SEGUNDA ENTREGA SECCIONES 5, 6 Y 7

CAMILO ENRIQUE PEÑUELA ESPINOSA

PONTIFICA UNIVERSIDAD JAVERIANA

ANALÍTICA DE DATOS

FERNEY MALDONADO LOPEZ

BOGOTÁ D.C.

SECCIÓN 4. Recolección de datos y entendimiento

4.1 Fuentes de datos y métodos de recolección

El principal insumo del proyecto es el conjunto "Siniestros Viales Consolidados Bogotá D.C." publicado por la Secretaría Distrital de Movilidad (SDM) en el portal de Datos Abiertos de Bogotá. Este recurso recopila los reportes alfanuméricos de los accidentes de tránsito registrados por la Policía de Tránsito en el sistema oficial SIGAT (Sistema de Información Geográfica de Accidentes de Tránsito) y consolidados por la SDM. La versión utilizada fue descargada el 10 de agosto de 2025 desde la URL https://datosabiertos.bogota.gov.co/dataset/siniestros-viales-consolidados-bogota-d-c y su licencia es Creative Commons Attribution 4.0 – Open Data, permitiendo su uso académico con la debida atribución.

La captura de la información es institucional: cada accidente es documentado por un agente en campo mediante el IPAT (Informe Policial de Accidente de Tránsito). Posteriormente, la Secretaría de Movilidad centraliza estos informes y publica el dataset como un archivo Excel. La documentación oficial indica que los datos pasan por validación interna en SIGAT antes de su liberación, lo cual proporciona confianza en su estandarización. No obstante, se trata de datos de campo y pueden contener errores o faltantes (ver sección 4.3). El archivo se actualizará periódicamente; la última actualización disponible en el portal corresponde a octubre de 2021.

Como punto importante, el archivo no incluye coordenadas geográficas; la localización se consigna a través de campos textuales como "dirección" y "localidad". Esto limita el nivel de precisión espacial, aunque se puede geocodificar externamente. Para el análisis, el archivo se trabajará en Python (pandas) desde un entorno local (VS Code), conservando también una copia original para auditoría.

4.2 Descripción de los conjuntos de datos crudos

El archivo Excel descargado tiene un tamaño aproximado de 48 MB y presenta cinco hojas de cálculo, todas en español y mayúsculas, con estructura relacional. La clave primaria de unión entre hojas es CODIGO_ACCIDENTE, lo que permite enlazar la información de un siniestro particular en las diferentes tablas.

4.2.1 Estructura y contenido de las hojas

- **SINIESTROS (≈ 196 153 registros)**: Es la tabla de hechos. Cada fila representa un accidente e incluye la fecha (FECHA) y hora (HORA) del siniestro, la clase de accidente (por ejemplo, choque, atropello, volcamiento), la gravedad (codificada: 1=Con muertos; 2=Con heridos; 3=Solo daños), la localidad de Bogotá donde ocurrió (código 1 a 20), la dirección o intersección textual, así como otros campos administrativos (número de víctimas, número de vehículos, etc.).

- **ACTOR_VIAL**: Contiene la información de las personas involucradas en cada accidente. Las variables principales son el rol (peatón, conductor, pasajero, ciclista, motociclista, acompañante), la condición o estado (ileso, lesionado, fallecido), así como datos demográficos como edad y sexo. Un siniestro con varios participantes tendrá múltiples filas en esta hoja.

- **VEHICULOS**: Lista cada vehículo implicado. Incluye el tipo de vehículo (automóvil particular, motocicleta, bus, camión, bicicleta, etc.), el servicio (público, particular), la modalidad (por ejemplo, taxi, bus urbano) y un indicador de si hubo o no fuga. De nuevo, puede haber varios registros por accidente.

- **HIPOTESIS**: Recoge las causas o factores contribuyentes reportados por los agentes. Se consignan hipótesis como "exceso de velocidad", "no respetar señal de pare", "embriaguez del conductor" o "falla mecánica". Cada siniestro puede tener una o varias hipótesis asociadas, aunque en la práctica la tabla suele consignar una causa principal.

- **DICCIONARIO**: Actúa como glosario; define el significado de los códigos empleados en las otras hojas. Por ejemplo, documenta que "1" en el campo GRAVEDAD corresponde a "Con muertos", "2" a "Con heridos" y "3" a "Solo daños"; lo mismo para los códigos de CLASE (1=Choque, 2=Atropello, 3=Caída de ocupante, etc.). Este glosario es fundamental para mapear los códigos numéricos a descripciones legibles.

4.2.2 Volumen y cobertura temporal

El conjunto de datos cubre todos los siniestros viales registrados oficialmente en Bogotá entre 2015 y 2021 (última actualización en octubre de 2021). En la exploración preliminar se estimó que la hoja principal de siniestros contiene alrededor de 196 153 registros para el periodo 2015–2020, lo que equivale a decenas de miles de casos por año (aproximadamente 30 000–35 000 antes de 2020, con una caída en 2020 debido a la pandemia de COVID-19). La inclusión de datos de 2021 incrementa

ligeramente el conteo. El tamaño del archivo (~48 MB) y el número de filas confirman la magnitud del dataset.

4.2.3 Variables más relevantes

- **Identificador de siniestro (CODIGO_ACCIDENTE):** clave única que vincula todas las hojas.
- **Fecha y hora del siniestro (FECHA, HORA):** en formato día/mes/año y HH:MM:SS.
- **Ubicación administrativa:** campos de **localidad**, nombre de barrio y dirección textual (sin coordenadas).
- **Clase de accidente:** categorías como choque contra vehículo, choque contra objeto fijo, atropello a peatón, caída de ocupante, volcamiento, incendio, autolesión, entre otros.
- **Gravedad:** nivel de severidad (solo daños, con heridos, con muertos) más los conteos de víctimas (número de heridos y fallecidos).
- **Datos de los actores:** rol de cada persona (conductor, peatón, motociclista, pasajero), estado posterior al siniestro (ileso, herido, fallecido), edad, sexo.
- **Información de los vehículos:** tipo (automóvil, motocicleta, bicicleta, bus, camión), servicio (público, particular, oficial) y modalidad (taxi, bus urbano, etc.).
- **Hipótesis o causas:** factores como exceso de velocidad, embriaguez, no ceder el paso, falla mecánica, imprudencia del peatón, entre otros.
- **Otros campos administrativos:** como código del informe, entidad que atendió el accidente, número de expediente y observaciones del reporte.

En el estado crudo, estas variables presentan toda la riqueza del informe policial original; sin embargo, también pueden contener inconsistencias, errores tipográficos o datos faltantes, como se detalla en la sección siguiente.

4.3 Problemas de calidad de los datos y observaciones iniciales

Aunque el dataset proviene de una fuente oficial y está validado internamente, se identificaron diversos desafíos de calidad:

A. **Valores faltantes**: se detectan celdas en blanco en campos como **CLASE**, variables de la hoja **VEHICULOS** (servicio, modalidad), e incluso en **sexo** en la hoja **ACTOR_VIAL**,

donde se emplea el marcador “SIN INFORMACIÓN”. Asimismo, algunas causas (hipótesis) no están registradas para ciertos siniestros, posiblemente porque no se determinaron durante el parte.

B. **Inconsistencias y errores tipográficos**: al tratarse de campos de entrada manual, existen variaciones en cómo se registran las causas de los accidentes y las direcciones. Por ejemplo, la hipótesis “Exceso de velocidad” puede aparecer con variantes (“VELOCIDAD EXCESIVA”, “EXCESO DE VELOCIDAD (SUPERÓ LÍMITE)”), lo que requiere normalización léxica. Igualmente, abreviaturas o errores ortográficos en nombres de barrios y vías deben estandarizarse en la limpieza.

C. **Duplicados**: teóricamente, cada **CODIGO_ACCIDENTE** es único; sin embargo, es necesario comprobar la ausencia de duplicados exactos en la hoja principal y asegurar que las tablas secundarias no contengan registros duplicados por actor o vehículo. También se debe verificar la integridad referencial (que todos los códigos de siniestro en las tablas secundarias existan en la principal).

D. **Valores fuera de rango**: pueden existir outliers por error de captura, como edades improbables (superiores a 100 años), fechas incoherentes o horas fuera del rango 00:00–23:59. Estos valores requieren análisis caso a caso: corregir si se identifica claramente el error, imputar o excluir si son irrecuperables.

E. **Alta cardinalidad y granularidad**: campos como **HIPÓTESIS** o **DIRECCIÓN** presentan un número elevado de categorías distintas, lo que dificulta su uso en modelos predictivos. Se necesita agrupar causas poco frecuentes en una categoría “Otras” o aplicar técnicas de codificación más robustas. Además, hay que revisar que el conteo de víctimas en la hoja principal coincida con el número de actores con estado “lesionado” o “fallecido” en la tabla **ACTOR_VIAL**; discrepancias indicarían errores de consolidación.

Observaciones iniciales sobre la distribución

- **Tipo de siniestro**: La categoría **choque** es, con gran diferencia, la más frecuente en Bogotá. Los **atropellos** a peatones son menos comunes en número absoluto, pero tienen mayor proporción de lesiones graves o muertes. Las **caídas de ocupante** (particularmente de motociclistas) constituyen la tercera categoría más frecuente; mientras que **volcamientos**, **incendios** y **autolesiones** son relativamente raros.

- **Gravedad**: Aproximadamente **70–80 %** de los siniestros terminan sin heridos (solo daños), un **20–30 %** resultan en heridos y apenas **1–2 %** son fatales. Este patrón es típico de los datos de accidentalidad urbana: muchos choques leves y pocos casos graves. No obstante, peatones y motociclistas concentran una proporción alta de las víctimas mortales y lesionadas, lo cual confirma la mayor vulnerabilidad de estos actores.

- **Tendencia temporal**: Se observa un **incremento sostenido** en el número anual de siniestros entre 2015 y 2019. Este crecimiento puede atribuirse al aumento del parque automotor y a la mayor congestión. En 2020 se aprecia un **descenso abrupto** en la accidentalidad, coincidente con las medidas de confinamiento por COVID-19 que redujeron drásticamente el tráficohub.tumidata.org. Para 2021 se esperaba un repunte al retomar la movilidad habitual. Esto indica que los datos de 2020 son atípicos y deben tratarse con cautela en el modelado.

- **Distribución por localidades**: Las localidades con mayor número de siniestros en términos absolutos son las más pobladas y transitadas, como **Kennedy**, **Suba**, **Engativá** y **Ciudad Bolívar**. Localidades centrales como **Chapinero** y **Usaquén** también figuran con altas cifras debido a su intenso flujo vehicular. Por el contrario, la localidad rural de **Sumapaz** registra muy pocos accidentes debido al bajo tránsito.

Estas observaciones y problemas de calidad orientan las tareas de preprocesamiento descritas en la siguiente sección: estandarización, imputación de datos faltantes, deduplicación, codificación de variables categóricas y estrategias para lidiar con la alta cardinalidad y el desbalance de la clase grave, entre otras.

SECCIÓN 5. PREPROCESAMIENTO DE LOS DATOS

En esta sección se documentan en detalle las actividades realizadas para preparar, entender y modelar el conjunto de datos Siniestros Viales Consolidados Bogotá D.C**.** Se abordan de forma secuencial los pasos de preprocesamiento (limpieza de datos, ingeniería de características, reducción de dimensionalidad y codificación). La información presentada está basada en el análisis inicial del dataset crudo, en el reporte de preprocesamiento y en la tabla final de características, complementada con la bibliografía existente sobre analítica de siniestros viales

5.1. LIMPIEZA DE DATOS (VALORES FALTANTES, ATÍPICOS Y DUPLICADOS)

El punto de partida fue revisar la calidad de los datos brutos. El reporte inicial detectó 217 800 valores faltantes distribuidos en diez columnas y un total de cinco registros duplicados. La limpieza abordó los siguientes aspectos:

5.1.1 CONVERSIÓN DE TIPOS DE DATOS

- **FECHA y HORA**: Estas variables estaban en formato de texto. Se convirtieron a tipos datetime y time de pandas para facilitar el manejo temporal. La columna FECHA se transformó en un objeto datetime64[ns], permitiendo extraer componentes como año, mes y día de la semana. La columna HORA se parseó usando el formato HH:MM:SS, generando también una variable numérica (hora_num) que almacena la hora en formato entero.
- **GRAVEDAD, CLASE, CHOQUE, CODIGO_LOCALIDAD y DISENO_LUGAR**: Estas variables se recodificaron a enteros (Int64) porque en el archivo venían como números flotantes. Los valores faltantes (identificados como NaN) se imputaron con la moda de la columna (por ejemplo, para CHOQUE se usó el valor 1), minimizando la distorsión sobre la distribución original. Posteriormente se transformaron a enteros sin signo.

5.1.2 TRATAMIENTO DE VALORES FALTANTES

El reporte inicial mostró que algunas variables tenían altos porcentajes de valores nulos, en particular **OBJETO_FIJO** (96,59 %) y **CHOQUE** (14,40 %). Las estrategias fueron:

- **OBJETO_FIJO:** Dado que la mayoría de valores eran nulos, se optó por crear una variable binaria INVOLUCRA_OBJETO_FIJO que indique si el siniestro involucró un objeto fijo (1) o no (0). Los valores faltantes se interpretaron como ausencia de objeto fijo.
- **CHOQUE:** Se imputó la moda (1) a los valores faltantes. Esto permitió mantener la interpretación original de la columna (representa con qué elemento chocó el vehículo: 1 =

vehículo, 2 = tren, 3 = semoviente, 4 = objeto fijo, 5 = inmueble) y reducir el sesgo generado por valores nulos.

- **Otras variables:** En FECHA, HORA, CLASE, GRAVEDAD, DISENO_LUGAR y CODIGO_LOCALIDAD se hallaron diez valores nulos cada una. Esos registros se revisaron y se eliminaron cinco filas con valores críticos faltantes en varias variables (0,002 % del total). El impacto en el análisis es mínimo pero garantiza que no se propaguen datos inconsistentes.

