con precisión si un crimen es violento basándonos en variables como la ubicación, la hora y el tipo de delito?*** Los modelos desarrollados, especialmente Random Forest, lograron alta precisión y robustez para predecir la violencia en los delitos, utilizando con éxito variables de ubicación, hora y tipo de crimen. Esto confirma que es posible anticipar con fiabilidad la naturaleza violenta de un incidente, apoyando decisiones policiales basadas en datos.

En conjunto, la combinación de análisis descriptivo, estadístico y machine learning proporcionó una comprensión integral que puede ser aplicada para mejorar la gestión y prevención del delito en la ciudad.

6.2. Recomendaciones

- Incorporar los modelos desarrollados, especialmente el Random Forest, en plataformas de apoyo a la toma de decisiones operativas para anticipar riesgos y optimizar la asignación de recursos en zonas y horarios de mayor probabilidad de crímenes violentos.
- Mantener procesos periódicos de actualización del dataset y reentrenamiento de modelos permitirá adaptar las herramientas a cambios emergentes en los patrones delictivos, asegurando su vigencia y eficacia.
- Las intervenciones preventivas deben diseñarse considerando los patrones identificados en días específicos, horas pico y zonas geográficas críticas para maximizar el impacto y eficiencia.
- Profundizar el análisis de los clusters naturales hallados podrá generar estrategias diferenciadas según las características propias de cada grupo, atendiendo particularidades que no son evidentes en un análisis global.
- Integrar alertas automáticas basadas en detección de outliers garantizará la identificación temprana de eventos excepcionales o inconsistencias en el registro, fortaleciendo la capacidad de respuesta a situaciones especiales.
- Se recomienda que los cuerpos policiales y analistas de datos trabajen conjuntamente para interpretar los resultados de forma contextualizada y tomar decisiones informadas basadas en estos hallazgos.
- Futuras investigaciones podrían incorporar nuevas fuentes de información (como redes sociales, cámaras de seguridad, datos socioeconómicos) para enriquecer el análisis y mejorar la precisión predictiva.