




# MINERÍA DE DATOS APLICADA AL ANÁLISIS DE DELITOS EN LOS ÁNGELES (2020-2025)

PROYECTO FINAL - DATA MINING

**GRUPO  
3**

 Karen Torrico  
 Karem Huacota  
 Yesika Luna

 Elvis Miranda  
 Ivan Mamani

Agosto, 2025

# CONTENT



01

Introducción

02

Definición del Problema

03

Problemas Resueltos

04

Métodos Utilizados

05

Resultados del Análisis

06

Conclusiones y  
Recomendaciones

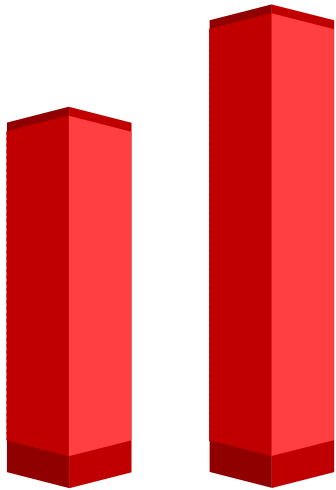
# Introducción

01



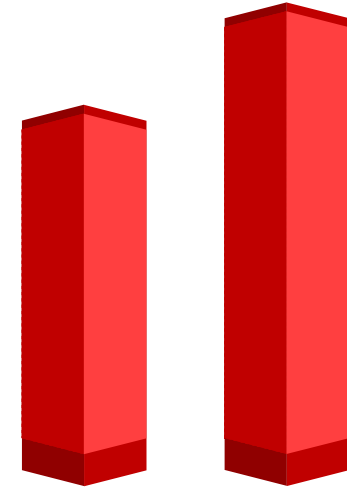
# Importancia del Análisis Delictivo

## Políticas Públicas y Seguridad



El análisis facilita la prevención, optimiza recursos policiales y mejora la confianza comunitaria.

## Enfoque en Los Ángeles



La ciudad es uno de los centros urbanos más grandes de EE.UU., con datos desde 2020 hasta 2025.



# Datos y Desafíos

---

## Amplio Dataset

Más de un millón de registros  
(1,004,991 filas) distribuidos en 28  
columnas.

1

## Variables Clave

Fecha, hora, tipo de crimen, demografía de  
víctimas, ubicación geográfica y armas  
utilizadas.

2

## Retos de Manejo y Procesamiento

Se requieren procesos rigurosos de  
limpieza y transformación para  
asegurar la calidad.

3



# Procesamiento de Datos

---

## Limpieza y Transformación

Eliminación de datos incompletos, normalización y creación de nuevas variables.

## Aseguramiento de la Calidad

Es fundamental para asegurar la calidad y consistencia de la información.



# Resultados y Avances

## Técnicas Aplicadas

Minería de datos, aprendizaje automático supervisado/no supervisado.



## Modelos Predictivos

Modelos con alta precisión para predecir la violencia en crímenes representan un avance.



# Definición del Problema



02





# Objetivos del Análisis

---

## **Análisis del Comportamiento Delictivo**

Analizar el comportamiento de los delitos reportados en Los Ángeles desde 2020 a 2025 y predecir si un crimen será violento o no.

1

## **Desarrollo de Modelos Predictivos**

Anticipar la violencia utilizando las variables disponibles en el dataset.

2



# Preguntas Clave

---

## Tipos de Delitos Frecuentes

¿Cuáles son los delitos más frecuentes en Los Ángeles durante el periodo?

## Patrones Temporales y Geográficos

¿Existen patrones temporales/geográficos en crímenes violentos (hora, día, zonas)?

## Predicción de la Violencia

¿Es posible predecir si un crimen es violento con la ubicación, hora y tipo de delito?

01

03

02



# Descripción del Dataset

---

## Fuente de Datos

Dataset oficial "Crime Data from 2020 to Present" (Los Ángeles).

## Actualización Periódica

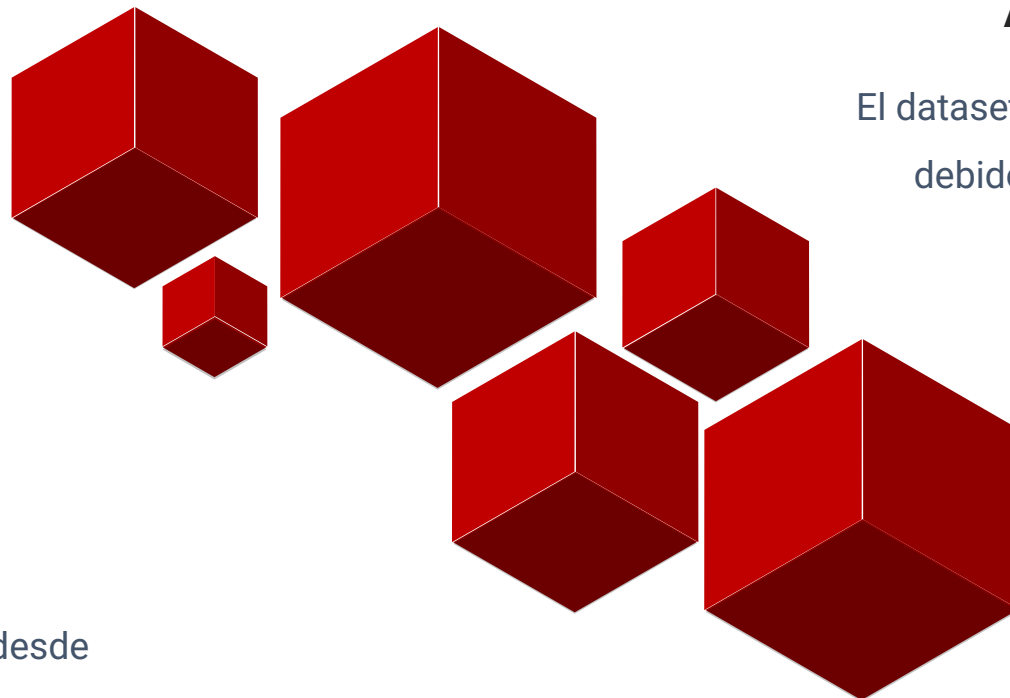
El dataset requiere limpieza y transformación debido a su tamaño y heterogeneidad.

## Cobertura Temporal

Incidentes delictivos reportados desde 2020 hasta 2025.

## Variables Relevantes

Fechas, horas, ubicación, tipo de crimen, demografía de víctimas e información de armas.



# Problemas Resueltos

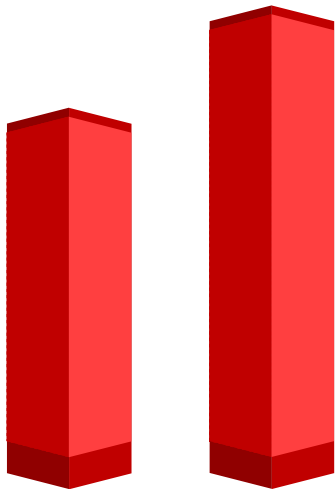


03



# Carga y Exploración Inicial de Datos

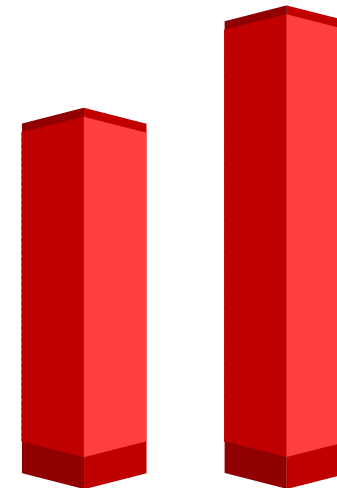
## Carga del Dataset



Carga de un gran dataset con más de un millón de registros y 28 columnas.



## Conocer la estructura inicial



Es necesario identificar la estructura y calidad de la información para planificar la limpieza.





# Gestión de Valores Faltantes

## Identificación de Valores Nulos

Evaluar la cantidad de valores nulos en cada variable.