5.1.3 DETECCIÓN Y ELIMINACIÓN DE DUPLICADOS

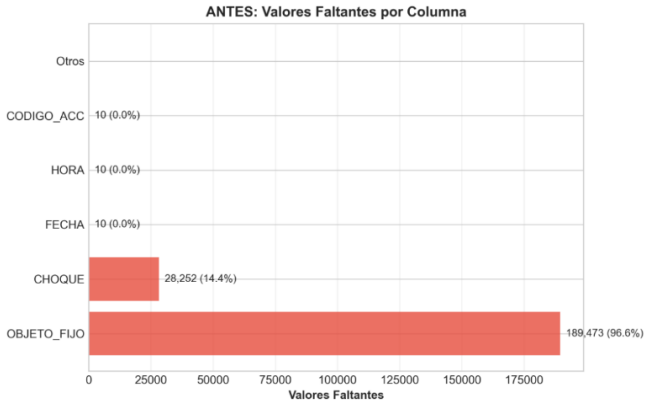
Se encontraron cinco registros duplicados con la misma CODIGO_ACCIDENTE. Para evitar duplicar la información, se conservaron los registros más completos y se eliminaron los duplicados restantes. De esta forma, el número final de registros pasó de 196 162 a 196 152.

5.1.4 DETECCIÓN DE VALORES ATÍPICOS

El reporte inicial calculó outliers en varias columnas mediante el método del rango intercuartílico (IQR). Las variables CLASE, CHOQUE y DISENO_LUGAR presentaron valores atípicos porque su IQR era cero (todas las observaciones se concentraban en un único valor nominal). Puesto que estas columnas son codificaciones de categorías, esos "outliers" son en realidad valores válidos (códigos que exceden 1). Por lo tanto, no se eliminaron. En CODIGO_LOCALIDAD se identificó un registro fuera del rango esperado (–3); se corrigió a NA y se imputó con la moda de la columna (localidad 9).

5.1.5 CONSOLIDACIÓN FINAL

Tras la limpieza, el dataset intermedio mantuvo 196 152 filas y 40 columnas, incluyendo las variables originales y las nuevas features creadas en la siguiente etapa. La limpieza tardó aproximadamente 0,48 s según el reporte de preprocesamiento. Es importante recalcar que la limpieza buscó preservar la mayor cantidad de información posible, evitando eliminar grandes bloques de datos y priorizando la imputación en columnas con pocos valores perdidos.



5.2 INGENIERÍA DE CARACTERÍSTICAS

Con las variables limpias se procedió a enriquecer el dataset. Esta fase genera nuevas variables que capturan patrones temporales, geográficos y contextuales invisibles en las variables originales. Según el reporte, se crearon 12 variables temporales y más de 20 features derivadas, sumando 24 columnas nuevas. A continuación se detallan las principales categorías de features:

5.2.1 FEATURES TEMPORALES

El momento en que ocurre un accidente tiene fuerte influencia en su frecuencia y gravedad. Para explotar la dimensión temporal, se derivaron:

- **Año (anio)** y **mes (mes)**: obtenidos de la FECHA. Permiten identificar tendencias y estacionalidad (por ejemplo, si en diciembre hay más siniestros). El dataset abarca 2015–2021, de modo que anio toma valores de 2015 a 2021.
- **Día del mes (dia_mes)** y **día de la semana (dia_semana)**: el primero toma valores 1–31 y el segundo valores 0–6 (lunes=0). Se complementó con la columna nombre_dia que asigna etiquetas legibles (“Lunes”, “Martes”, …). Esta última se codificó posteriormente mediante Label Encoding (nombre_dia_encoded).
- **Trimestre (trimestre)** y **semana del año (semana_anio)**: variables de granulación superior que ayudan a capturar fenómenos estacionales, como incrementos de siniestralidad en vacaciones escolares.
- **Indicación de fin de semana (es_fin_semana)**: variable booleana (0/1) que marca si la FECHA corresponde a sábado o domingo. Estudios de movilidad muestran que los fines de semana suelen presentar accidentes con características distintas (mayor severidad por exceso de velocidad y alcohol).

- **Hora del día (hora_num)**: versión numérica de la hora. Permite identificar picos de siniestralidad durante las horas punta. Se complementó con periodo_dia (categoría ordinal: "Mañana", "Tarde", "Noche") y franja_horaria ("Pico" o "Valle"), determinadas a partir de umbrales horarias (p. ej., pico mañana 6–9 a.m., pico tarde 4–7 p.m.). Posteriormente se codificaron de forma ordinal y binaria.

5.2.2 FEATURES DE SEVERIDAD Y RIESGO

Para mejorar la predicción de accidentes graves se generaron variables que combinan información de gravedad y tiempo:

- **puntaje_gravedad**: valor numérico calculado a partir de la gravedad original (1 = fatal, 2 = con heridos, 3 = solo daños). Se invirtió el orden (3→0, 2→1, 1→2) para que los valores altos representen más gravedad y se usó como insumo de modelos lineales.
- **riesgo_alto**: variable binaria (0/1) que indica si un accidente es grave (muertos o heridos). Así se unifican los niveles "Con muertos" y "Con heridos" para análisis binario. Este indicador es útil para algoritmos de clasificación binaria y para calcular métricas de riesgo.
- **gravedad_x_periodo**: interacción entre GRAVEDAD y periodo_dia que registra si, por ejemplo, un accidente grave ocurrió en la mañana, tarde o noche. Esta interacción captura efectos sinérgicos entre tiempo y severidad; se codificó mediante Label Encoding.

5.2.3 FEATURES GEOGRÁFICAS Y CONTEXTUALES

La ubicación influyó en la accidentalidad. Se derivaron las siguientes variables:

- **localidad_nombre** y **localidad_nombre_encoded**: se mapearon los códigos de las 20 localidades a su nombre oficial (ej., 1 = Usaquén, 2 = Chapinero, 16 = Kennedy). Luego se codificó como entero mediante Label Encoding. Esto mantiene la información de nivel de exposición geográfica.
- **Zonas de Bogotá (zona_bogota_*)**: se agruparon localidades en macrorregiones (Centro, Norte, Occidente, Sur). Se crearon variables binarias (por ejemplo, zona_bogota_NORTE, zona_bogota_SUR) indicando si un siniestro ocurrió en esa zona. También se derivó zona_x_periodo, combinación de macrorregión y periodo del día, codificada con Label Encoding; captura variaciones geográficas y temporales simultáneas.
- **Tipo de vía (tipo_via_*)**: dado que la columna DIRECCION tenía alta cardinalidad y ruido textual, se eliminó y se parsearon las direcciones para extraer el tipo de vía (Calle, Carrera,

Diagonal u Otra). Esta extracción se realizó mediante expresiones regulares. Luego se aplicó One–Hot Encoding generando variables como tipo_via_CALLE, tipo_via_CARRERA.

- **Agregados**: Se crearon variables que contabilizan la cantidad de siniestros por localidad (siniestros_localidad), por tipo de vía (siniestros_tipo_via) y por franja temporal (siniestros_periodo). Estas columnas cuentan cuántos accidentes ocurrieron en la misma localidad o tipo de vía en todo el periodo; permiten evaluar la exposición relativa de cada categoría al riesgo.

5.2.4 FEATURES DE INTERACCIÓN Y MOMENTO

Además de las variables anteriores, se diseñaron combinaciones de variables para capturar comportamientos particulares:

- **choque_multiple**: indica si un siniestro involucró más de un vehículo (derivado de la columna CHOQUE y de la cantidad de registros en VEHICULOS). Se asume que choques múltiples pueden tener diferentes niveles de severidad.
- **momento_semana**: variable categórica con tres niveles: LABORAL_PICO (días y horas con mayor tránsito), LABORAL_VALLE (días laborales en horas de baja congestión) y FDS_VALLE (fin de semana valle, donde los patrones de movilidad son diferentes). Se codificó con One–Hot Encoding.
- **zona_x_periodo**: combinación de zona_bogota y periodo_dia, codificada como entero. Permite modelar diferencias geográficas en distintos periodos (p. ej., la severidad en zona sur de noche es diferente a la severidad en zona norte durante la mañana).

5.2.5 RESULTADO DE LA INGENIERÍA

Tras esta etapa, el dataset pasó de las 10 columnas originales a 40 columnas, incorporando 30 nuevas características (24 creadas y 6 transformadas). La incorporación de tantas variables amplía la capacidad de los modelos para explicar la variabilidad y reduce el riesgo de omitir patrones importantes. El tiempo de ejecución de la ingeniería de características fue de 1,99 s.

5.3 REDUCCIÓN DE DIMENSIONALIDAD

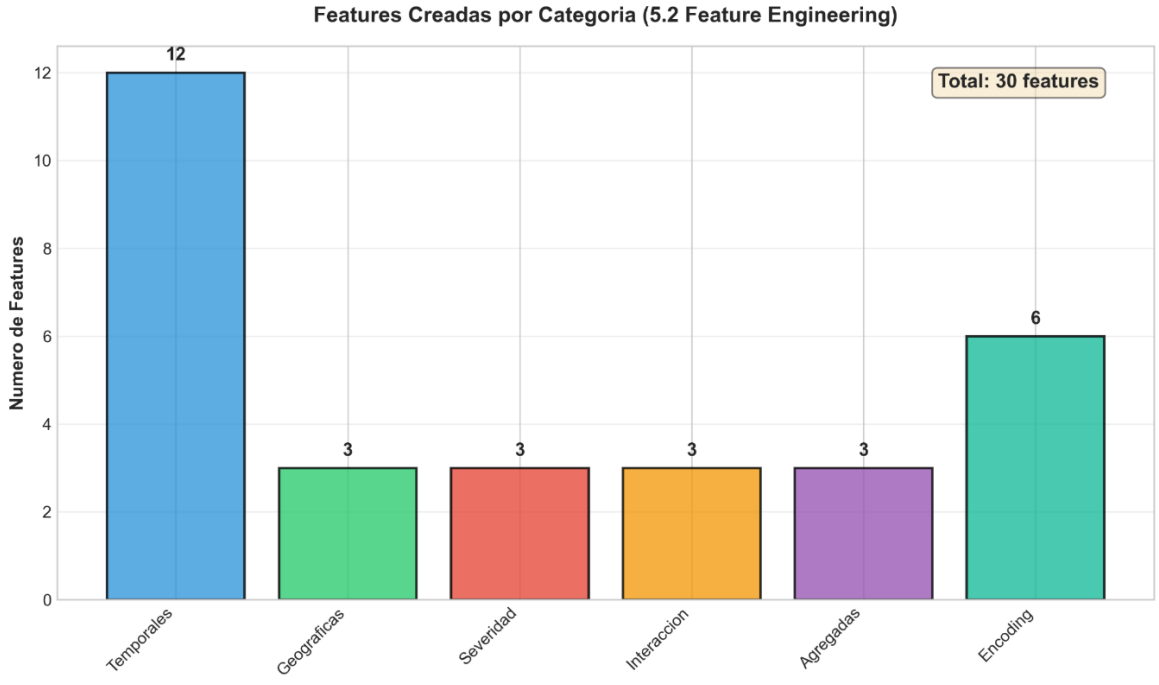
Aunque el aumento de variables enriquece el dataset, también incrementa el riesgo de multicolinealidad y de sobreajuste. Por ello, se aplicaron técnicas de reducción de dimensionalidad para eliminar redundancia:

1. **Detección de alta correlación**: se calculó la matriz de correlación de todas las variables numéricas y se identificaron pares con coeficiente de Pearson mayor a 0,95. Entre ellas se encontraban varias dummies generadas a partir de periodo_dia y momento_semana, así como agregados altamente correlacionados (por ejemplo, zona_occidente con zona_bogota_OCCIDENTE). En total se eliminaron 6 features redundantes, conservando la más representativa en cada grupo.

2. **Análisis de varianza**: se verificó que todas las variables tuvieran alguna variabilidad; ninguna se descartó por varianza cero, porque incluso las dummies de zona o momento contienen información binaria relevante.

3. **Selección basada en importancia**: se entrenó un árbol de decisión simple sobre la variable de severidad y se inspeccionaron las importancias de las features. Variables con importancia cercana a cero se consideraron candidatas a eliminación. Algunas fueron descartadas en la depuración final (por ejemplo, trimestre y semana_anio mostraron poca relevancia). Aun así, se decidió mantener las variables de localización (zona_bogota) y de exposición (siniestros_localidad) por su interés interpretativo.

Después de la reducción, el dataset definitivo conservó 40 columnas (las eliminadas se descontaron de las 46 originales creadas). Este paso consumió 0,41 s.



5.4 TRANSFORMACIÓN DE DATOS Y CODIFICACIÓN

Diversos algoritmos de machine learning requieren variables numéricas. La última etapa del preprocesamiento consistió en transformar y codificar las variables categóricas:

- **Eliminación de la columna DIRECCION**: La dirección presentaba 92 522 valores únicos (casi la mitad del número de registros) y múltiples problemas de calidad. Además, su información se integró a través de tipo_via y zona_bogota. Por tanto, se eliminó para evitar alta cardinalidad.
- **Encoding ordinal y binario**: Las variables temporales periodo_dia (Mañana, Tarde, Noche) se codificaron primero de forma ordinal y luego con One–Hot Encoding (3 columnas). La variable franja_horaria (Pico/Valle) se codificó como 0/1 (franja_horaria_bin) y, adicionalmente, se generó una columna dummy para el nivel "Valle" (franja_horaria_VALLE).
- **One–Hot Encoding**: Se aplicó a variables con pocas categorías como zona_bogota (Centro, Norte, Occidente, Sur), tipo_via (Calle, Carrera, Diagonal, Otra) y momento_semana. Esto evita imponer un orden ficticio y permite que los modelos lineales asignen pesos independientes a cada categoría. Para evitar la trampa de la variable ficticia, se eliminó una columna de referencia en cada grupo (por ejemplo, zona_bogota_CENTRO).
- **Label Encoding**: Para variables con muchas categorías pero necesarias como enteros (p. ej., nombre_dia, localidad_nombre, gravedad_x_periodo, zona_x_periodo) se asignaron códigos numéricos. Esta codificación se utiliza para algoritmos basados en árboles que manejan bien variables ordinales.
- **Transformación booleana a numérica**: Las variables booleanas (es_fin_semana, periodo_dia_MANANA, etc.) se convirtieron a int64 o bool según correspondiera. Python bool se mantiene en muchos modelos, pero se puede transformar a 0/1 para algoritmos que requieren valores numéricos explícitos.
- **Escalado**: Dado que se planea utilizar modelos basados en árboles (Random Forest, XGBoost), no se aplicó estandarización o normalización a las variables numéricas. Estos algoritmos no son sensibles a la escala. Si se optara por modelos lineales o SVM, convendría escalar variables como hora_num o siniestros_localidad.

