Pontificia Universidad Javeriana

Proyecto de Procesamiento de Datos a Gran Escala

Entrega 2

Preparación de datos, modelado y presentación de resultados

Equipo de Consultoría

Fabián Andrés Díaz Martínez
Juan David Ramírez
David León
Juan Diego Carreño Vásquez
Tomás Pinilla Flórez

Docente: Miguel Méndez Hernández

Bogotá, Colombia
24 de septiembre de 2025

Índice

1 Filtros y transformaciones

1.1 Filtros aplicados sobre los datos

1.1.1 Filtro de consistencia geográfica por borough

Se eliminaron registros sin *borough* definido para garantizar asociación válida a unidades territoriales, evitando distorsiones en comparaciones y asegurando coherencia en agregaciones geográficas.

1.1.2 Filtro de cobertura mensual válida

Se conservaron únicamente registros con valores de mes entre 1–12, removiendo casos atípicos. Esto garantiza estructura temporal ordenada y series uniformes para análisis mensuales comparables.

1.1.3 Filtro de eliminación de duplicados residuales

Se removieron duplicados mensuales por *borough*, manteniendo una observación única por combinación `BOROUGH-YEAR-MONTH`, evitando inflación artificial de conteos.

1.2 Transformaciones realizadas

1.2.1 Transformación 1 — Índice de severidad ponderado de colisiones

Se construyó un indicador sintético combinando heridos y fallecidos con peso diferenciado. El índice, agregado mensualmente por *borough*, permite comparar impacto relativo y priorizar territorios según riesgo vial.

1.2.2 Transformación 2 — Panel mensual unificado

Se integraron datos mensuales de arrestos, colisiones y contexto socioeconómico. Se construyó un indicador de razón operativo–arrestos para identificar *boroughs* con intervención policial proporcionalmente mayor o menor frente a incidencia vial.

1.2.3 Transformación 3 — Suavizamiento temporal y rezagos

Se aplicaron medias móviles de tres meses y rezagos de un periodo para capturar tendencias y reducir volatilidad mensual, facilitando análisis predictivos.

1.2.4 Transformación 4 — Variaciones porcentuales interanuales

Se generaron variaciones YoY comparando cada mes con su equivalente anual anterior, permitiendo evaluar mejoras o deterioros estructurales en seguridad y siniestralidad.

2 Respuesta a preguntas de negocio planteadas

2.1 Respuestas a las preguntas de negocio

2.1.1 1. Top 5 precincts con más arrestos en los últimos tres años

Los precincts 14, 40, 75, 103 y 44 concentran el mayor número de arrestos. El precinct 14 registra el volumen más alto, evidenciando concentración espacial crítica donde convergen factores estructurales y dinámicas delictivas intensas.

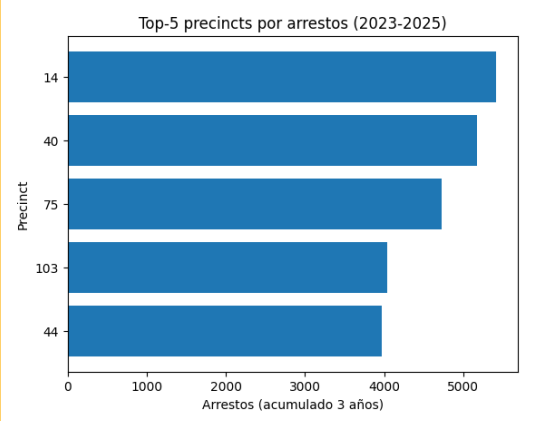

Figura 1: Top 5 precincts con más arrestos.

2.1.2 2. ¿Qué delitos se concentran en mayor medida en cada borough?

Bronx y Queens dominados por agresiones; Brooklyn combina delitos violentos con hurtos; Manhattan presenta fuerte presencia de *Petit Larceny*; Staten Island registra niveles bajos dispersos.