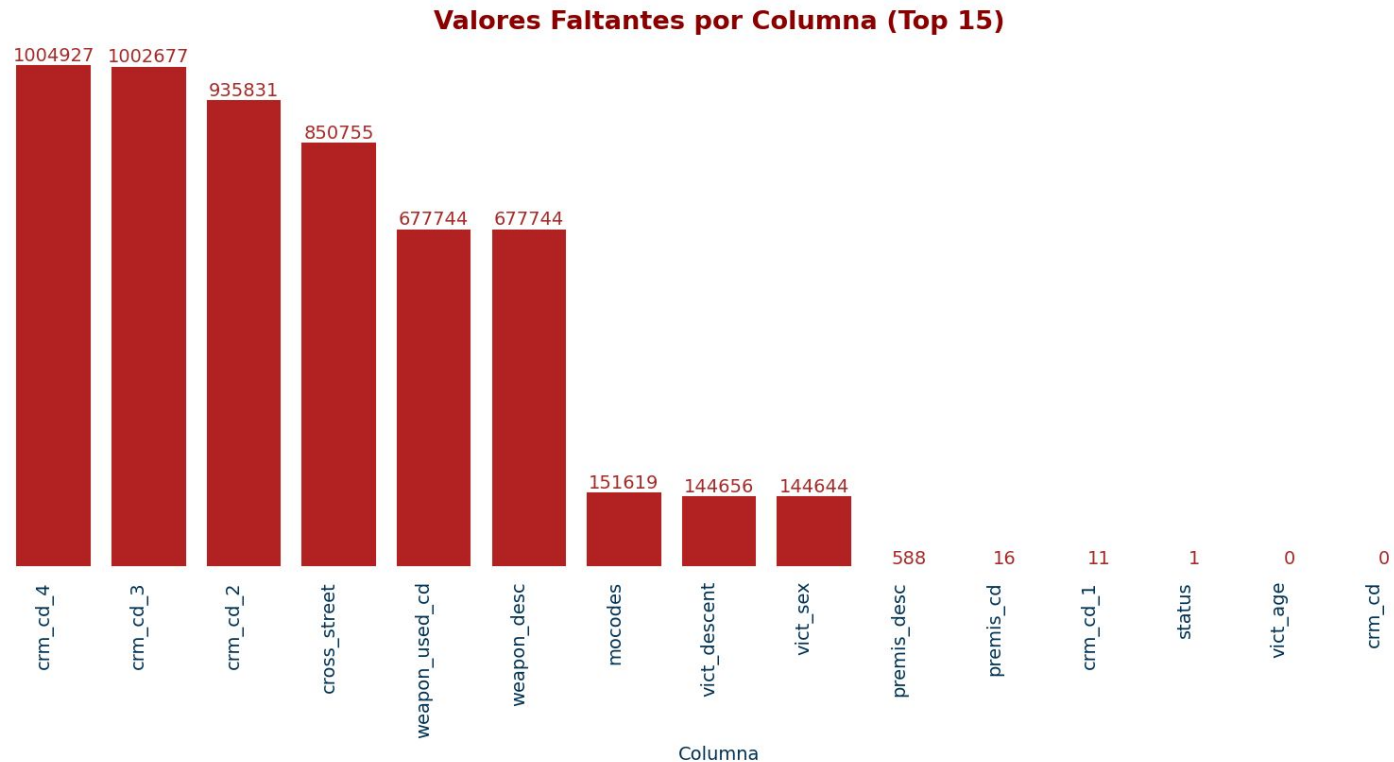
## Estrategias de Eliminación o Imputación

Para evitar sesgos o errores en el análisis, se requiere identificar y definir filtros.





# Valores Faltantes por Columna



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.



# Transformación de Datos

## Parseo de Fechas Frecuentes

Conversión de columnas de fechas a formato datetime. La conversión permite extraer información temporal.

## Parseo de Hora

Normalización y extracción de la hora para crear la variable numérica hour\_occ.

## Selección de Fecha Base

Se determina qué columna de fecha será la referencia para el análisis temporal.

## Normalización de Variables Temporales

Extracción de variables como año, mes, día de la semana y fecha del mes.







# Limpieza de Datos



## Limpieza de las coordenadas

Se realiza búsqueda de limpieza de las columnas de latitud y longitud creando un indicador.

1

2

3

4



## Creación de variable objetivo

Se crea un variable binaria indicando si el delito fue violento para la clasificación.

## Limpieza de Variables Categóricas



Se normalizaron columnas como area\_name para reducir inconsistencias y variabilidad.

## Limpieza final de registros



Es necesario eliminar registros duplicados y aquellos con falta de datos importantes.

# Métodos Utilizados

04

# Análisis Descriptivo

## Correlaciones Numéricas

---

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

## Histograma de Edad de Víctima

---

Se estudió la distribución de la edad de las víctimas.

## Distribución por Día de la Semana

---

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

## Tendencia Mensual de Incidentes

---

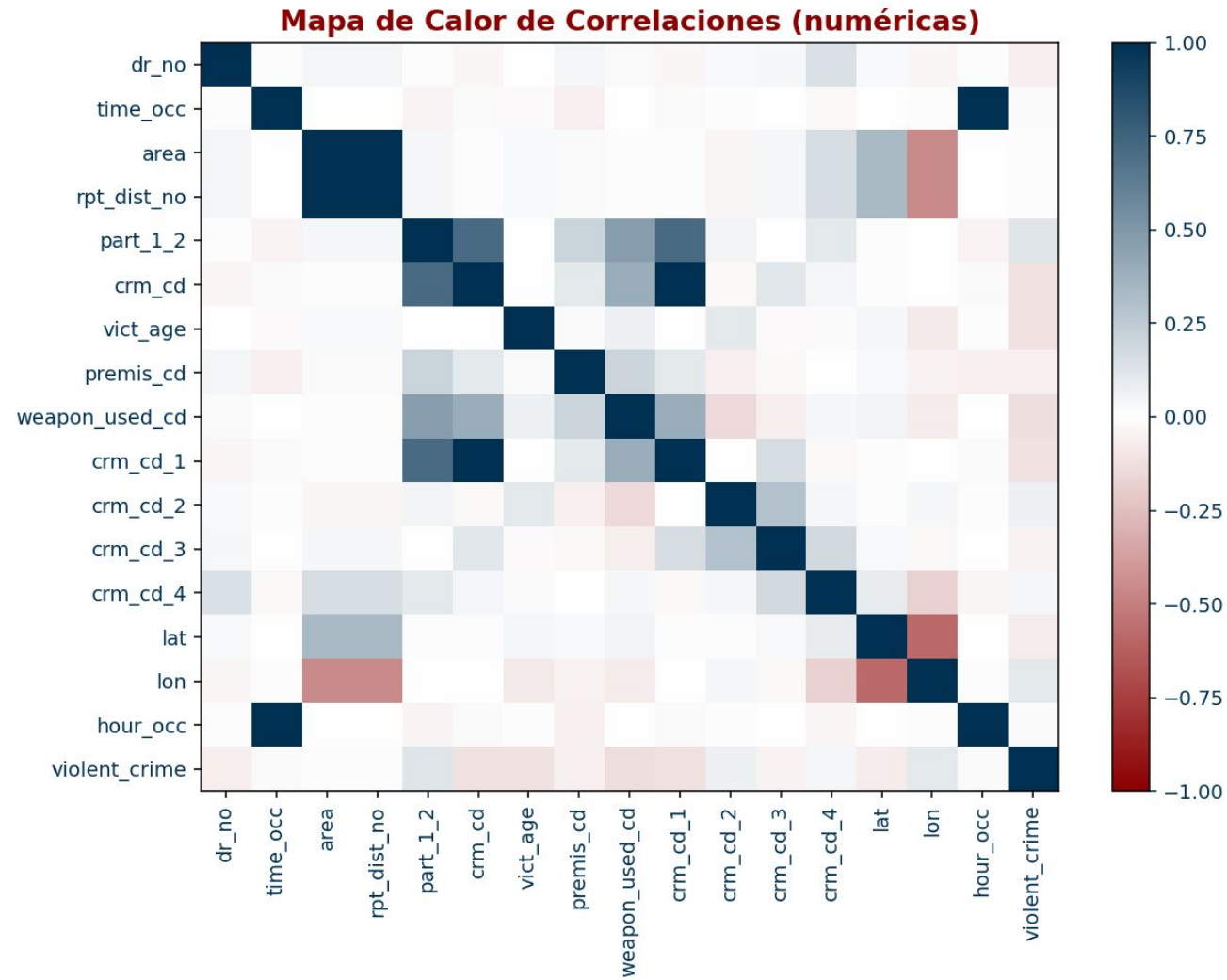
Se analizó la evolución del número de incidentes por mes.

## Top 15 Tipos de Crimen

---

Se obtuvieron y visualizaron los 15 tipos de delito más frecuentes.