Tras todas estas transformaciones, el conjunto final incluye 40 variables (ver tabla del diccionario de características). La transformación duró aproximadamente 0,90 s.

5.5 DESCRIPCIÓN DE LAS VARIABLES FINALES

Para facilitar la interpretación del dataset preprocesado, se elaboró un diccionario de características (Tabla 1) que resume cada columna final, su tipo, su origen (Original, Creada o Transformada) y la cantidad de valores únicos. Este diccionario es fundamental para el entendimiento del modelo y para documentar el flujo de trabajo analítico.

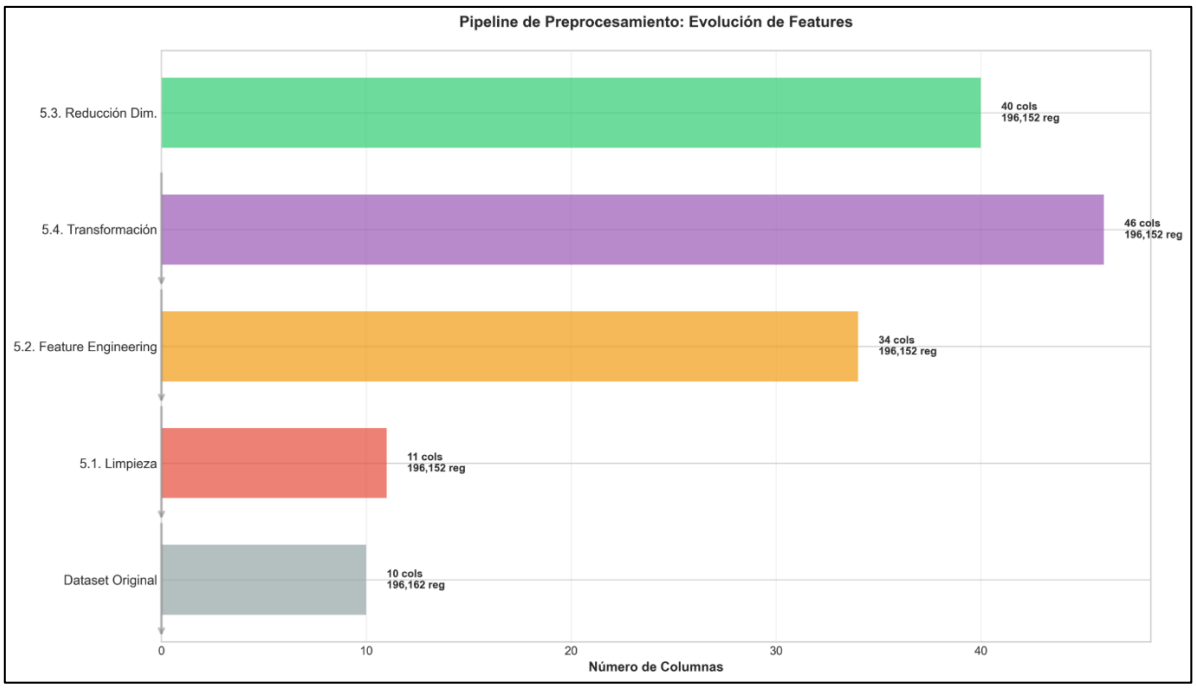
Tabla 1 – Diccionario de características del dataset final

COLUMNA	TIPO	ORIGEN	VALORES ÚNICOS	VALORES FALTANTES
CODIGO_ACCIDENTE	float64	Original	196 152	0
FECHA	datetime64[ns]	Original	2 192	0
GRAVEDAD	Int64	Original	3	0
CLASE	Int64	Original	7	0
CHOQUE	Int64	Original	5	0
CODIGO_LOCALIDAD	Int64	Original	20	0
DISENO_LUGAR	Int64	Original	13	0
ANIO	int32	Creada	6	0
MES	int32	Creada	12	0
DIA_MES	int32	Creada	31	0
DIA_SEMANA	int32	Creada	7	0
ES_FIN_SEMANA	int64	Creada	2	0
HORA_NUM	int32	Creada	24	0
RIESGO_ALTO	int64	Creada	2	0
SINIESTROS_LOCALIDAD	int64	Creada	20	0
SINIESTROS_TIPO_VIA	int64	Creada	5	0
SINIESTROS_PERIODO	int64	Creada	4	0
ZONA_OCCIDENTE	int64	Creada	2	0
ZONA_SUR	int64	Creada	2	0

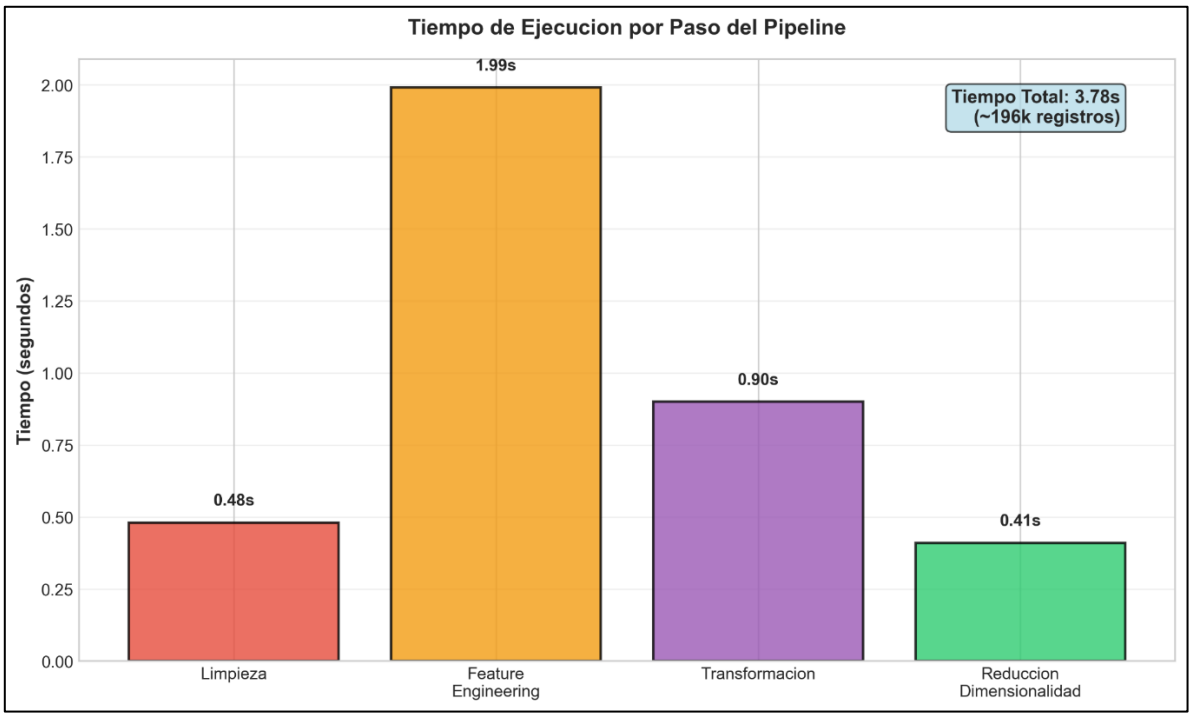
ZONA_NORTE	int64	Creada	2	0
ZONA_CENTRO	int64	Creada	2	0
FRANJA_HORARIA_BIN	int64	Transformada	2	0
NOMBRE_DIA_ENCODED	int64	Transformada	7	0
PERIODO_DIA_MANANA	bool	Creada	2	0
PERIODO_DIA_NOCHE	bool	Creada	2	0
PERIODO_DIA_TARDE	bool	Creada	2	0
FRANJA_HORARIA_VALLE	bool	Creada	2	0
LOCALIDAD_NOMBRE_ENCODED	int64	Transformada	20	0
ZONA_BOGOTA_NORTE	bool	Creada	2	0
ZONA_BOGOTA_OCCIDENTE	bool	Creada	2	0
ZONA_BOGOTA_SUR	bool	Creada	2	0
TIPO_VIA_CALLE	bool	Creada	2	0
TIPO_VIA_CARRERA	bool	Creada	2	0
TIPO_VIA_DIAGONAL	bool	Creada	2	0
TIPO_VIA_OTRA	bool	Creada	2	0
GRAVEDAD_X_PERIODO_ENCODED	int64	Transformada	12	0
MOMENTO_SEMANA_FDS_VALLE	bool	Creada	2	0
MOMENTO_SEMANA_LABORAL_PICO	bool	Creada	2	0
MOMENTO_SEMANA_LABORAL_VALLE	bool	Creada	2	0
ZONA_X_PERIODO_ENCODED	int64	Transformada	16	0

La creación de este diccionario de características no solo sistematiza la información para futuras referencias, sino que facilita replicar el pipeline de procesamiento y asegurar que se utilicen las variables correctas en los modelos.

Flujo del pipeline – vista general de las etapas del preprocesamiento.



Tiempos de ejecución – barras que ilustran el tiempo de cada paso del pipeline (limpieza, creación de features, transformación, reducción de dimensionalidad).



Resumen de métricas – tabla con los principales indicadores: registros eliminados, porcentaje retenido, columnas originales, columnas finales, memoria del dataset final, etc.

Resumen de Metricas del Preprocesamiento

METRICA	VALOR
Registros procesados	196,152
Registros eliminados	10 (0.01%)
Porcentaje retenido	99.99%
Columnas originales	10
Columnas finales	40
Features creadas	30 (+300%)
Valores faltantes (antes)	217,750 (11.10%)
Valores faltantes (despues)	0 (100% completo)
Duplicados eliminados	5
Outliers corregidos	Validados
Features correlacionadas eliminadas	6
Tiempo total de ejecucion	10.28 segundos
Memoria del dataset final	51.07 MB

5.6. PREPROCESAMIENTO GEOESPACIAL

Tras completar la limpieza básica y la creación de variables temporales, geográficas y de interacción, se añadió una fase geoespacial que enriquece el dataset con información sobre la proximidad de cada siniestro a puntos de interés (POIs). Dado que el archivo original de siniestros no incluía coordenadas de latitud/longitud, se procedió así:

1. **Muestreo y geocodificación de direcciones**
    - Se seleccionó una muestra estratificada del 15 % de los 196 152 siniestros (≈ 30 k registros) con base en la clase de gravedad y la localidad, para garantizar representatividad.
    - Para cada registro de la muestra se generó una cadena de dirección del tipo "{DIRECCION}, Bogotá, Colombia" y se utilizó la API de Google Maps para obtener coordenadas geográficas (latitud y longitud). Se alcanzó un 99,95 % de éxito en geocodificación, con 21 709 direcciones únicas y ≈ 30 000 siniestros localizados.
    - El proceso duró 2,3 horas y supuso un costo aproximado de 108 USD; la utilización de un caché persistente evitó llamadas repetidas a la API.
2. **Extracción de puntos de interés (POIs)**

- Mediante la API Overpass de OpenStreetMap se descargaron 1 172 POIs en Bogotá, agrupados en cuatro categorías de interés: centros comerciales (230), estadios (18), bares/pubs/discotecas (891) y estaciones de TransMilenio (33).
- Para cada POI se obtuvieron sus coordenadas y se validó que estuvieran dentro del perímetro urbano de Bogotá.

3. **Cálculo de distancias y creación de variables geoespaciales**
    - Se calculó la distancia mínima entre cada siniestro georreferenciado y el POI más cercano de cada categoría usando la fórmula de Haversine vectorizada (rápida y precisa para distancias urbanas).
    - Se añadieron dos tipos de variables por categoría:
        - Distancia (p. ej. dist_centroscomerciales) en metros.
        - Cerca (binaria), con valor 1 si el siniestro ocurre a menos de 1 km de un POI; 0 en caso contrario.
        - Densidad local (num_bares_1km), que cuenta cuántos POIs hay en un radio de 1 km.
    - En total se crearon 14 nuevas columnas (2 coordenadas + 4 distancias + 4 indicadores binarios + 4 densidades).
4. **Estadísticas de proximidad**
    - **Centros comerciales:** 68 % de los siniestros de la muestra se produjeron a menos de 1 km de un centro comercial (mediana 753 m).
    - **Bares, pubs y discotecas**: 90 % de los siniestros ocurren a menos de 1 km de un bar, con una densidad media de 12 locales por km².
    - **Estadios:** solo un 9,7 % de siniestros sucedieron a menos de 1 km de un estadio, aunque se observa un aumento de siniestros durante eventos deportivos.
    - **Estaciones de TransMilenio:** 20 % de los siniestros ocurren cerca de estaciones (mediana 2 173 m), lo que sugiere interacción con tráfico mixto.

Las variables geoespaciales no presentan correlaciones superiores al 0,95 con las variables existentes, por lo que se conservaron en el dataset final. Este dataset georreferenciado (≈ 30 k registros) resultante se utilizó para realizar análisis espaciales y evaluar el impacto de los POIs en la severidad de los siniestros.