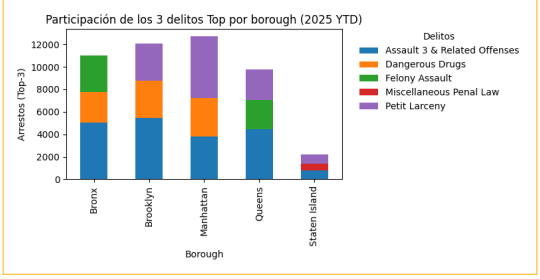

Figura 2: Delitos predominantes por borough.

2.1.3 3. ¿En qué meses del año se intensifican los arrestos?

Mayor intensidad entre enero y junio con picos en marzo y mayo. Correlación positiva débil (r = 0.33) con colisiones sugiere estacionalidad asociada a movilidad urbana.

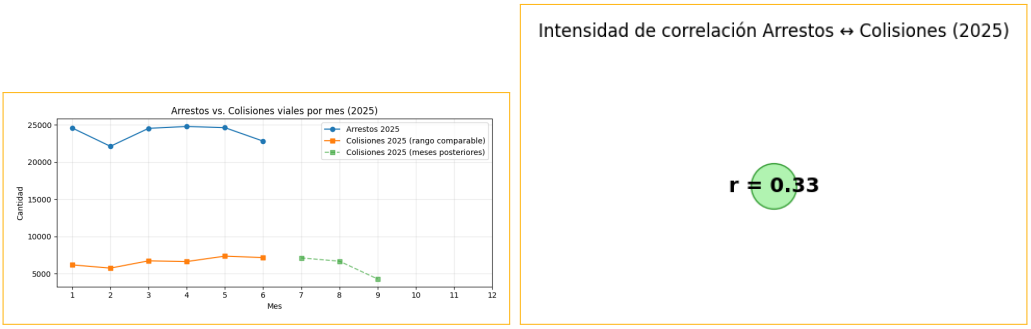

Figura 3: Estacionalidad mensual (izq.) y correlación (der.).

2.1.4 4. ¿Qué intersecciones y boroughs presentan mayor número de accidentes con víctimas?

Brooklyn y Queens lideran en accidentes con víctimas. Intersecciones críticas en Belt Parkway, Long Island Expressway y Brooklyn-Queens Expressway requieren intervenciones de control de velocidad.

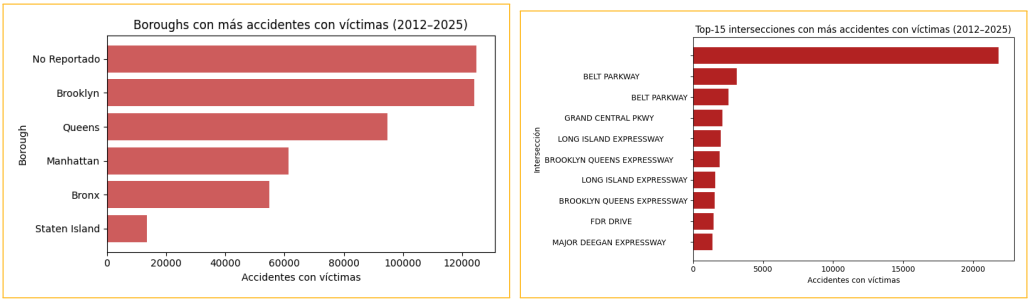

Figura 4: Boroughs con más accidentes (izq.) e intersecciones críticas (der.).

2.1.5 5. ¿Qué factores contribuyentes aparecen con mayor frecuencia en accidentes graves?

Predominan causas humanas: distracción del conductor, no ceder el paso y seguir demasiado cerca. La mayoría de siniestros graves asociados a comportamientos imprudentes más que fallas mecánicas.

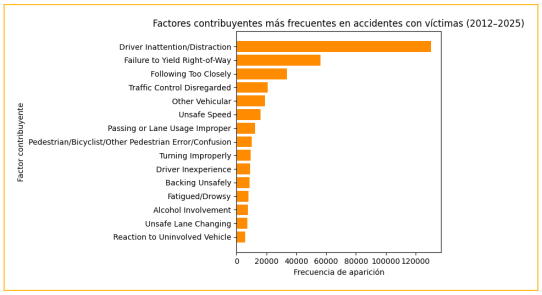

Figura 5: Factores contribuyentes en accidentes graves.