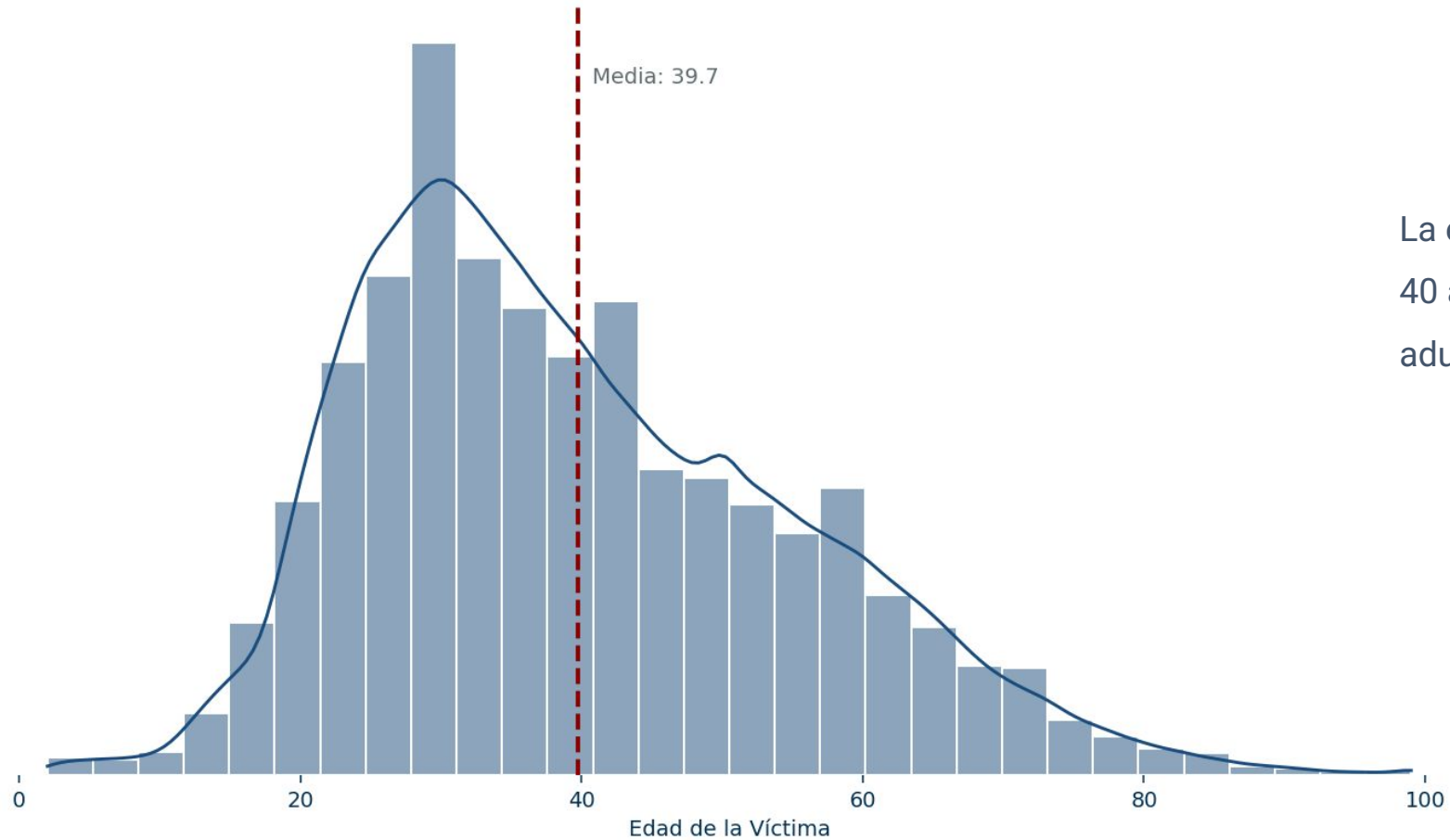
# Análisis Descriptivo



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

# Análisis Descriptivo

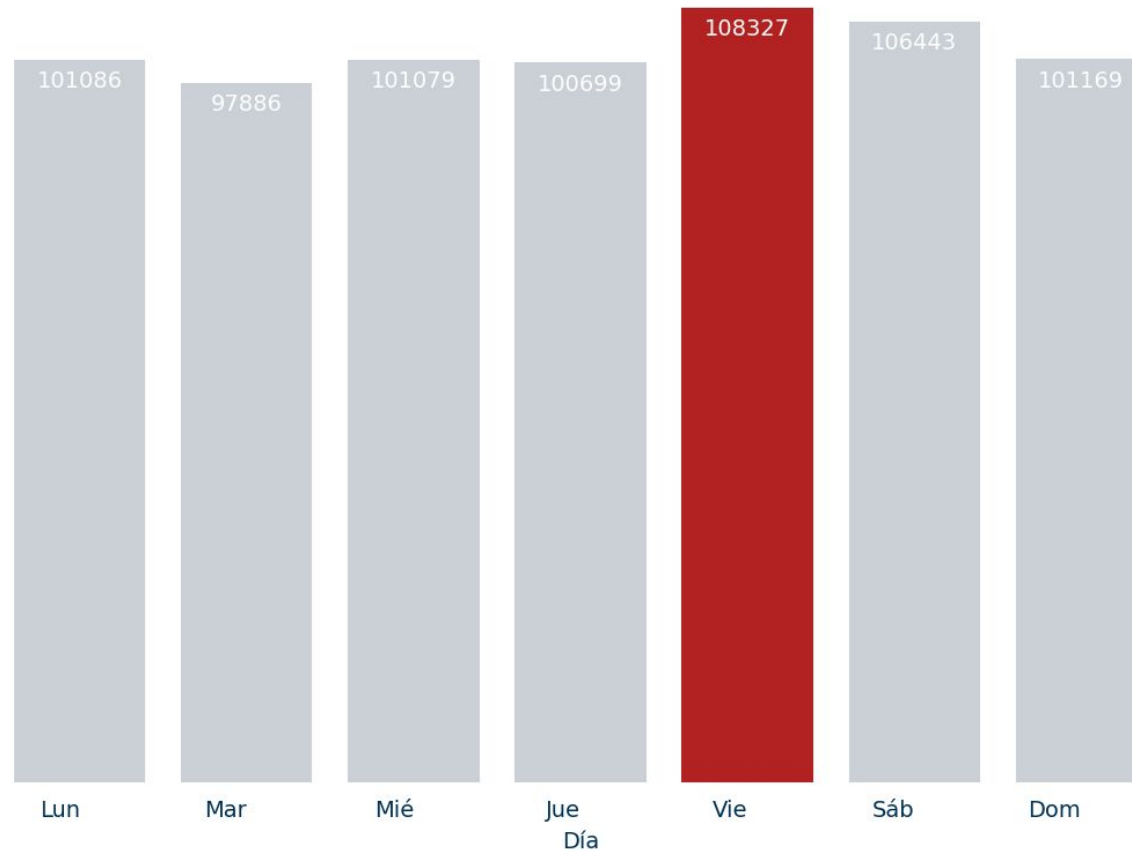
**Distribución de la Edad de Víctimas**



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

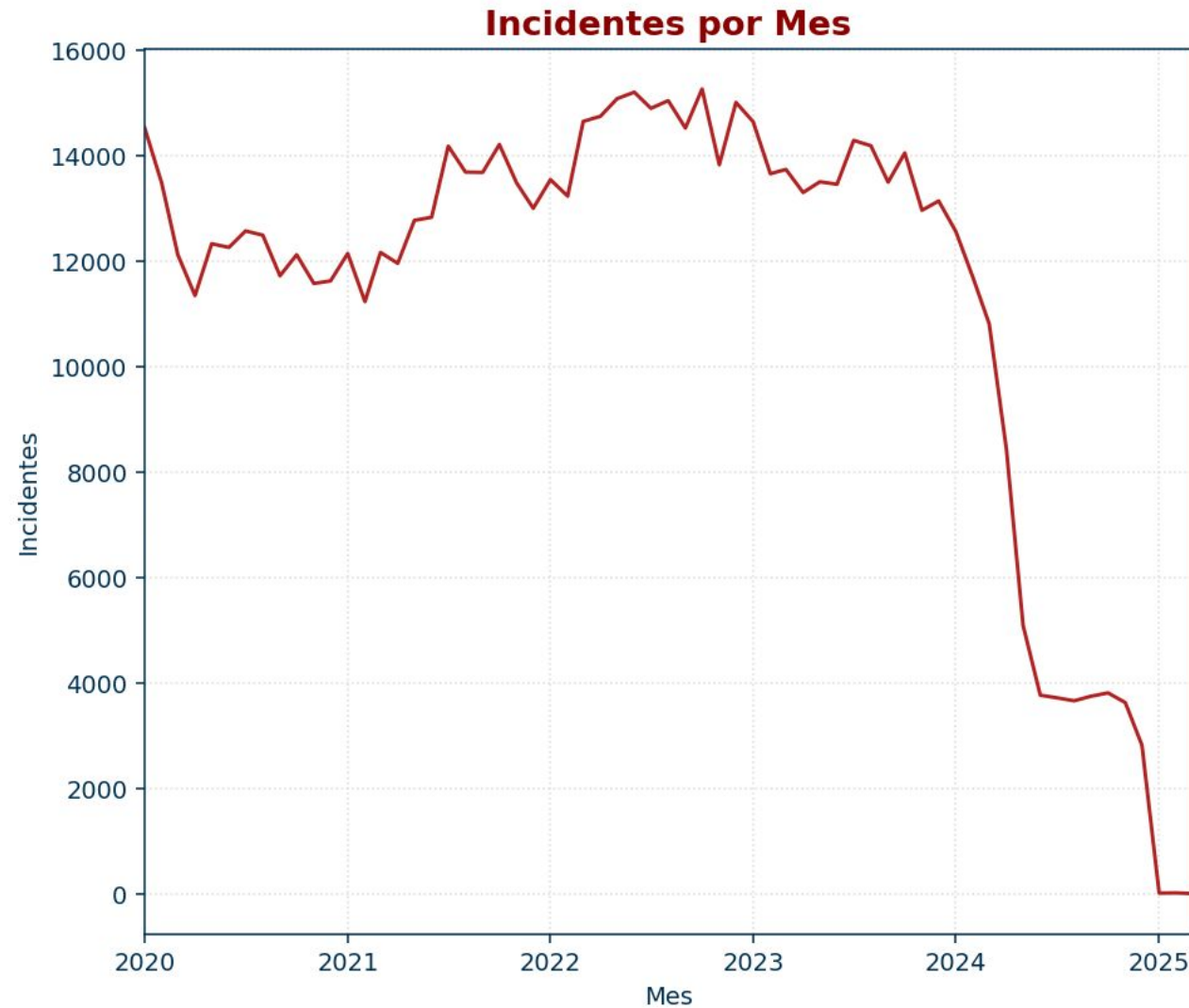
# Análisis Descriptivo

Incidentes por Día de la Semana (0=Lun)



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 días viernes.