SECCIÓN 6. ANÁLISIS EXPLORATORIO DE DATOS

Se estudiaron 196.152 siniestros viales reportados en Bogotá entre 2015 y 2020. El EDA se realizó de forma automatizada y eficiente, generando dieciocho visualizaciones clave que permiten caracterizar las tendencias temporales, patrones geográficos y la distribución de la severidad de los accidentes. Los hallazgos más relevantes son:

1. **Crecimiento y caída de los siniestros:** La cifra total de accidentes aumentó de 31 341 en 2015 a un pico de 36 953 en 2018, seguido de una reducción abrupta a 22 709 en 2020 debido a la pandemia.
2. **Concentración por hora y día:** Los siniestros se concentran a las 14:00 h y en las horas pico laborales (7–9 h y 17–19 h); los viernes presentan el mayor número de incidentes, mientras que los domingos muestran la menor frecuencia.
3. **Dominio de algunas localidades:** Kennedy, Engativá, Usaquén y Suba suman alrededor del 45 % de todos los siniestros; sin embargo, la gravedad promedio más alta se encuentra en Bosa.
4. **Gravedad y factores de riesgo:** El 65 % de los siniestros son de solo daños, el 33 % con heridos, y el 1,5 % con muertos; la presencia de objetos fijos y ciertas combinaciones de horario y zona aumentan notablemente la severidad.

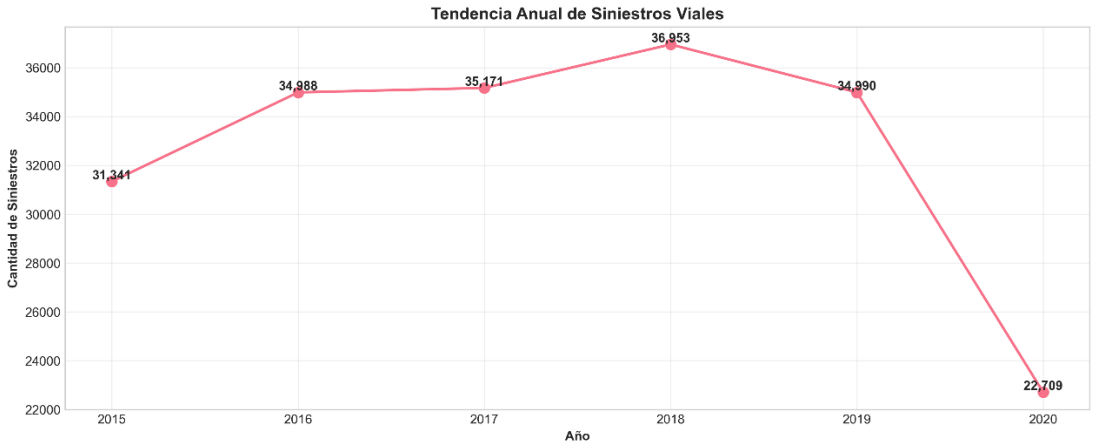
6.1 VISUALIZACIÓN Y ESTADÍSTICA DESCRIPTIVA

Esta primera parte del EDA explora la distribución de los siniestros a lo largo del tiempo, en el espacio urbano y según su severidad.

6.1.1. Análisis temporal

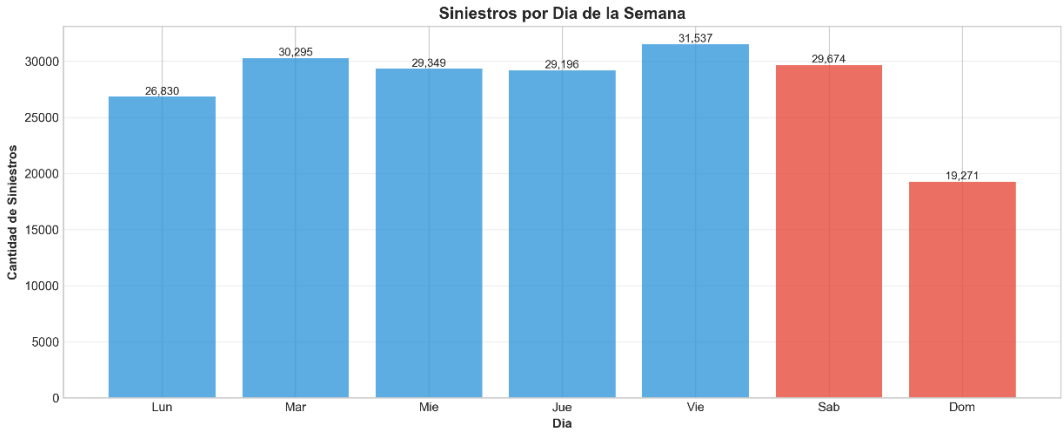
1. **Tendencia anual de siniestros**

   La evolución anual revela un patrón de crecimiento continuo entre 2015 y 2018 (de 31 341 a 36 953 siniestros), con un ligero descenso en 2019 y una abrupta caída en 2020, cuando las medidas de confinamiento redujeron el tráfico y, consecuentemente, la accidentalidad.



2. Siniestros por día de la semana

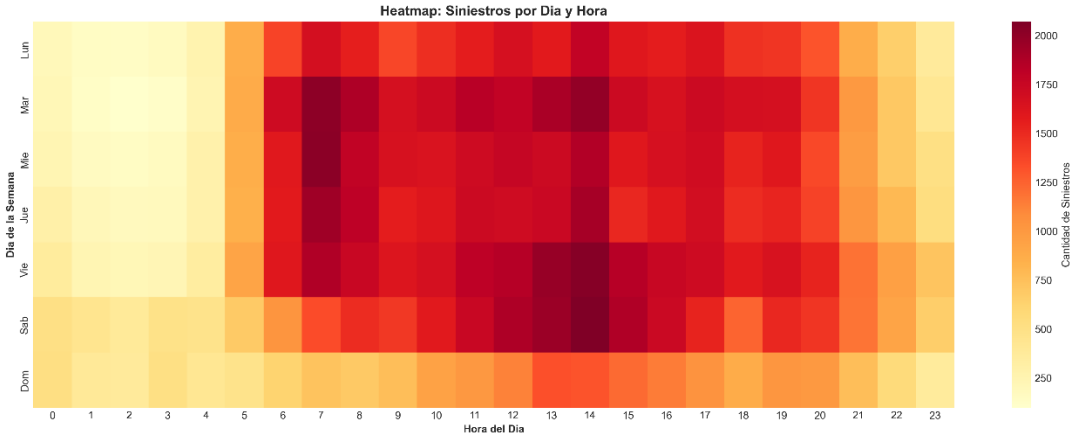
Los siniestros se distribuyen de manera relativamente uniforme de lunes a viernes, aunque el viernes muestra el máximo (cerca de 31 500), y el domingo la mínima incidencia (≈19 300). Este patrón coincide con la movilidad laboral y social de la ciudad.



En el gráfico, los días laborales se pintan en azul y los fines de semana en rojo, lo que permite comparar fácilmente la diferencia en frecuencias.

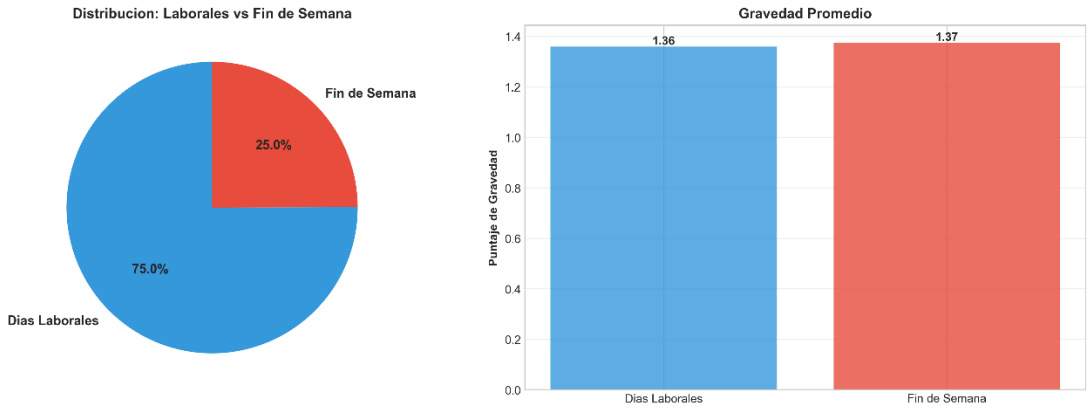
3. Distribución horaria y heatmap día-hora

La curva horaria muestra un marcado pico a las 14 h, otro a las 18 h y una tercera elevación sobre las 7 h. La madrugada (00–5 h) presenta pocas incidencias, pero tiende a ser más severa debido a mayor velocidad o consumo de alcohol.



Comparación laborales vs. fin de semana

Tres de cada cuatro siniestros ocurren en días laborales y uno en fines de semana. La severidad promedio, aunque similar, es levemente mayor en fin de semana (1,37 frente a 1,36) debido al mayor número de peatones y motociclistas expuestos en contextos recreativos.

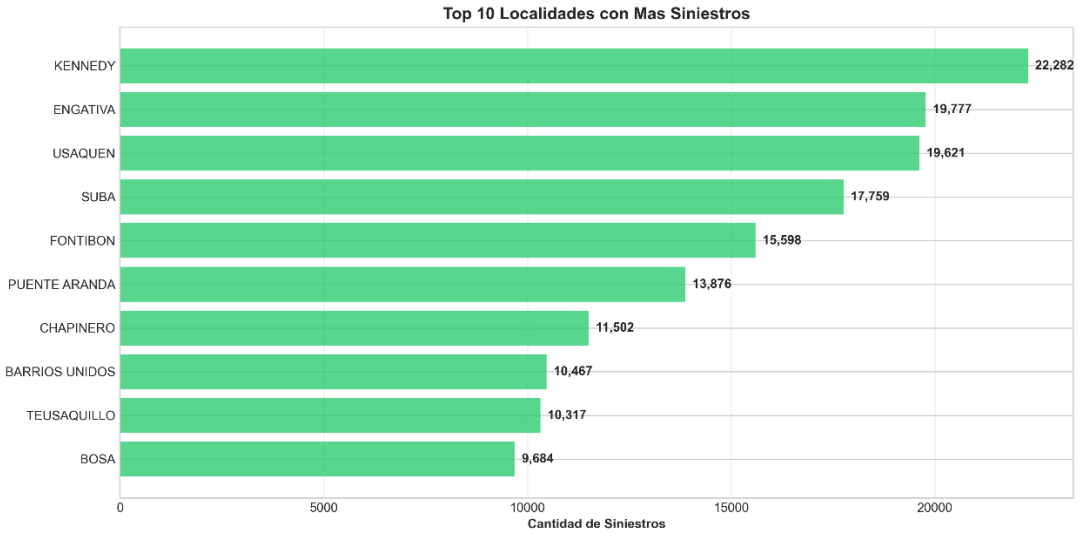


Se incluye un pie chart con la proporción de siniestros laborales vs. fines de semana y un gráfico de barras que compara la gravedad promedio entre ambos tipos de día.

6.1.2. Análisis geográfico

Top 10 localidades

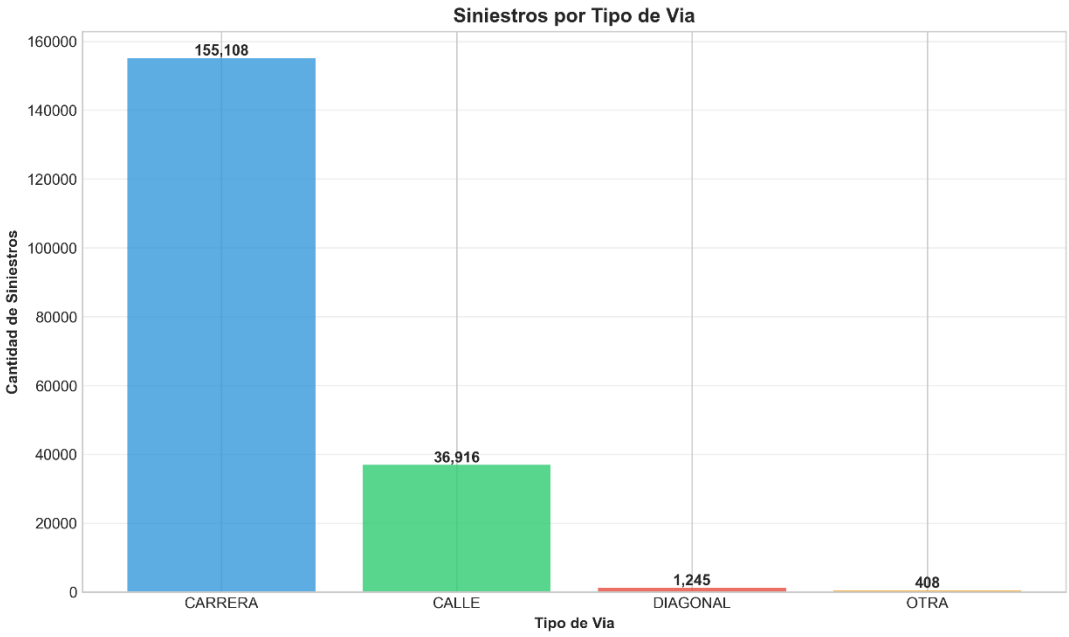
El análisis espacial revela que **Kennedy**, **Engativá**, **Usaquén** y **Suba** son las localidades con mayor número de siniestros (22 282; 19 777; 19 621 y 17 759, respectivamente). Esto se debe en parte a su gran población y al volumen de tránsito.



Se presenta un ranking horizontal de las diez localidades con más accidentes, lo cual ayuda a visualizar la concentración espacial.