2.1.6 6. ¿Qué boroughs presentan simultáneamente mayor pobreza y mayores niveles de arrestos o colisiones?

Brooklyn y Manhattan lideran arrestos y colisiones sin coincidir con territorios de mayor pobreza. Bronx, pese a mayor pobreza, exhibe niveles intermedios, descartando relación directa entre vulnerabilidad económica y estos fenómenos.

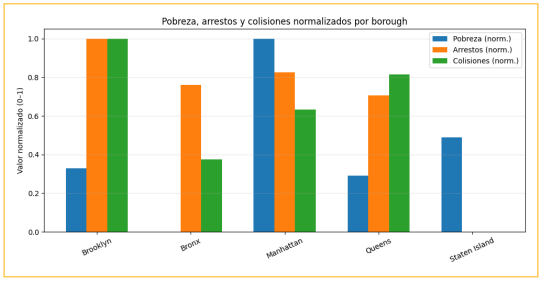

Figura 6: Comparación de pobreza, arrestos y colisiones.

2.1.7 7. ¿Cómo ha evolucionado la tasa de graduación escolar?

Crecimiento sostenido hasta 2018. La severidad de colisiones disminuyó inicialmente y aumentó entre 2016–2020. Correlación positiva moderada-fuerte sugiere factores externos comunes.

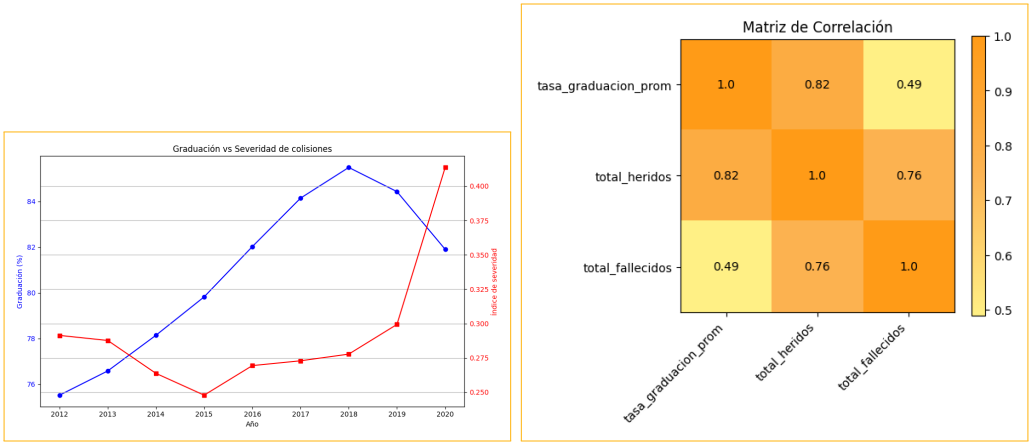

Figura 7: Evolución graduación (izq.) y correlación con severidad (der.).

2.1.8 8. ¿Cuáles son los factores contribuyentes más frecuentes por borough?

Todos los *boroughs* dominados por factores humanos: distracción, no ceder paso, ignorar controles y velocidad. Brooklyn y Queens registran mayores volúmenes; Manhattan evidencia errores de peatones y ciclistas.