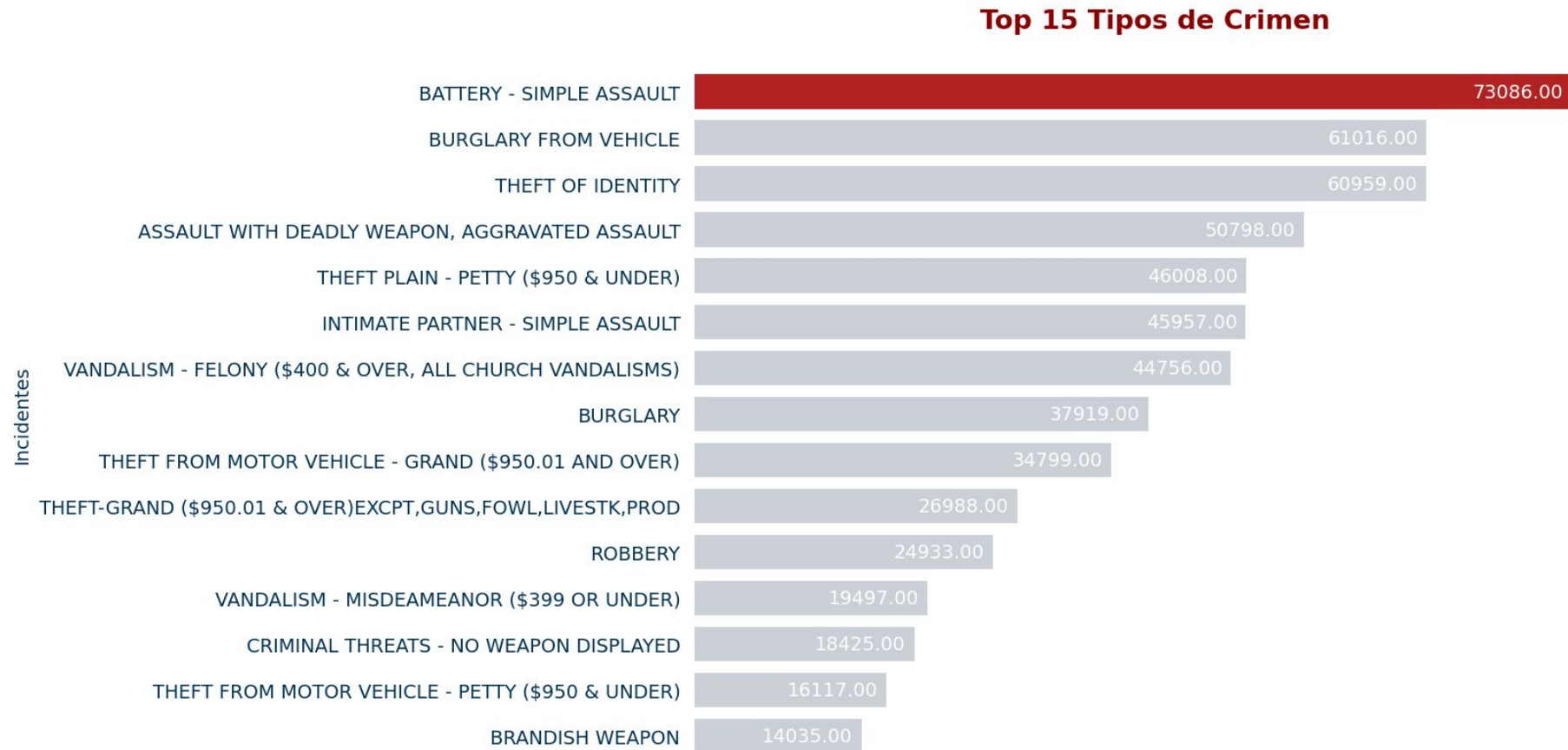
# Análisis Descriptivo



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



# Análisis Descriptivo



Los datos indican que el ***delito más común es por Agresión simple*** (Battery - Simple Assault) seguido por ***robo de vehículo*** (Burglary from Vehicle) y otros tipos de agresiones, resaltando áreas prioritarias para políticas de seguridad.



# Modelado Estadístico

Se ajustó un modelo de regresión lineal para explicar las variaciones mensuales.

**Regresión Lineal OLS sobre Incidencias Mensuales**

01

Se graficaron los residuos a lo largo del tiempo.

**Residuos del Modelo OLS**

02

**Ajuste del Modelo OLS**

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

03

**Regresión Logística para Odds Ratios**

Se utilizó regresión logística para modelar la probabilidad de que un delito sea violento.

04

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

**Machine Learning: Random Forest**

05

# Modelado Estadístico

## Regresión Lineal OLS sobre Incidencias Mensuales

[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.

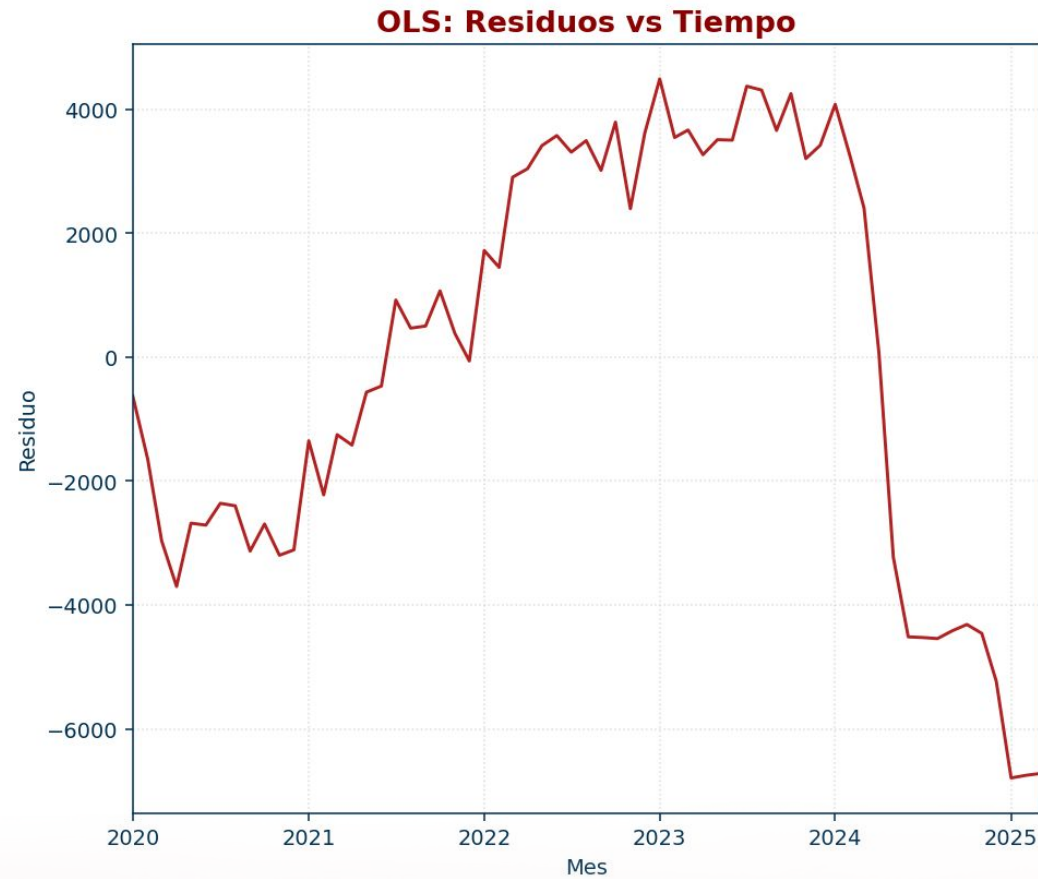
El modelo de regresión lineal muestra que aproximadamente **el 36% de la variabilidad en los incidentes ( $R^2 = 0.361$ ) se explica por las variables año y mes.**

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.**

# Modelado Estadístico

## Residuos del Modelo OLS



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

# Modelado Estadístico

## Machine Learning: Random Forest

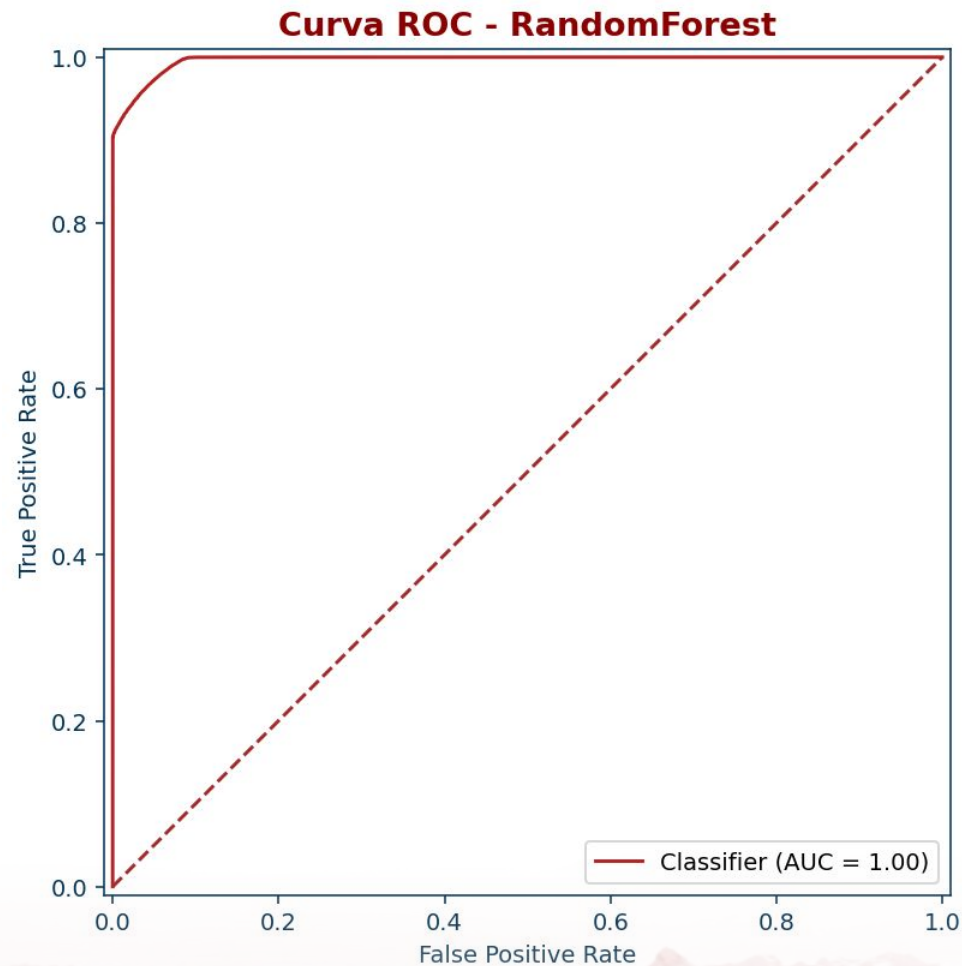
```
[RandomForest] Métricas:  
{  
  "accuracy": 0.9673219941676318,  
  "precision": 0.9648120367102012,  
  "recall": 0.9340267395668703,  
  "f1": 0.9491698317959848,  
  "roc_auc": 0.9966567254312784  
}
```