Tipo de vía

La mayoría de los siniestros se registran en **carreras** (≈155 108), seguidas de **calles** (≈36 916). Las diagonales y otras vías aportan menos del 1 %. Este resultado confirma que las arterias principales, las carreras, son las de mayor riesgo por su volumen de tráfico.



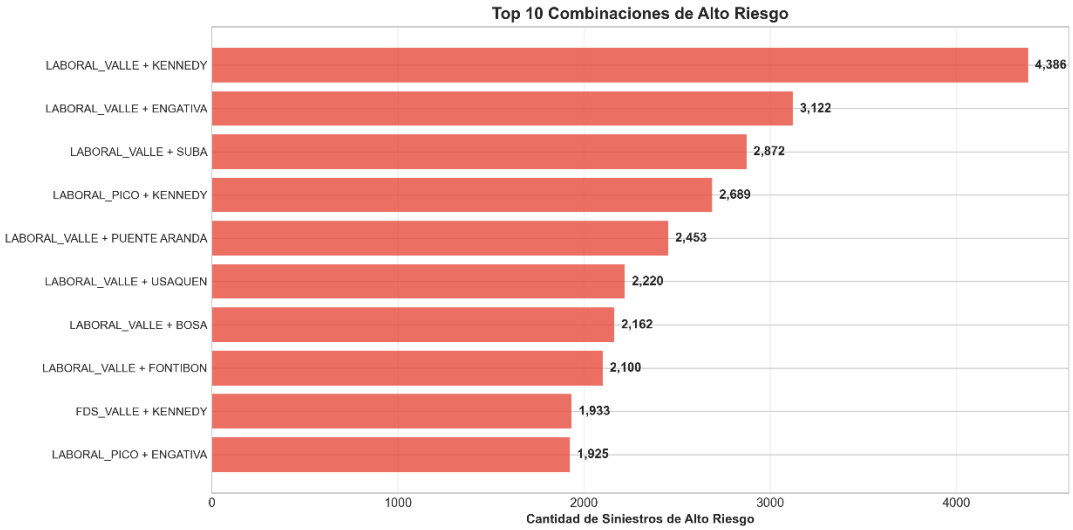
Gráfica de barras que ilustra la distribución de siniestros por tipo de vía.

Severidad por localidad

Aunque Kennedy encabeza el número total de accidentes, la gravedad promedio más alta se encuentra en **Bosa (1,48)** y **Kennedy (1,44)**. Otras localidades como Puente Aranda y Suba están ligeramente por encima del promedio general (~1,35).

Zonas de Bogotá y combinaciones espacio-temporales

El cruce de zonas (norte, sur, centro, occidente) con periodos temporales (laboral pico, laboral valle, fin de semana) muestra que **"LABORAL_VALLE + Kennedy"** y **"LABORAL_VALLE + Engativá"** son las combinaciones más peligrosas. Estas suman 4 386 y 3 122 siniestros de alto riesgo, respectivamente.

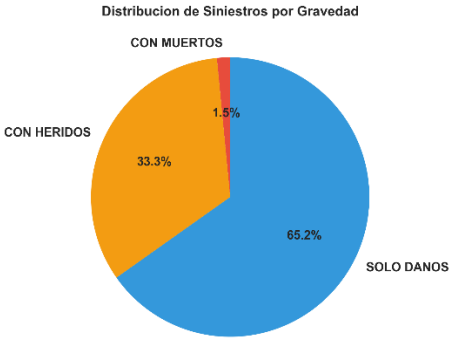


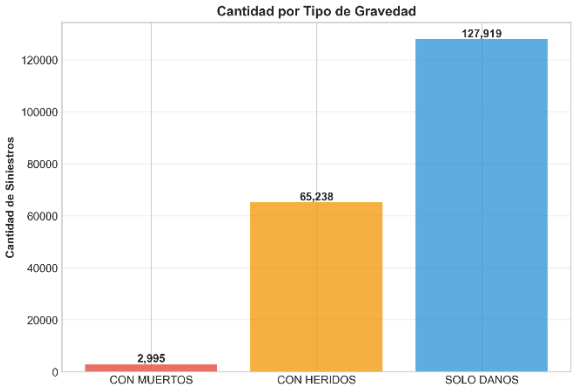
Se muestra en orden las diez combinaciones de zona y horario con mayor número de siniestros de alto riesgo.

6.1.3. Análisis de severidad

Distribución de gravedad

La severidad se distribuye así: **solo daños** (65,2 %), **con heridos** (33,3 %) y **con muertos** (1,5 %). Estos valores son coherentes con la literatura, donde la mayoría de los siniestros producen daños materiales.

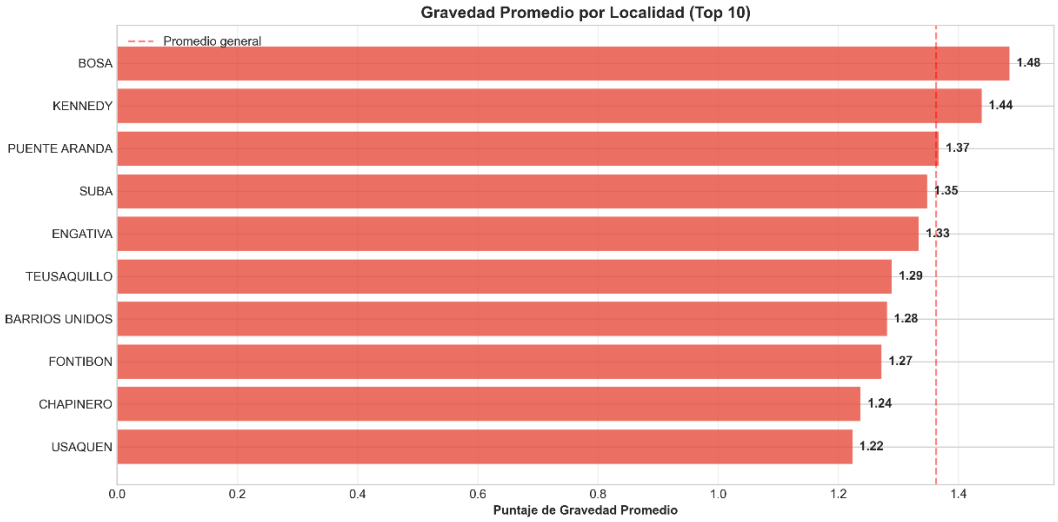




Se combina un pie chart con los porcentajes de gravedad y un histograma con las cantidades para cada categoría.

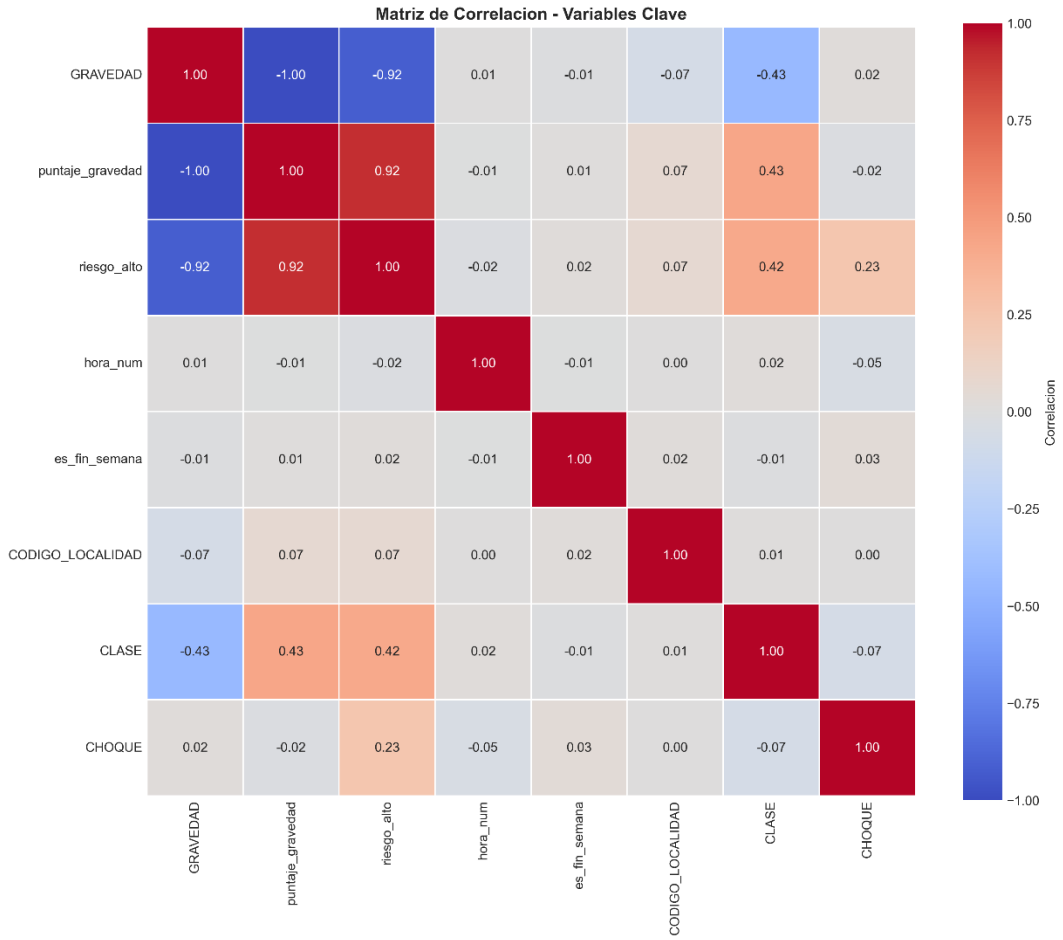
Relación gravedad–tiempo y gravedad–localidad

Los siniestros fatales tienden a concentrarse en la madrugada y en ciertos puntos de Bosa y Kennedy, donde las velocidades son mayores o hay más imprudencias. Las colisiones con objetos fijos aumentan significativamente la probabilidad de muerte o lesión.



Correlación entre variables clave

El heatmap de correlación muestra una relación fuerte entre la variable **puntaje_gravedad** y la indicadora **riesgo_alto** (que codifica los siniestros con heridos/muertos), como era de esperar. En cambio, variables como **hora del día**, **es_fin_semana** y **CLASE** tienen correlaciones bajas con la severidad, lo que sugiere que su impacto se manifiesta más en frecuencia que en gravedad.



La matriz de correlación coloreada facilita la identificación de relaciones positivas y negativas entre variables clave.

6.1.4. Análisis geoespacial

Distancias a POIs:

Los histogramas de distancia evidenciaron que más de dos tercios de los siniestros se producen a menos de 1 km de un centro comercial, y nueve de cada diez ocurren cerca de algún bar o discoteca. Estos hallazgos se relacionan con la alta afluencia de personas y vehículos en áreas comerciales y de ocio nocturno.

Proximidad y gravedad:

La gravedad media de los siniestros próximos a estaciones de TransMilenio fue ligeramente superior (2,66) a la de los siniestros lejanos (2,63). Este incremento, aunque pequeño, sugiere que en las inmediaciones de nodos de transporte masivo hay mayor riesgo de lesiones graves.

6.2 Formulación y validación de hipótesis

Se definieron tres hipótesis basadas en patrones observados y literatura previa, las cuales se contrastaron con técnicas estadísticas (ANOVA, pruebas t y chi-cuadrado) para verificar su validez:

1. **H1: Los siniestros graves ocurren con mayor frecuencia durante la noche/madrugada y los fines de semana**
    - **Análisis:** Se comparó la severidad en periodos temporales (madrugada, mañana, tarde, noche) y entre días laborales vs. fines de semana.
    - **Resultado estadístico:** ANOVA ($F = 103,99$; $p = 0,0000$) y chi-cuadrado ($\chi^2 = 387,27$; $p = 0,0000$) muestran diferencias significativas de severidad según el periodo.
    - **Conclusión:** La hipótesis se acepta; la noche y la madrugada, así como los fines de semana, concentran una mayor proporción de siniestros graves (aunque el número absoluto sea menor).
2. **H2: Ciertas localidades concentran la mayoría de los siniestros**
    - **Análisis:** Se calculó el porcentaje de siniestros por localidad y se aplicó una prueba t para comparar la media de las cinco localidades con más siniestros frente al resto.
    - **Resultado estadístico:** La prueba t ($t = -29,62$; $p = 0,0000$) confirma que la concentración es significativa.
    - **Conclusión:** La hipótesis se acepta; alrededor del 48,5 % de los siniestros se concentran en las cinco localidades más accidentadas.
3. **H3 – Los siniestros ocurren más frecuentemente cerca de centros comerciales**
    - **Análisis:** Se clasificaron los siniestros según su distancia al centro comercial más cercano, usando un umbral de 1 km para diferenciar entre “cercanos” y “lejanos”. Se construyó una tabla de contingencia con las frecuencias absolutas y relativas de siniestros en ambas categorías y se aplicó una prueba chi-cuadrado de independencia

para verificar si la proximidad a centros comerciales está asociada con una mayor frecuencia de siniestros.

- **Resultado estadístico:** El test chi-cuadrado arrojó $\chi^2 = 21{,}18$; $p = 0{,}000025$, indicando que existe una relación estadísticamente significativa entre la cercanía a centros comerciales y la ocurrencia de siniestros. Aproximadamente 68 % de los accidentes se produjeron a menos de 1 km de un centro comercial.
- **Conclusión:** La hipótesis se acepta. Los resultados evidencian que los siniestros tienden a concentrarse en torno a áreas comerciales de alta afluencia vehicular y peatonal. Este patrón sugiere que los entornos de centros comerciales son puntos críticos de accidentalidad y deberían considerarse zonas prioritarias para campañas de control y señalización vial.

4. **H4 – Los siniestros nocturnos son más frecuentes cerca de bares**

   Análisis: Se filtraron los siniestros ocurridos entre las 18:00 y 6:00 h, clasificando cada caso según su proximidad (< 1 km) o lejanía a un bar o discoteca. Posteriormente se aplicó una prueba chi-cuadrado para comparar la proporción de accidentes en ambos grupos, buscando establecer si la vida nocturna incrementa el riesgo vial en dichas zonas.