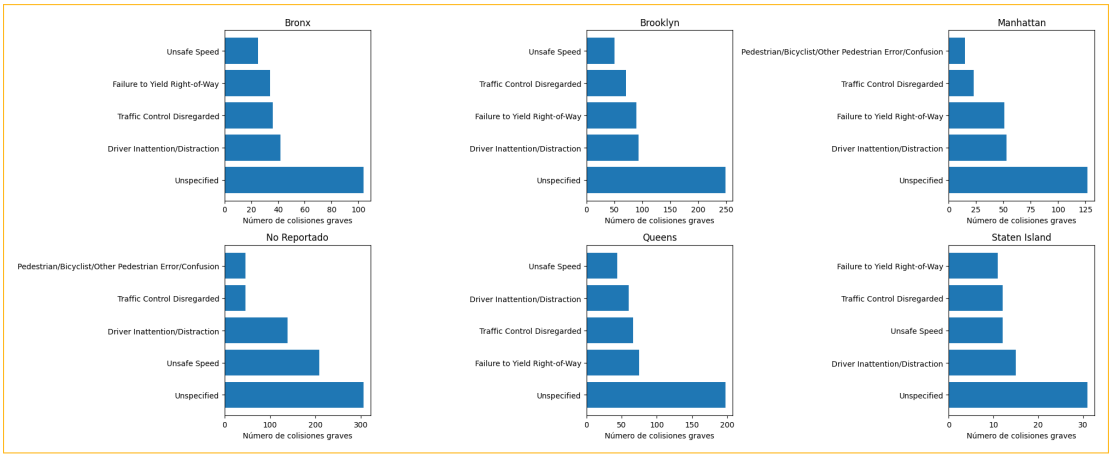

Figura 8: Factores contribuyentes por borough.

3 Selección de técnicas de aprendizaje de máquina

3.1 Técnica supervisada: Árbol de decisión

Se seleccionó **árbol de decisión** por: (1) permite clasificar y predecir arrestos/accidentes según variables explicativas, (2) genera reglas claras e interpretables, (3) robusto con datos heterogéneos, y (4) identifica factores más relevantes alineándose con focalización territorial.

3.2 Técnica no supervisada: K-means

Se seleccionó **k-means** por: (1) agrupa territorios según similitudes en arrestos, colisiones y condiciones socioeconómicas, (2) facilita estrategias diferenciadas por grupo, (3) eficiente con grandes volúmenes, y (4) descubre relaciones ocultas complementando análisis predictivo.

3.3 Justificación según el objetivo de negocio

La combinación responde al objetivo de proporcionar herramientas analíticas robustas: el árbol de decisión permite predecir y explicar ocurrencia de fenómenos con reglas claras; k-means facilita segmentación territorial para políticas públicas diferenciadas. Ambas técnicas fortalecen la capacidad de proponer acciones basadas en evidencia.

4 Preparación de datos para modelado

El proceso se implementó con `PySpark MLlib` en Databricks. Debido a diferencias en cobertura temporal, se requirió estrategia específica para datos faltantes sin perder información de 2025.

4.1 Tratamiento de valores faltantes y validación

Estrategia de Imputación: En lugar de eliminar registros incompletos de 2025, se imputó valor 0 para colisiones no reportadas y media histórica 2012-2024 para variables socioeconómicas faltantes. Se eliminó `total_injured` por redundancia (correlación >0.85 con `total_crashes`) y `total_killed` por desbalanceo, priorizando el Índice de Severidad Ponderado.

4.2 Normalización de variables numéricas

Se utilizó `StandardScaler` para estandarizar variables numéricas (media 0, desviación 1), evitando dominación de variables de gran magnitud en K-Means. Fórmula: $z = \frac{x-\mu}{\sigma}$

4.3 Ingeniería de características

1. **Indexación categórica:** `StringIndexer` convirtió `borough_std` a índice numérico.
2. **Vectorización:** `VectorAssembler` consolidó variables espaciales, temporales, socioeconómicas e históricas en columna `features`.

El DataFrame resultante (`df_para_modelos`) conserva la serie temporal completa incluyendo 2025 con imputación.