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.
- 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.
- 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

# Modelado Estadístico

## Machine Learning: Random Forest

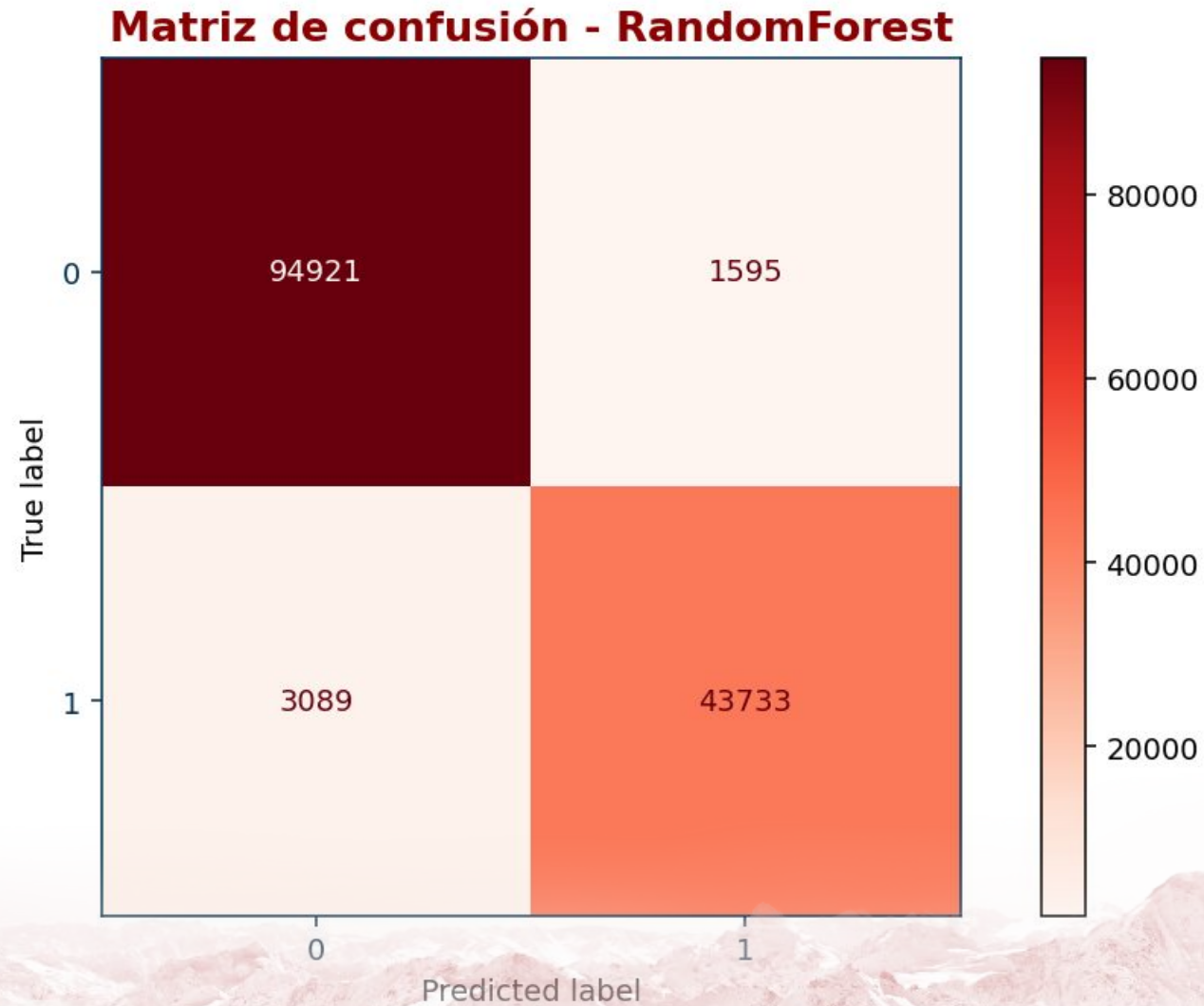


La curva destaca una casi perfecta discriminación entre clases, lo que respalda su aplicación práctica.



# Modelado Estadístico

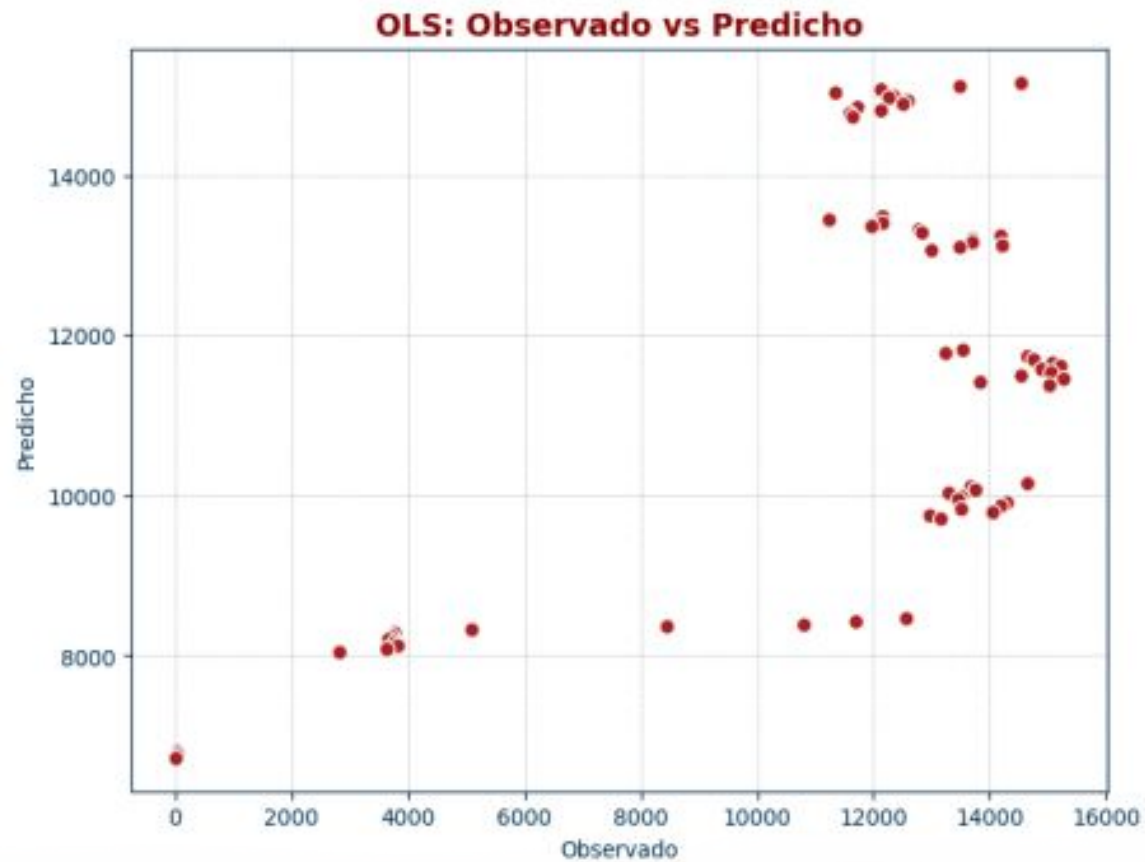
## Machine Learning: Random Forest



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

# Modelado Estadístico

## Ajustes del Modelo OLS



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

# Modelado Estadístico

## Regresión Logística para Odds Ratios

[Logit] Coeficientes y OR:

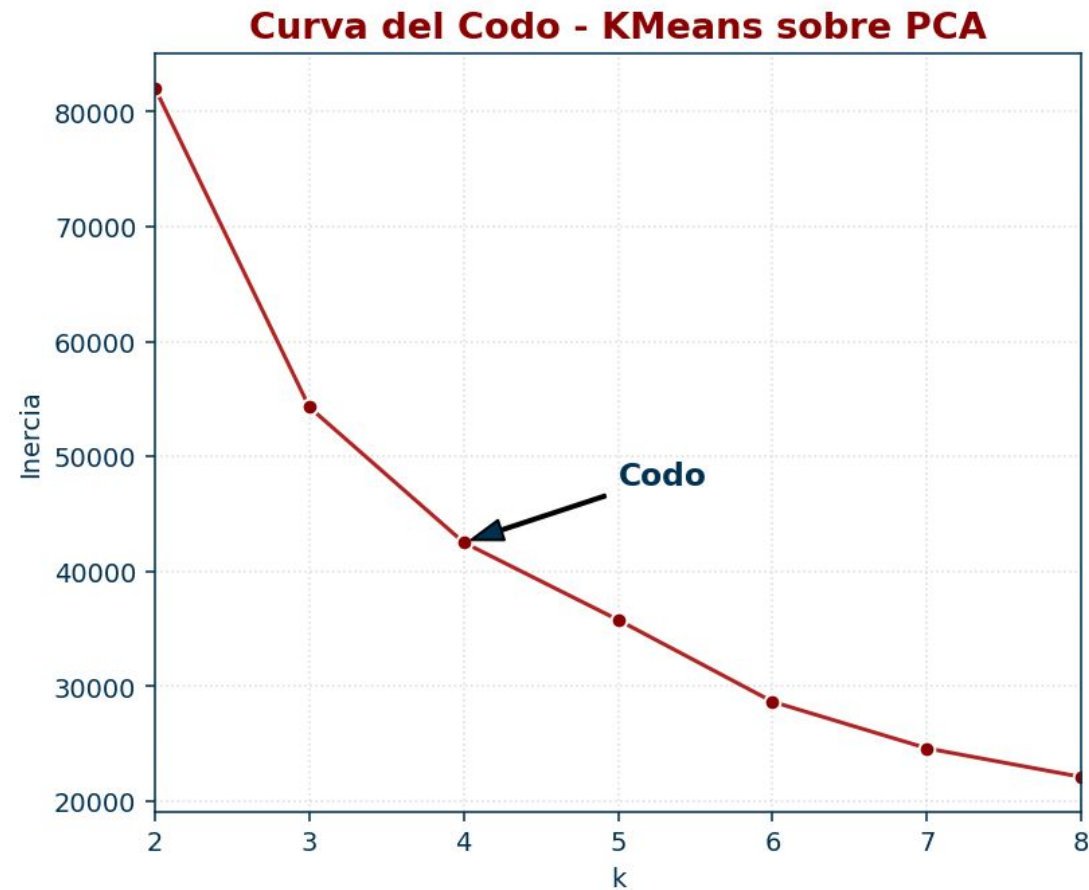
	coef	OR		pvalue
premis_desc_top_SIDEWALK	0.816005	2.261447	premis_desc_top_SIDEWALK	7.059446e-169
const	0.402085	1.494938	const	3.421455e-41
area_name_top_Southeast	0.103444	1.108984	area_name_top_Southeast	2.080242e-12
dayofweek	0.037483	1.038194	dayofweek	2.428328e-173
hour_occ	0.007551	1.007579	hour_occ	4.200617e-76
month	-0.001302	0.998699	month	8.932637e-02
premis_desc_top_PARK/PLAYGROUND	-0.005294	0.994720	premis_desc_top_PARK/PLAYGROUND	8.910349e-01
premis_desc_top_HOTEL	-0.007476	0.992552	premis_desc_top_HOTEL	8.539403e-01
premis_desc_top_GAS STATION	-0.015569	0.984551	premis_desc_top_GAS STATION	7.201576e-01
area_name_top_Newton	-0.168000	0.845354	area_name_top_Newton	4.337834e-28
premis_desc_top_RESTAURANT/FAST FOOD	-0.318516	0.727227	premis_desc_top_RESTAURANT/FAST FOOD	1.506591e-17
area_name_top_Rampart	-0.336758	0.714082	area_name_top_Rampart	8.532860e-104
area_name_top_Other	-0.375281	0.687096	area_name_top_Other	2.425636e-195
area_name_top_Southwest	-0.436110	0.646547	area_name_top_Southwest	3.156892e-198
area_name_top_Mission	-0.515067	0.597461	area_name_top_Mission	3.139760e-224
premis_desc_top_MULTI-UNIT DWELLING (APARTMENT,...	-0.521508	0.593625	premis_desc_top_MULTI-UNIT DWELLING (APARTMENT,...	4.031200e-78
area_name_top_Olympic	-0.522139	0.593250	area_name_top_Olympic	2.981539e-254
premis_desc_top_STREET	-0.542824	0.581105	premis_desc_top_STREET	2.608526e-85
premis_desc_top_OTHER BUSINESS	-0.628441	0.533423	premis_desc_top_OTHER BUSINESS	6.425413e-91
premis_desc_top_Other	-0.657661	0.518062	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.



# Modelado No Supervisados

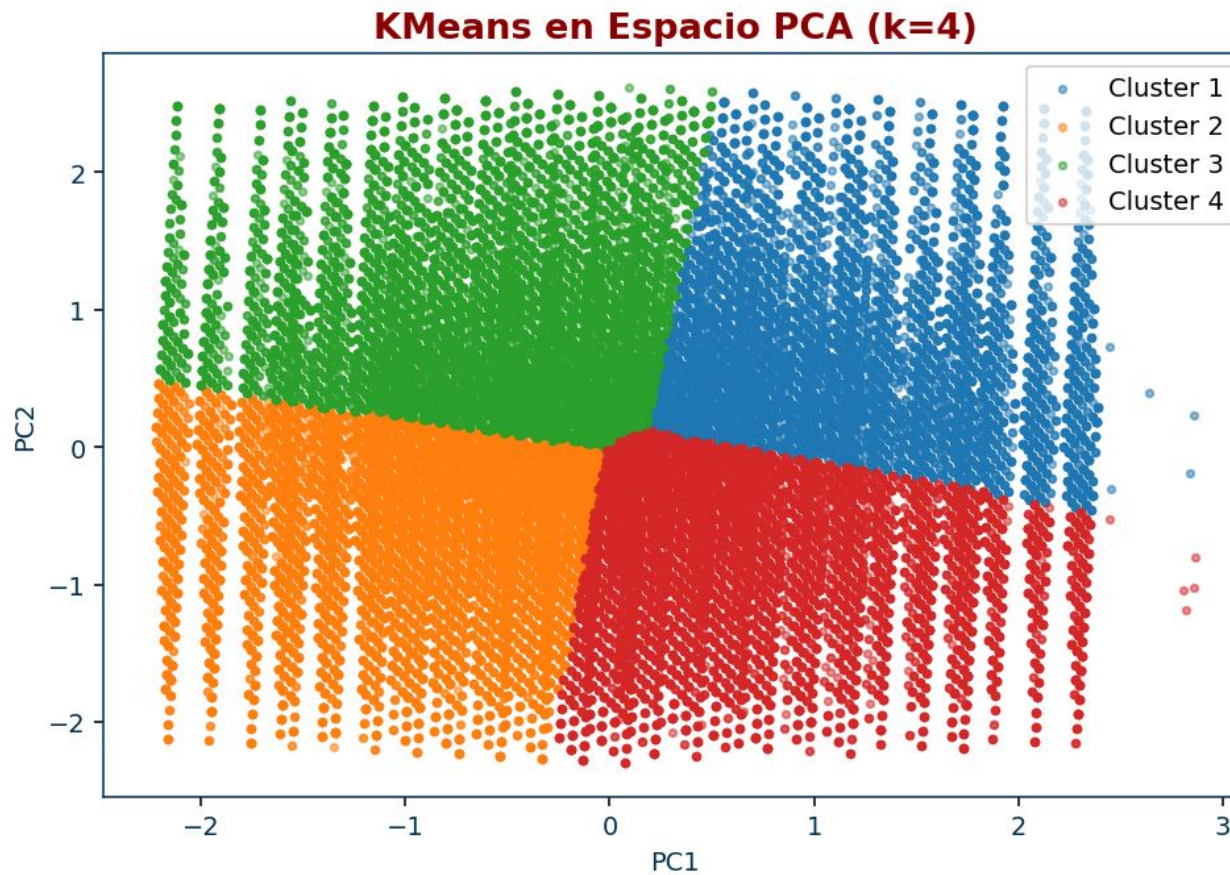
## PCA (Principal Component Analysis)