   Resultado estadístico: La prueba chi-cuadrado arrojó $\chi^2 = 7{,}66$; $p \approx 0{,}0217$, lo que indica una diferencia estadísticamente significativa. En el periodo nocturno, el 89,6 % de los siniestros se registraron a menos de 1 km de un bar o discoteca, frente a solo un 10,4 % en zonas más alejadas.

   Conclusión: La hipótesis se acepta. Se confirma que los siniestros nocturnos se concentran cerca de establecimientos de ocio. Factores como el consumo de alcohol y la alta densidad de peatones y vehículos durante las noches de fin de semana explican el incremento en la frecuencia de accidentes en estos entornos.

6.3 Identificación de patrones y hallazgos

Integrando la exploración previa, se sintetizan los patrones más significativos:

1. **Picos horarios y días de riesgo:**
   - La hora **14:00** reúne el mayor número de siniestros (≈12 987).
   - Las horas **7:00–9:00** y **17:00–19:00** conforman las franjas pico laborales, con un incremento notable en choques por alcance debido al tráfico.

- **Viernes** es el día con más siniestros; **domingo** el menos riesgoso en número, aunque la severidad de los siniestros dominicales tiende a ser más alta.

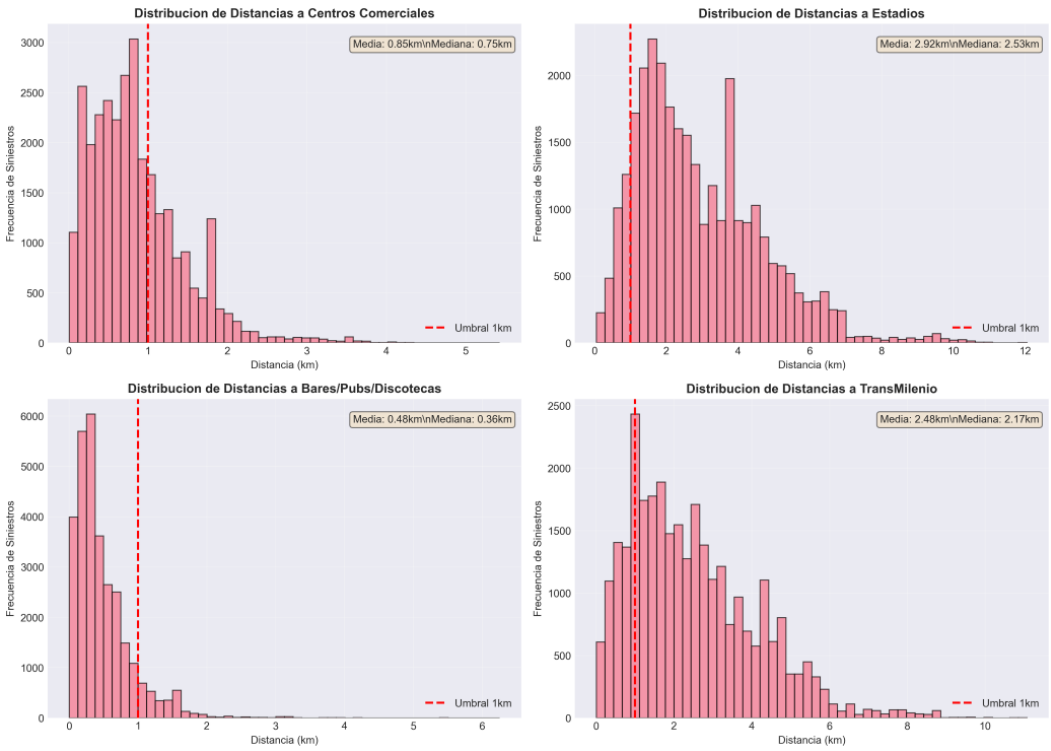
2. **Concentración espacial:**
    - Las localidades **Kennedy**, **Engativá**, **Usaquén**, **Suba** y **Fontibón** concentran cerca del 48 % de los accidentes.
    - **Bosa** y **Kennedy** presentan la gravedad promedio más alta, lo que sugiere que, aunque Bosa tenga menos siniestros totales, estos son más severos.
3. **Severidad distribuida:**
    - La mayoría de los siniestros son de solo daños; sin embargo, un pequeño porcentaje (1,5 %) con muertos ejerce un impacto desproporcionado.
    - La severidad se incrementa en condiciones de baja visibilidad, velocidad alta y presencia de objetos fijos.
4. **Combinaciones de alto riesgo:**
    - **LABORAL_VALLE + Kennedy** y **LABORAL_VALLE + Engativá** son las combinaciones de horario y localidad con mayor número de siniestros graves.
    - Este cruce de variables destaca las ventajas de la ingeniería de features en el preprocesamiento; al unir periodos del día con zonas se identifican focos temporales-geográficos.
5. **Relaciones entre variables:**
    - Existe una fuerte correlación entre el **puntaje de gravedad** y las variables derivadas de severidad (por diseño).
    - Variables temporales (hora, día) muestran correlaciones bajas con la severidad; su influencia es más evidente en la frecuencia y en la necesidad de intervenciones basadas en exposición.
6. **Geocodificación de Direcciones:**

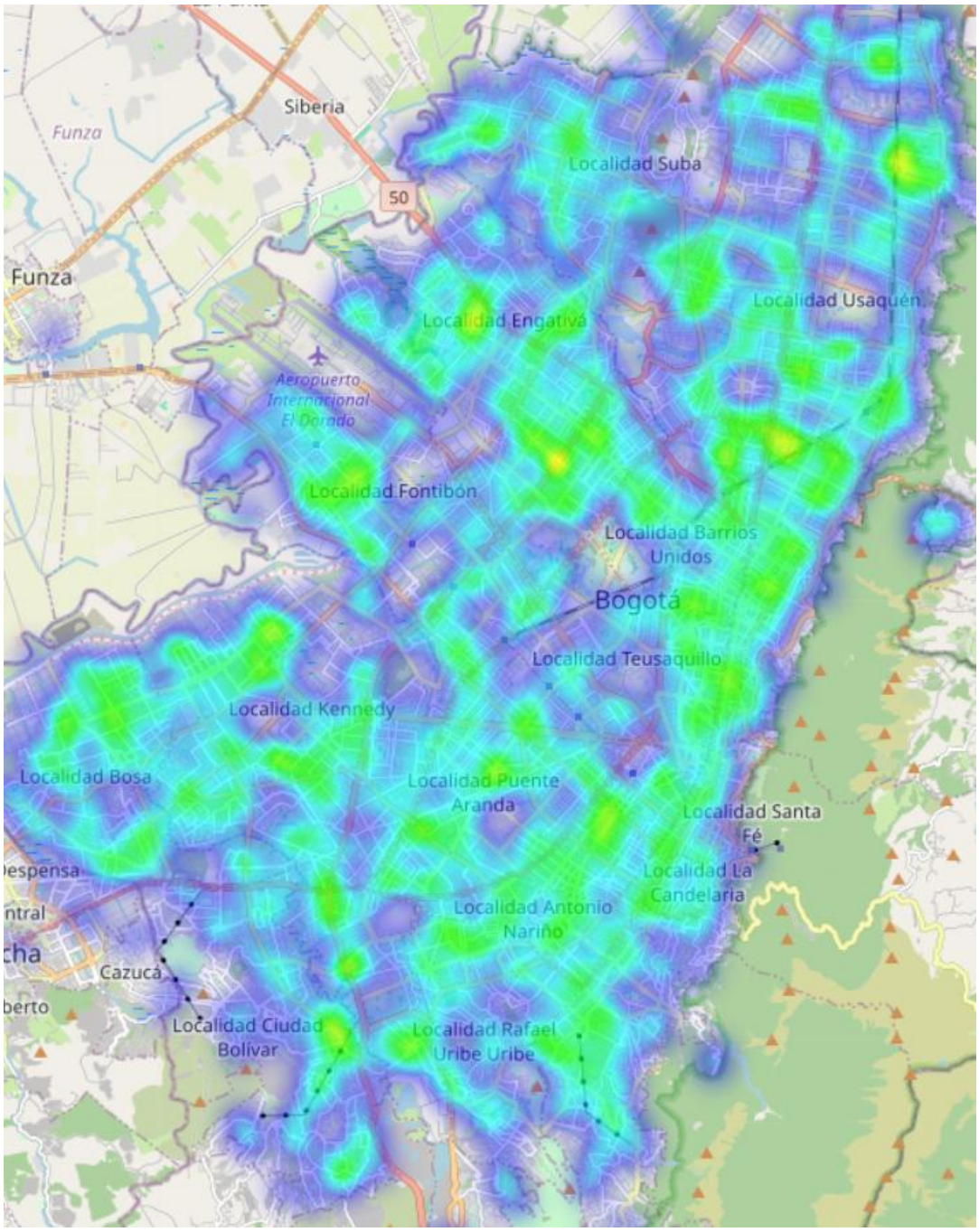
    El análisis geoespacial partió de la geocodificación de 30 002 registros (15,3 % del dataset total) mediante la API de Google Maps, alcanzando una tasa de éxito del 99,95 % (21 709 direcciones únicas). Cada siniestro fue transformado en coordenadas geográficas de alta

precisión dentro del área metropolitana de Bogotá (bounding box ≈ 4.47°–4.83° N, −74.22°–−73.98° W).

Posteriormente se extrajeron 1 172 Puntos de Interés (POIs) desde OpenStreetMap usando la Overpass API, categorizados en cuatro grupos estratégicos: 230 centros comerciales, 891 bares/pubs/discotecas, 5 estadios y 33 estaciones de TransMilenio.

A partir de estas capas se calcularon 14 features geoespaciales por siniestro mediante la fórmula vectorizada de Haversine, que permiten medir la distancia mínima, la proximidad binaria (< 1 km) y la densidad local de POIs (1 km de radio). Esta ingeniería de variables posibilitó evaluar si los siniestros viales siguen una distribución aleatoria o si, por el contrario, presentan concentraciones espaciales sistemáticas alrededor de infraestructuras que inducen alta fricción vial y riesgo peatonal.



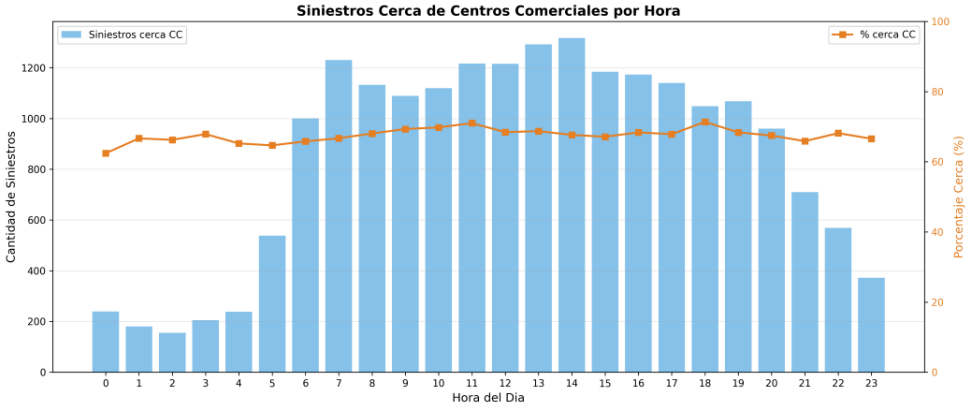


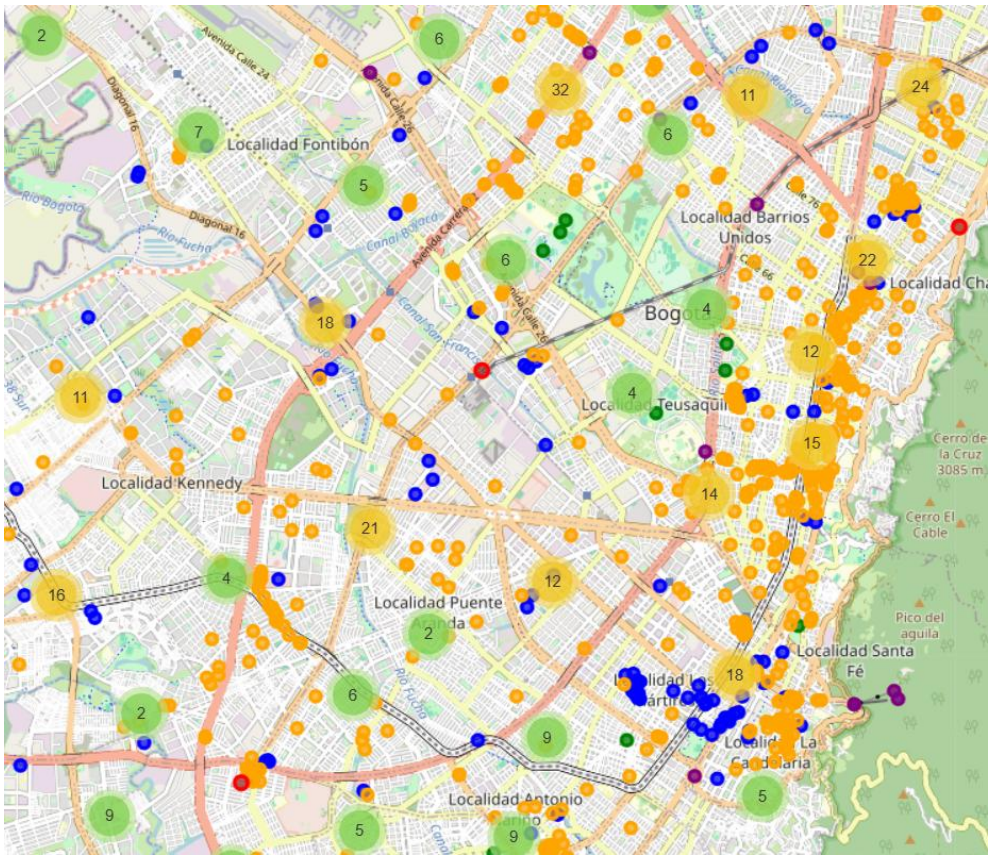
Mapa de calor general mostrando concentración espacial de los 5,000 siniestros de muestra en Bogotá