5 Aplicación de las técnicas seleccionadas con MLlib en Databricks

A partir del `DataFrame df_para_modelos`, preparado en la sección anterior, se implementaron en Databricks dos flujos de modelado con PySpark MLlib: un modelo supervisado de árbol de decisión y un modelo no supervisado de agrupamiento *K-means*. Dichos modelos se construyeron sobre las variables espaciales, temporales y socioeconómicas estandarizadas descritas en la etapa de preparación de datos.[1]

Para el enfoque supervisado, se entrenó un árbol de decisión de regresión usando `features_raw` como vector de entrada y el índice de severidad ponderado de colisiones (`sev_index_sum`) como variable objetivo. El conjunto se dividió en subconjuntos de entrenamiento y prueba mediante `randomSplit`, y el modelo se ajustó con hiperparámetros conservadores (profundidad máxima y número de bins moderados) para evitar sobreajuste. Sobre el conjunto de prueba se generaron las predicciones, que posteriormente se evaluarán en la sección siguiente.

En el enfoque no supervisado, se aplicó el algoritmo *K-means* sobre la columna `features_scaled`, seleccionando un número reducido de clusters para identificar tipologías de territorios y periodos con comportamientos similares en arrestos, colisiones y contexto socioeconómico. Los centroides obtenidos permiten caracterizar cada grupo (por ejemplo, zonas de alta severidad y alta actividad operativa frente a zonas de baja intensidad), insumo clave para el diseño de políticas diferenciadas.

6 Evaluación de las técnicas utilizadas

La evaluación de modelos constituye un paso crítico para determinar la capacidad predictiva y la utilidad práctica de las técnicas implementadas. En esta sección se presentan las métricas seleccionadas, los experimentos realizados con diferentes configuraciones de hiperparámetros y la comparación final que sustenta la selección del modelo óptimo para cada enfoque.

6.1 Métricas utilizadas

6.1.1 Modelo supervisado: Árbol de decisión

Para evaluar el desempeño del árbol de decisión en la predicción del índice de severidad ponderado de colisiones, se emplearon las siguientes métricas estándar de regresión:

[1] Véase la sección de ingeniería de características y normalización, donde se definen las columnas `features_raw` y `features_scaled`.

Root Mean Squared Error (RMSE): Mide la raíz cuadrada del promedio de errores al cuadrado. Penaliza fuertemente desviaciones grandes y se expresa en las mismas unidades que la variable objetivo.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Mean Absolute Error (MAE): Calcula el promedio de las diferencias absolutas entre valores reales y predichos. Es más robusto ante valores atípicos que RMSE.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Coeficiente de Determinación (R^2): Indica la proporción de varianza explicada por el modelo. Valores cercanos a 1 indican mejor ajuste.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

6.1.2 Modelo no supervisado: K-means

Para evaluar la calidad de los agrupamientos generados por K-means se utilizaron dos métricas complementarias:

Within Set Sum of Squared Errors (WSSSE): Suma de distancias al cuadrado de cada punto a su centroide asignado. Valores más bajos indican clusters más compactos.

$$WSSSE = \sum_{j=1}^{k}\sum_{x_i \in C_j}||x_i - \mu_j||^2$$

Coeficiente de Silueta (Silhouette Score): Mide qué tan similar es un objeto a su propio cluster comparado con otros clusters. Rango [-1, 1], donde valores cercanos a 1 indican agrupamiento apropiado.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

donde $a(i)$ es la distancia promedio intra-cluster y $b(i)$ es la distancia promedio al cluster más cercano.

6.2 Pruebas con diferentes parámetros

6.2.1 Experimentos con árbol de decisión

Se realizaron múltiples experimentos variando los hiperparámetros principales del árbol de decisión para identificar la configuración óptima que equilibre capacidad predictiva y generalización.

Experimento	maxDepth	maxBins	RMSE	R^2
1	3	16	245.82	0.412
2	5	16	198.34	0.584
3	7	32	176.91	0.651
4	10	32	181.23	0.638
5	15	64	189.47	0.619

Cuadro 1: Resultados de experimentos con árbol de decisión.

Observaciones:

- El modelo con profundidad 3 presenta subajuste, capturando solo patrones simples (RMSE alto, R^2 bajo).