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

# Modelado No Supervisados

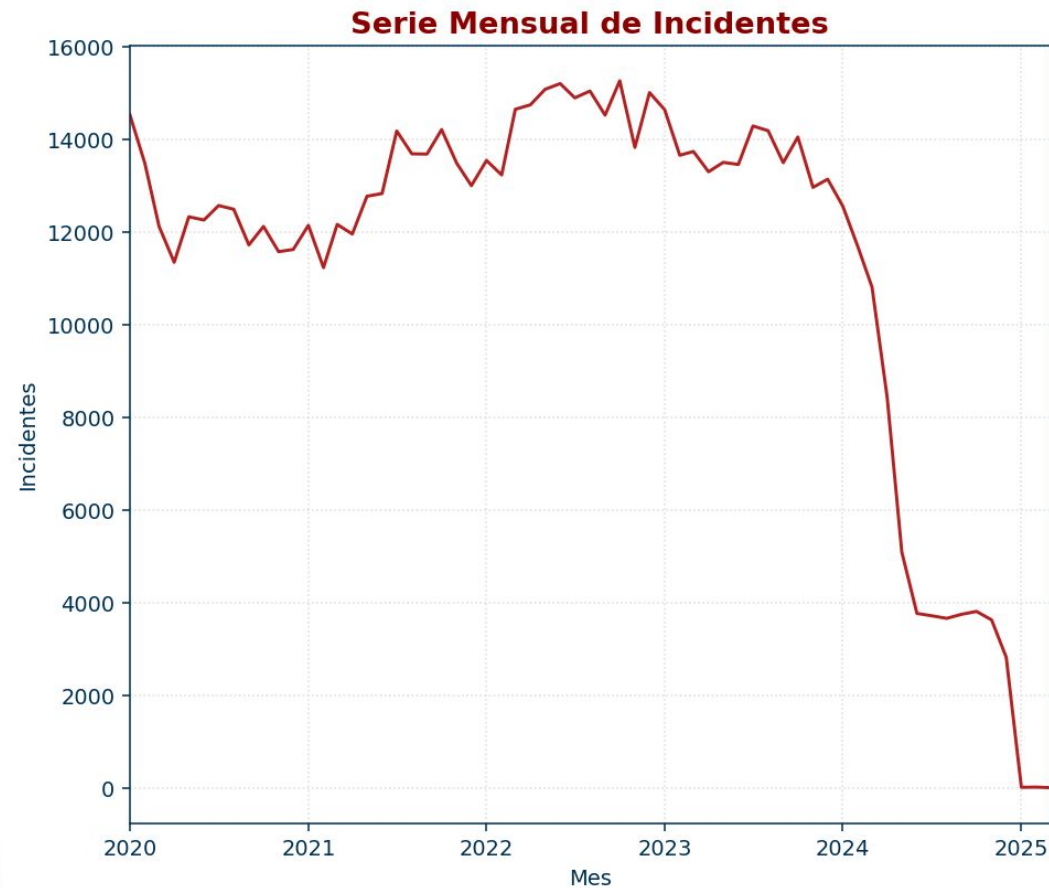
## KMeans



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.

# Modelado No Supervisados

## Series de Tiempo

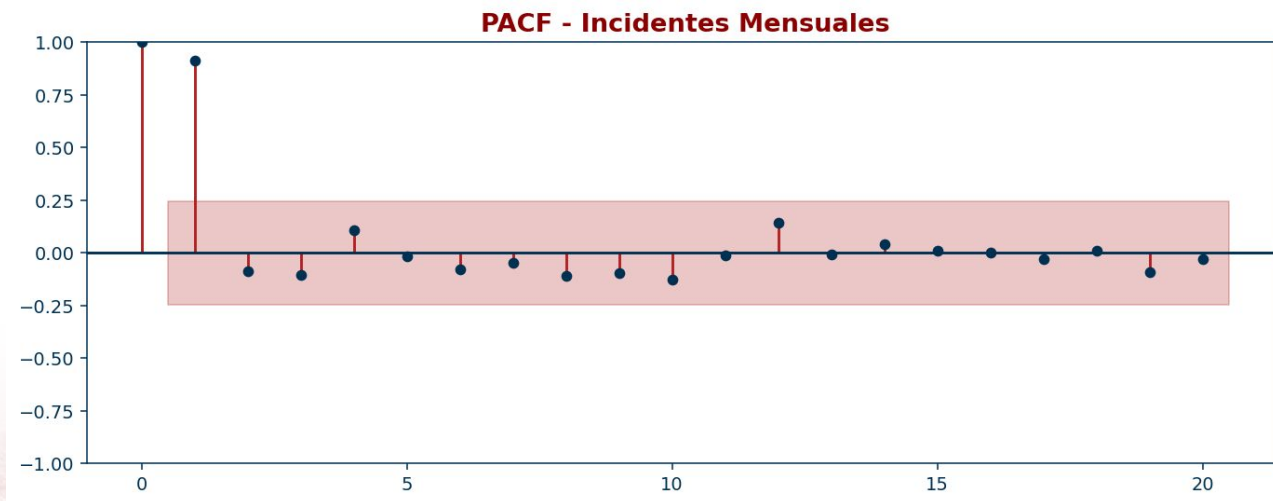
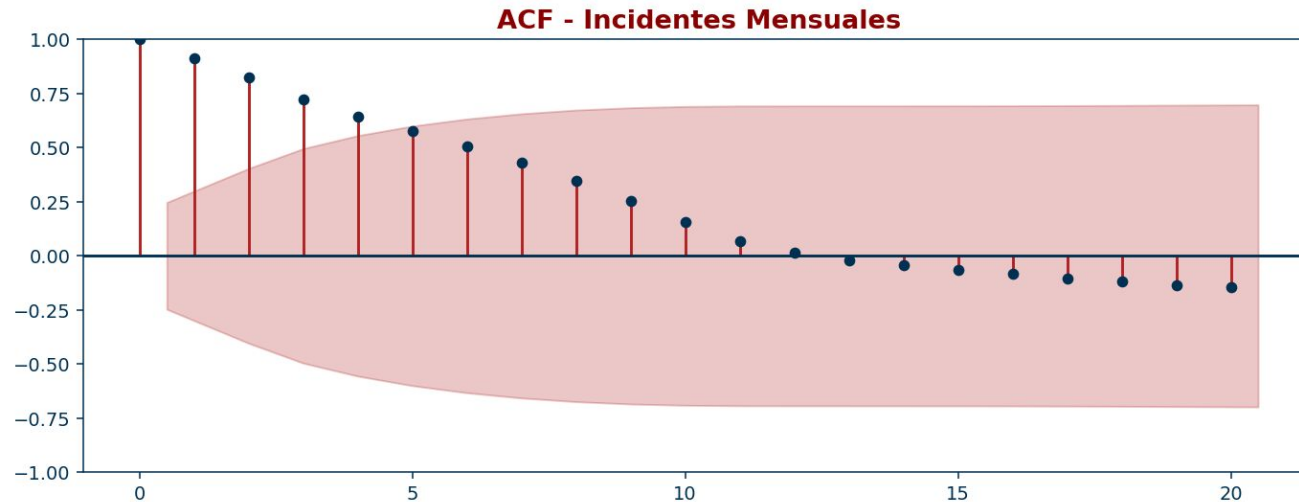


Este gráfico ilustra la dinámica de los incidentes delictivos a lo largo del tiempo, mostrando variaciones, tendencias y posibles picos estacionales. Esta representación permite identificar períodos de incremento o disminución del delito, apoyando la toma de decisiones para asignar recursos en momentos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



# Modelado No Supervisados

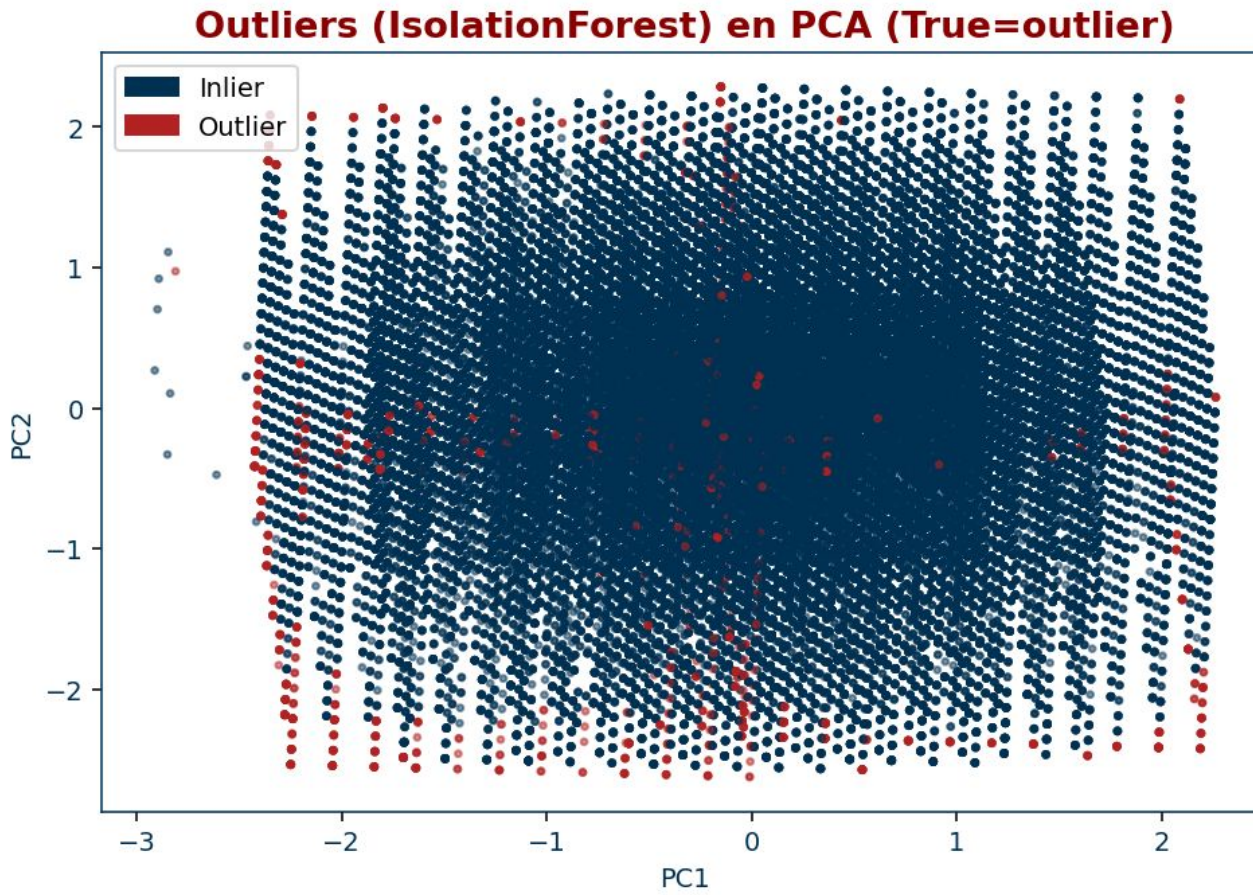
## ACF y PACF - Incidentes Mensuales



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.