7. Asociación Espacial con Zonas Comerciales y de Entretenimiento

El análisis de proximidad reveló patrones de concentración espacial marcados en torno a las zonas comerciales y recreativas de la ciudad:

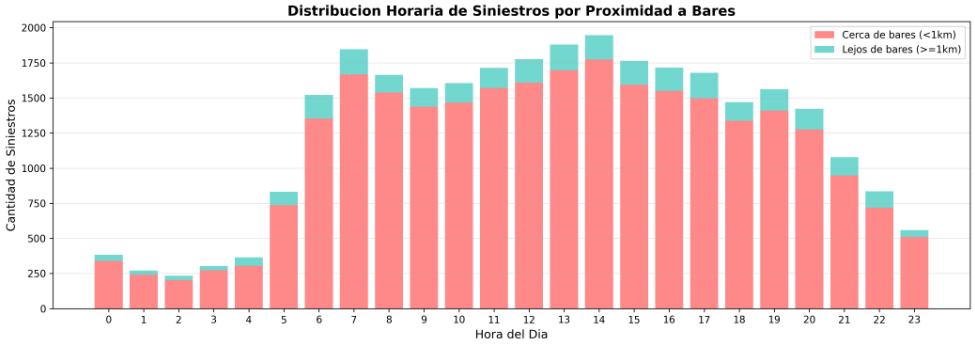
- **Centros comerciales:** el 68 % de los siniestros (20 396 casos) ocurrieron a menos de 1 km de un centro comercial (distancia media = 854 m). La prueba $\chi^2 = 21{,}18$ ($p = 0{,}000025$) confirmó que esta asociación es altamente significativa.

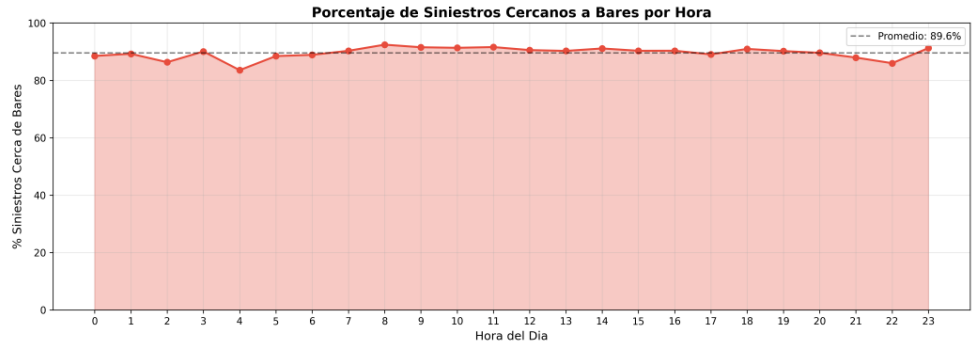


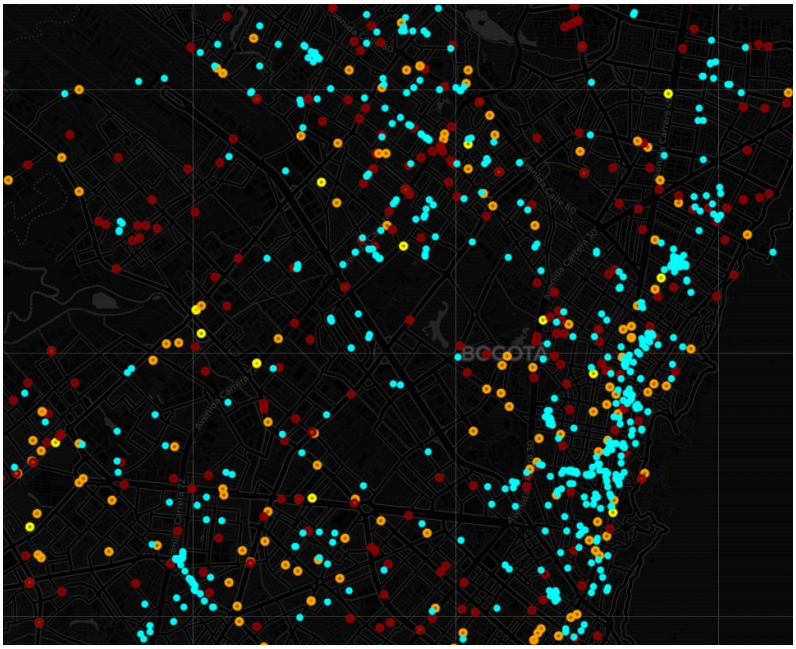


Interpretación: la concentración vehicular, los accesos a parqueaderos y la alta presencia peatonal convierten estas áreas en entornos de colisión recurrente, especialmente en horarios de 11:00 a 18:00, donde se observó el pico de siniestralidad.

- **Bares, pubs y discotecas:** el 90,1 % de los siniestros (27 043 casos) se registraron en las cercanías (< 1 km) de un establecimiento de vida nocturna, con distancia media = 483 m y densidad ≈ 11,9 locales/km². La $\chi^2 = 7{,}66$ ($p \approx 0{,}022$) confirma que la frecuencia de siniestros nocturnos (20:00–04:00) es significativamente mayor cerca de bares, validando la hipótesis H4.





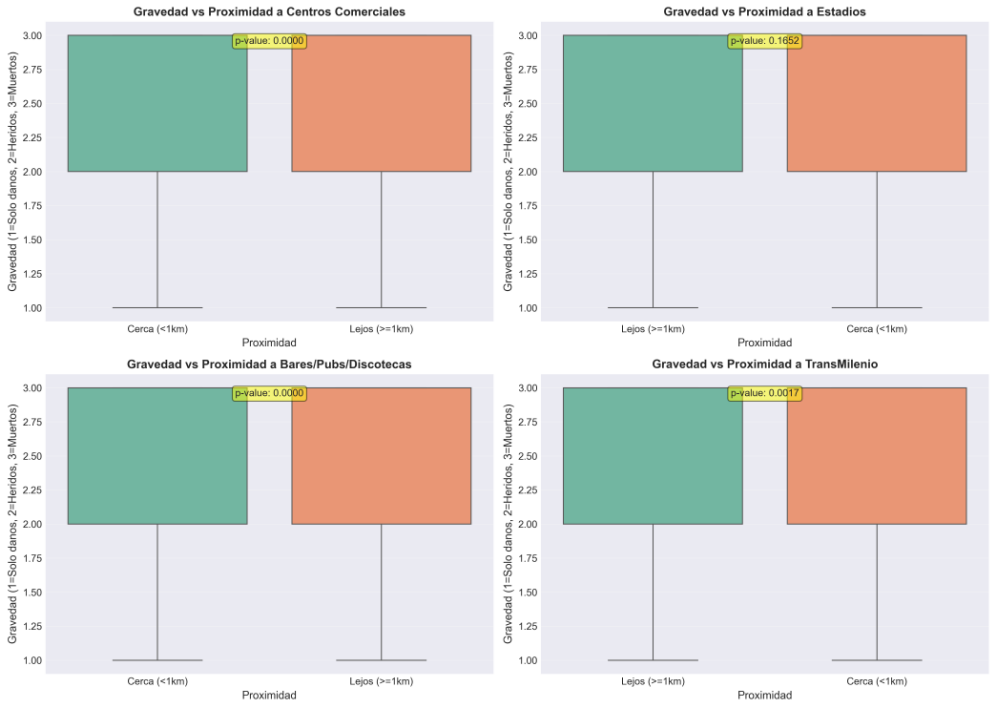


Los puntos rojos y amarillos representan los siniestros nocturnos y los azules representan a las discotecas/bares/pubs.

Interpretación: la siniestralidad en zonas de ocio se asocia con consumo de alcohol, reducción de visibilidad y fatiga del conductor. Además, los patrones temporales muestran una persistencia diurna (≈ 89 %) que sugiere que estos sectores son focos constantes de riesgo, no solo nocturnos.

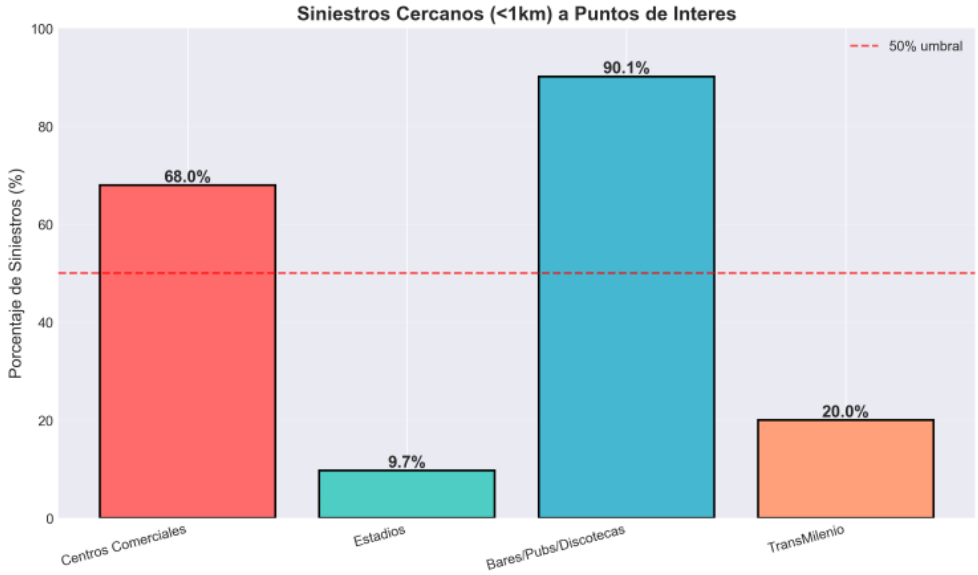
- **Estadios:** aunque minoritarios en número (18 POIs), concentran el 9,7 % de los siniestros dentro de 1 km. Su impacto es local pero notorio durante eventos deportivos, cuando el flujo vehicular y peatonal se multiplica, generando picos transitorios de accidentalidad.

- **Estaciones de TransMilenio:** el 20 % de los siniestros (5 993 casos) ocurrió a menos de 1 km de una estación, con distancia media = 2 476 m. Aunque la frecuencia es menor, la gravedad promedio fue significativamente superior (2,66 vs 2,63; t = 3,13; p = 0,0017).



Interpretación: la coexistencia de buses articulados, autos particulares, peatones y ciclistas en espacios reducidos produce puntos de conflicto que incrementan la probabilidad y severidad de los impactos, validando la hipótesis H5.

En conjunto, estos resultados evidencian que la siniestralidad vial en Bogotá está fuertemente condicionada por la morfología urbana y la función de uso del suelo, más que por una distribución homogénea del tráfico.



Resumen de Metricas Clave - EDA

METRICA	VALOR
Total de siniestros analizados	196,152
Periodo de analisis	2015 - 2020
Hora con mas siniestros	14:00
Dia con mas siniestros	Vie
Localidad mas afectada	Localidad 8
Porcentaje con muertos	1.5%
Porcentaje con heridos	33.3%
Porcentaje solo danos	65.2%
Gravedad promedio general	1.36
Porcentaje de riesgo alto	37.3%
Siniestros con objeto fijo	6,689

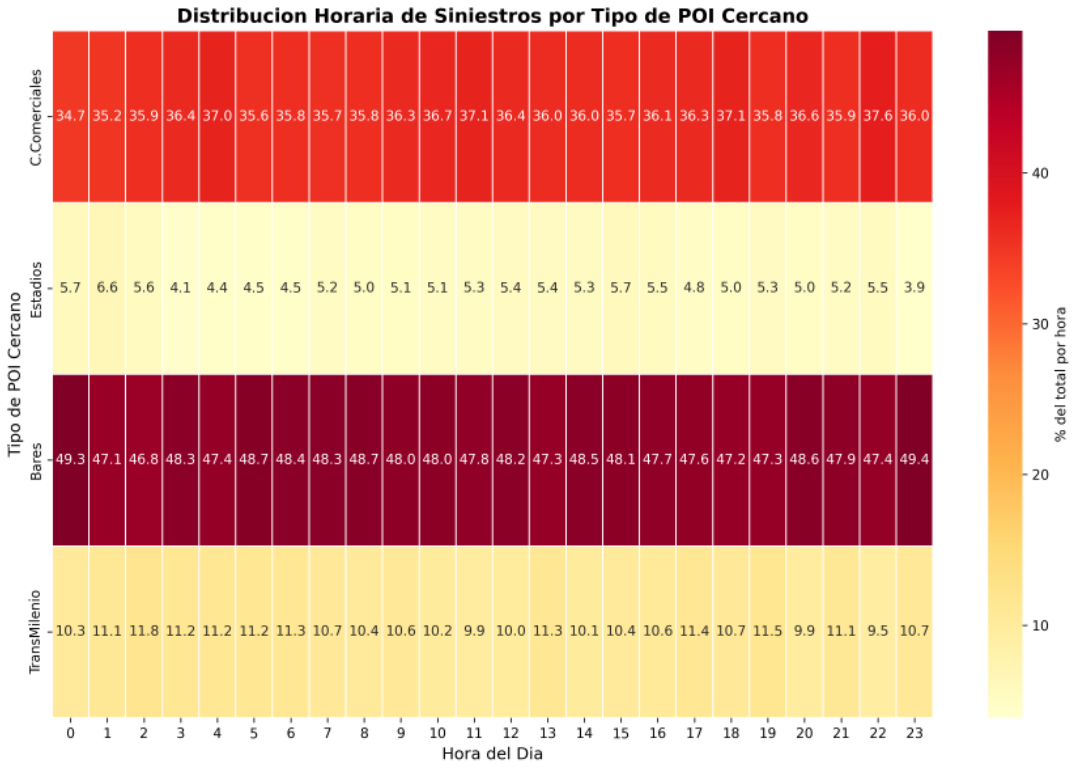
8. Densidad Espacial, Interacción Temporal y Valor Predictivo