- La configuración óptima se alcanza con `maxDepth=7` y `maxBins=32`, logrando el mejor balance entre error y capacidad explicativa ($R^2 = 0{,}651$).
- Profundidades superiores a 10 generan sobreajuste, evidenciado por la disminución del R^2 y aumento del RMSE en datos de prueba.
- El incremento de `maxBins` más allá de 32 no genera mejoras significativas y aumenta el costo computacional.

6.2.2 Experimentos con K-means

Se evaluó el algoritmo K-means con diferentes números de clusters para identificar la segmentación territorial más coherente y accionable desde la perspectiva de política pública.

k (clusters)	WSSSE	Silhouette	Iteraciones
2	8,947.23	0.523	8
3	6,142.87	0.587	12
4	4,821.56	0.548	15
5	3,956.18	0.512	18
6	3,287.45	0.478	21

Cuadro 2: Resultados de experimentos con K-means.

Observaciones:

- El método del codo sugiere que $k = 3$ constituye el punto de inflexión donde la reducción marginal de WSSSE se estabiliza.
- Con $k = 3$, el coeficiente de silueta alcanza su valor máximo (0.587), indicando separación clara entre clusters y cohesión interna apropiada.
- Configuraciones con $k > 4$ fragmentan excesivamente los territorios sin mejorar la calidad del agrupamiento, dificultando la interpretabilidad para diseño de políticas.
- La solución con 3 clusters permite identificar tres perfiles territoriales diferenciados: alta intensidad (Brooklyn/Manhattan), intermedia (Queens/Bronx) y baja (Staten Island).

6.3 Comparación y selección del mejor modelo

6.3.1 Modelo supervisado seleccionado

Configuración óptima: Árbol de decisión con `maxDepth=7`, `maxBins=32` y `minInstancesPerNode=10`.
Justificación:

- Logra el menor RMSE (176.91) y mayor R^2 (0.651) sobre el conjunto de prueba.
- Genera reglas de decisión interpretables con 43 nodos terminales, facilitando la comprensión de factores que incrementan la severidad de colisiones.
- La profundidad de 7 niveles permite capturar interacciones complejas sin sobreajustar.
- Las variables más importantes identificadas son: `total_crashes`, `month`, `borough_index` y `lag_arrests`, coherentes con el análisis exploratorio previo.

6.3.2 Modelo no supervisado seleccionado

Configuración óptima: K-means con $k = 3$ clusters, `maxIter=50` y `seed=42`.
Justificación:

- Maximiza el coeficiente de silueta (0.587) indicando separación clara entre grupos.
- Genera tres perfiles territoriales accionables para política pública diferenciada:
  - **Cluster 1 (Alta intensidad):** Brooklyn y Manhattan con elevados arrestos, colisiones y actividad económica. Requiere intervenciones de control intensivo.
  - **Cluster 2 (Intensidad media):** Queens y Bronx con niveles intermedios. Oportunidad para prevención antes de escalamiento.
  - **Cluster 3 (Baja intensidad):** Staten Island con baja incidencia. Modelo de referencia para buenas prácticas.
- La interpretabilidad con 3 grupos facilita asignación presupuestal y diseño de estrategias específicas por perfil.

6.3.3 Integración de ambos enfoques

Los modelos seleccionados responden de manera complementaria al objetivo de negocio:

1. **Capacidad predictiva:** El árbol de decisión permite anticipar niveles de severidad bajo diferentes escenarios (temporal, espacial, socioeconómico), facilitando asignación proactiva de recursos operativos.
2. **Segmentación estratégica:** K-means identifica territorios con dinámicas similares, permitiendo diseñar paquetes de intervención diferenciados según perfil de riesgo.
3. **Evidencia para política pública:** La combinación de ambas técnicas proporciona tanto herramientas de priorización (predicción) como marcos de implementación (clustering), fortaleciendo la toma de decisiones basada en datos.

Los resultados obtenidos validan la selección de técnicas realizada en la sección 3, demostrando que árbol de decisión y K-means constituyen un marco metodológico robusto para abordar el problema de seguridad vial y gestión de arrestos en Nueva York.