# Modelado No Supervisados

## Detección de Anomalías



Este gráfico facilita la visualización clara de la separación entre casos normales y anómalos, destacando la eficiencia del método para discriminar registros poco comunes

# Resultados del Análisis



05



# Hallazgos principales

---

- Se identificaron **patrones relevantes** en la frecuencia y en la **distribuição temporal y espacial** de los delitos.
- Los crímenes más frecuentes corresponden a **agresiones (“Battery - Simple Assault”) y robos**.
- La distribución temporal mostró **variaciones estacionales y cíclicas**, con fluctuaciones mensuales y diferencias entre días de la semana.
- Los delitos en Los Ángeles **no ocurren de forma aleatoria**, sino que están influenciados por **factores temporales y geográficos**.





# Modelos Predictivos y Variables Clave

---

- Se desarrollaron **modelos estadísticos y de machine learning** con alta capacidad predictiva para clasificar la violencia en delitos.
- El **Random Forest obtuvo el mejor rendimiento** con: **precisión 96.7%, recall 93.4%, F1-score 94.9% y AUC ~0.997**, mostrando excelente capacidad de diferenciación entre delitos violentos y no violentos.
- El **análisis de importancia de variables** destacó el peso de **factores temporales** (hora del día, día de la semana) y **geográficos** (área del delito).
- Estos resultados evidencian que **el contexto espacio-temporal es tan determinante como el tipo de crimen para anticipar la violencia**.
- La **regresión logística** aportó una interpretación estadística formal, permitiendo entender **el efecto independiente de cada variable significativa** en la ocurrencia de crímenes violentos.







# Validación del Objetivo y Aplicaciones Prácticas

- Los modelos confirmaron que ***es posible predecir si un crimen será violento usando variables de lugar, tiempo y tipo de delito.***
- Esta capacidad predictiva ***permite priorizar recursos policiales, diseñar programas preventivos*** focalizados y generar alertas tempranas.
- Los análisis no supervisados (***reducción dimensional y clustering***) identificaron ***grupos naturales de delitos***, facilitando estrategias diferenciadas por cluster.
- La ***detección de anomalías con Isolation Forest*** añadió un componente de vigilancia para eventos atípicos o errores de registro.
- El proyecto ***cumplió exitosamente sus objetivos***: generó ***modelos robustos***, extrajo ***patrones significativos*** y aportó herramientas prácticas para la seguridad pública.



# Conclusiones y Recomendaciones

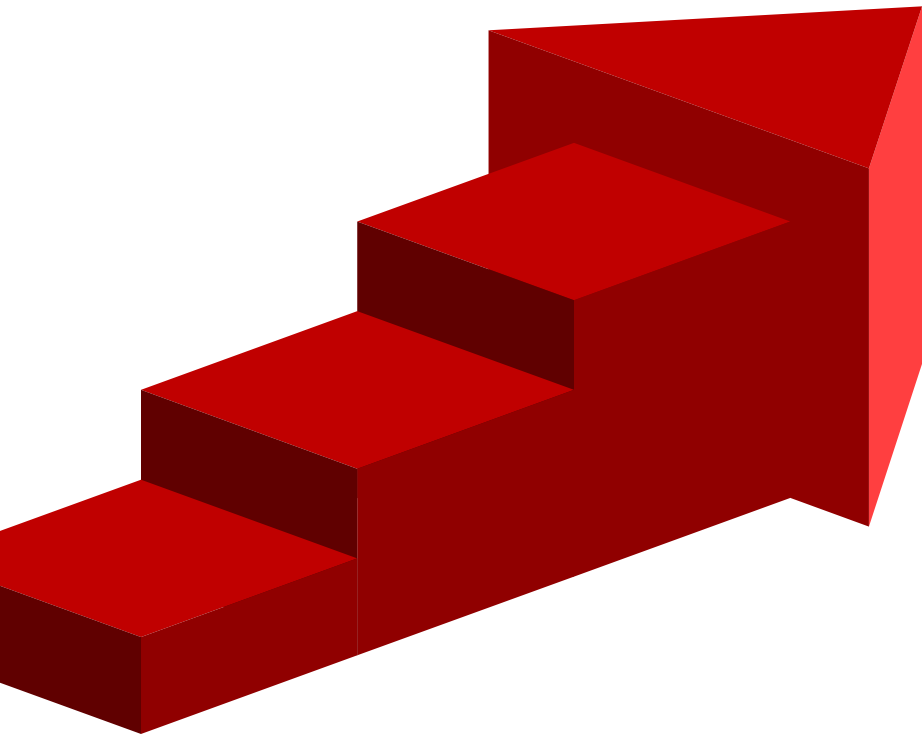


06



# Conclusiones

---

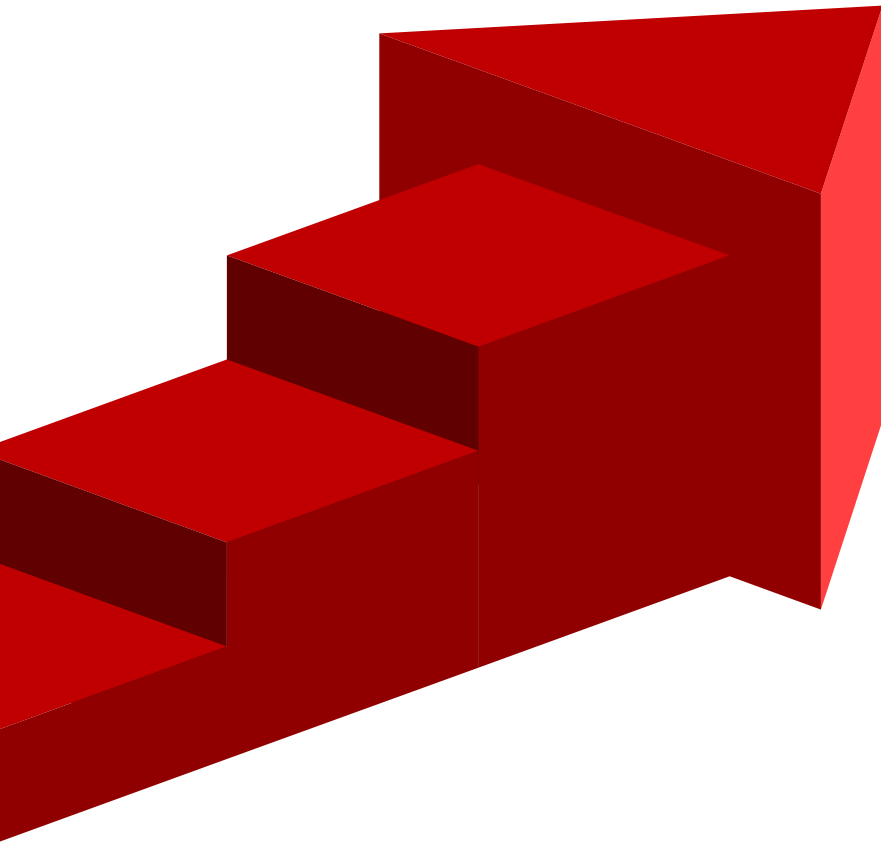


- El estudio *cumplió el objetivo de analizar el comportamiento delictivo en Los Ángeles* desde 2020 y construir modelos predictivos para identificar crímenes violentos.
- Tipos de delitos más frecuentes: *agresiones simples ("Battery - Simple Assault") y robos.*
- Patrones temporales: *mayor incidencia de crímenes violentos en fines de semana y noches (8 p.m. – 3 a.m.).*
- Patrones geográficos: *mayor concentración en áreas urbanas densamente pobladas.*
- Predicción de violencia: los modelos, especialmente *Random Forest*, mostraron *alta precisión y robustez al anticipar si un crimen será violento* basándose en ubicación, hora y tipo de delito.
- La *combinación de análisis descriptivo, estadístico y machine learning* brindó una *visión integral*, útil para mejorar la gestión y prevención del delito en la ciudad.



# Recomendaciones

---



- **Implementar y actualizar periódicamente los modelos predictivos** (especialmente Random Forest) para anticipar riesgos y optimizar recursos.
- **Diseñar intervenciones preventivas focalizadas según patrones de tiempo y espacio**, y profundizar el análisis de clusters para estrategias diferenciadas.
- **Integrar alertas automáticas de anomalías** para detectar eventos excepcionales y mejorar la capacidad de respuesta.
- **Promover la colaboración entre policías y analistas**, e incorporar nuevas fuentes de datos para enriquecer el análisis y aumentar la precisión predictiva.



**SCHOOL OF ENGINEERING**  
UNIDAD DE POSTGRADO FICCT - UAGRM  
PLATAFORMA EDUCATIVA

# Thank you!