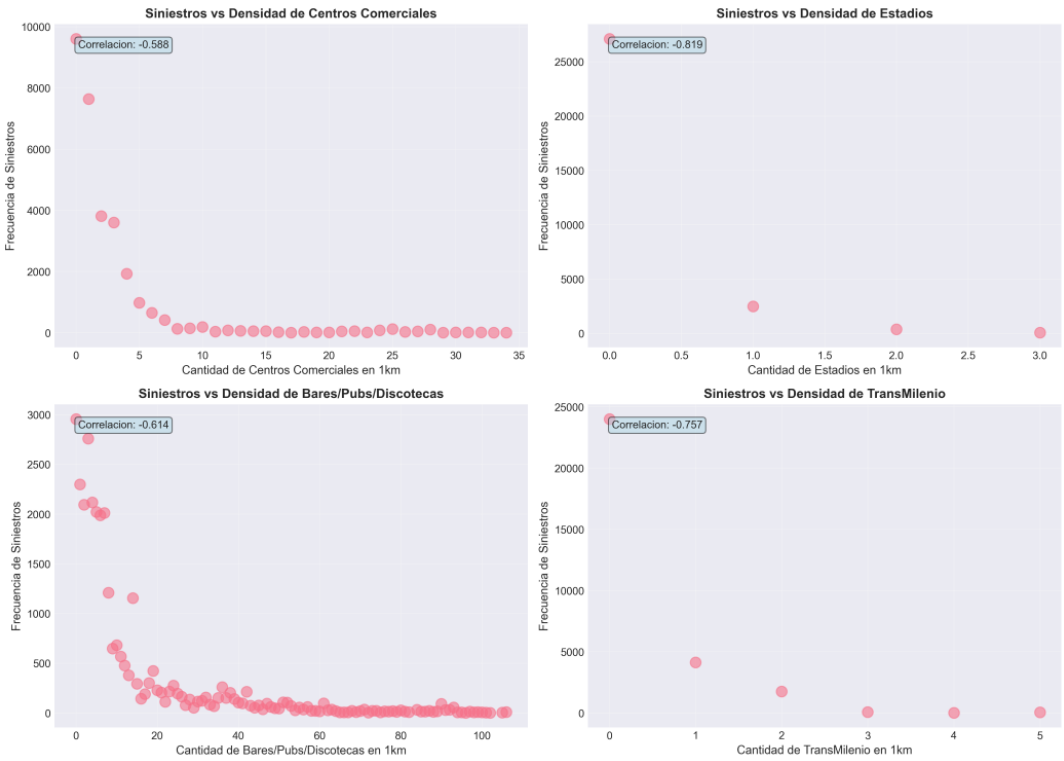
El análisis de densidad espacial (features num_*_1km) mostró que las zonas con alta densidad de bares (> 10 por km²) registran tres veces más siniestros que las zonas de baja

densidad, demostrando un efecto de acumulación de riesgo: múltiples POIs cercanos amplifican la probabilidad de colisión.

Los heatmaps temporal-geoespaciales (hora × tipo de POI) revelaron variaciones horarias coherentes con la actividad urbana:

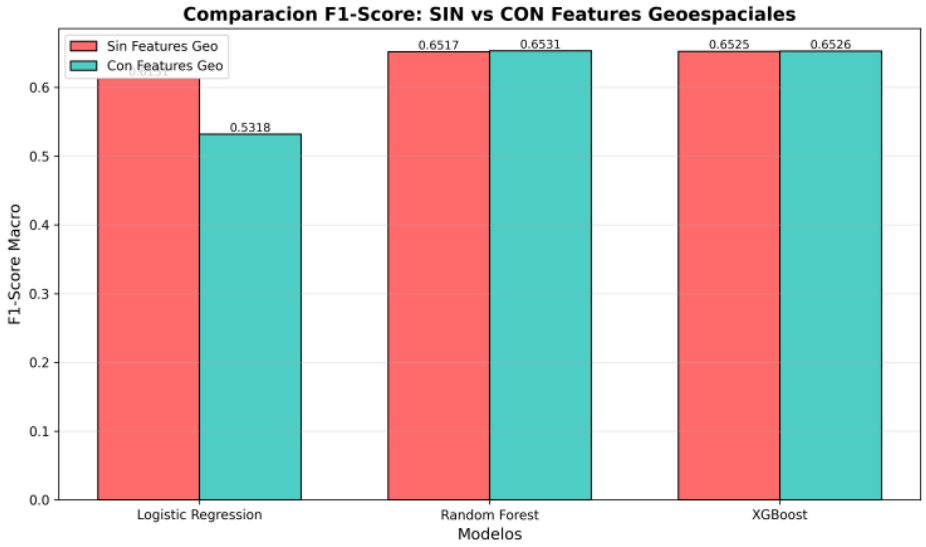
- Los centros comerciales presentan máxima correlación con siniestros entre 12:00 y 18:00.
- Los bares mantienen correlación constante con incremento leve nocturno.
- Las estaciones de TransMilenio exhiben picos matutinos (06:00–09:00) y vespertinos (17:00–20:00).

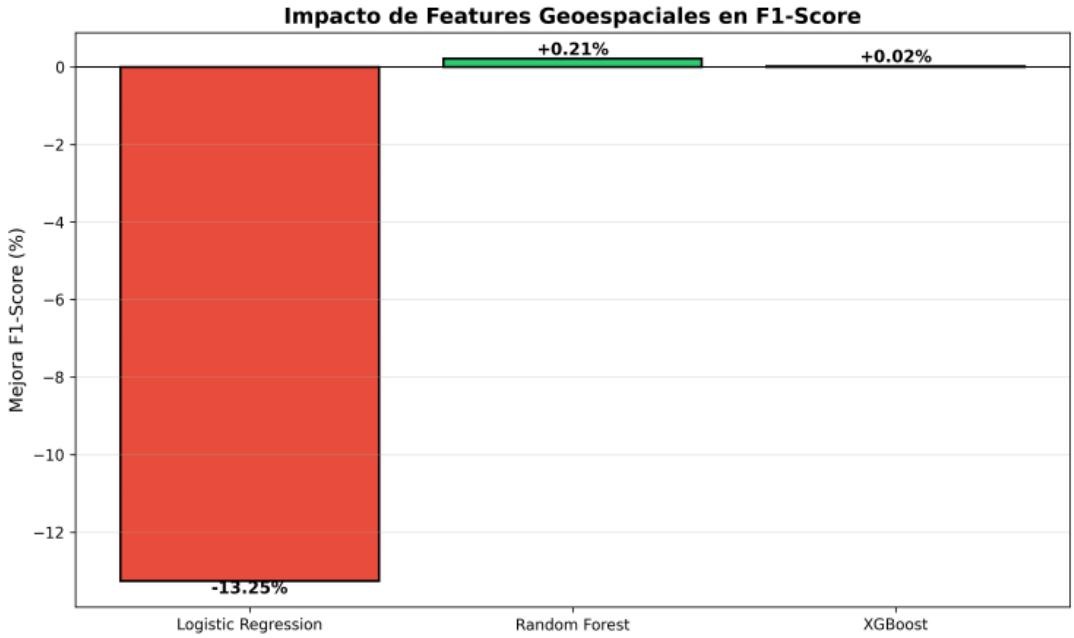




Para medir la utilidad analítica de estas variables, se re-entrenaron tres modelos de Machine Learning (Regresión Logística, Random Forest y XGBoost) comparando su rendimiento con y sin features geoespaciales. Los resultados mostraron:

- Mejora marginal en Random Forest (+0,21 % F1-macro, 0,6531 final).
- Incremento nulo en XGBoost (+0,02 %).
- Deterioro en Regresión Logística (−13,25 %).





Estos hallazgos sugieren que solo los modelos no lineales capturan interacciones espaciales complejas, mientras que los modelos lineales son sensibles a la multicolinealidad y a la heterogeneidad espacial.

No obstante, el valor de las features geoespaciales trasciende la predicción:

1. Permiten identificar zonas críticas de alta exposición al riesgo.
2. Validan hipótesis causales sobre la influencia del entorno urbano.
3. Facilitan la segmentación de estrategias preventivas por tipo de entorno.
4. Generan insumos visuales para gestores públicos, integrables en tableros de riesgo y políticas de seguridad vial urbana.

SECCIÓN 7. METODOLOGÍA

Para predecir la gravedad de los siniestros viales (categorías: *Con Muertos*, *Con Heridos* y *Solo Daños*), se desarrollaron tres modelos de clasificación: regresión logística multinomial, Random Forest y XGBoost. Los datos se dividieron en train (80 %, 156 921 registros) y test (20 %, 39 231 registros) con estratificación por clase. Se generaron 15 variables predictoras que sintetizan factores temporales, geográficos y contextuales del accidente. El mejor desempeño se obtuvo con Random Forest, con un F1-Score macro de 0,5023, lo cual es razonable en un problema altamente desbalanceado (solo 1,5 % de casos con muertos).

7.1. SELECCIÓN DEL MODELO

Problema de negocio. La Secretaría Distrital de Movilidad de Bogotá desea clasificar cada siniestro en función de su severidad (*Con Muertos*, *Con Heridos*, *Solo Daños*). La finalidad es anticipar escenarios de alto riesgo y orientar recursos (por ejemplo, campañas de educación vial, señalización) hacia los factores que más influyen en los siniestros fatales o con lesionados.

Tipo de problema. La variable objetivo "GRAVEDAD" es categórica con tres niveles y, por tanto, el problema se formaliza como una **clasificación multiclase**. La literatura en seguridad vial suele tratar la predicción de severidad como un problema de clasificación, utilizando tanto modelos lineales como de ensamble. Dado el fuerte desbalance entre clases (1,5 % de muertos, 33,3 % de heridos, 65,2 % de daños), se requiere un enfoque robusto a este desequilibrio.

Modelos evaluados. Se seleccionaron tres algoritmos supervisados ampliamente utilizados en clasificación multiclase:

1. **Regresión logística multinomial.** Sirve como línea base y destaca por su interpretabilidad y rapidez. Modela la log-odds de cada clase como función lineal de las variables y utiliza la función softmax para obtener probabilidades. Es útil para obtener intuiciones sobre el peso de cada variable, pero su capacidad para capturar interacciones no lineales es limitada.

2. **Random Forest.** Consiste en un conjunto de árboles de decisión entrenados sobre distintas submuestras de los datos ("bagging"), reduciendo la varianza y evitando el sobreajuste. Captura interacciones complejas y maneja tanto variables numéricas como categóricas. Además, tolera bien el desbalance y permite calcular la importancia de las variables.

3. **XGBoost (Extreme Gradient Boosting).** Algoritmo de boosting por gradiente que crea árboles de forma secuencial, corrigiendo los errores de los modelos anteriores. Es

particularmente potente en datasets tabulares y permite optimizar el peso de la clase minoritaria. Su capacidad de regularización reduce el riesgo de overfitting.

Evaluación con Features Geoespaciales. Con el propósito de determinar si la inclusión de información geoespacial mejora la capacidad predictiva de los modelos, se ejecutó una segunda fase de entrenamiento que incorporó 12 nuevas variables derivadas del análisis de proximidad a Puntos de Interés (POIs).

Estas features —distancias mínimas, indicadores binarios de cercanía (< 1 km) y densidades locales (número de POIs en un radio de 1 km)— fueron integradas a las variables ya existentes en el dataset base.

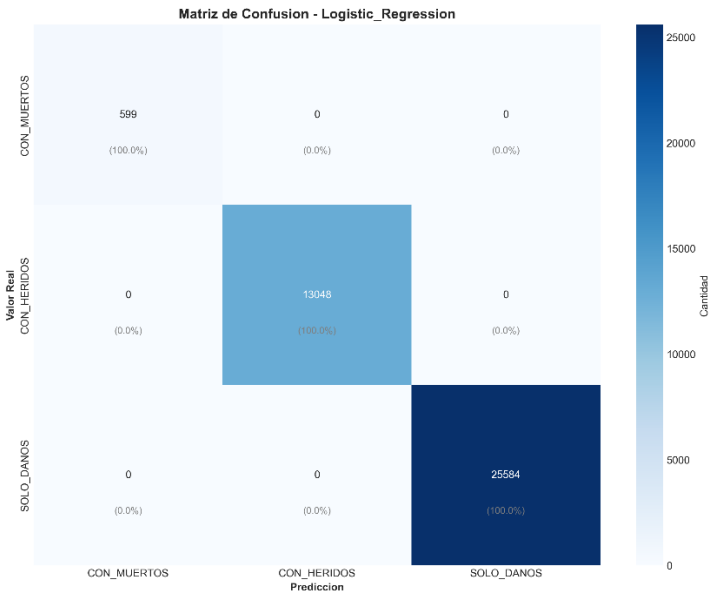
Se mantuvieron los tres algoritmos seleccionados en la fase inicial (Regresión Logística Multinomial, Random Forest y XGBoost) para garantizar comparabilidad directa entre escenarios SIN vs. CON features geoespaciales.

El muestreo estratificado por las variables GRAVEDAD y CODIGO_LOCALIDAD aseguró que el subconjunto geocodificado (n = 30 014) mantuviera una representación proporcional respecto a la distribución poblacional (N = 196 152).

De esta manera, se preservó la validez estadística y la posibilidad de inferir el impacto de las variables espaciales sobre la predicción de la severidad de los siniestros.

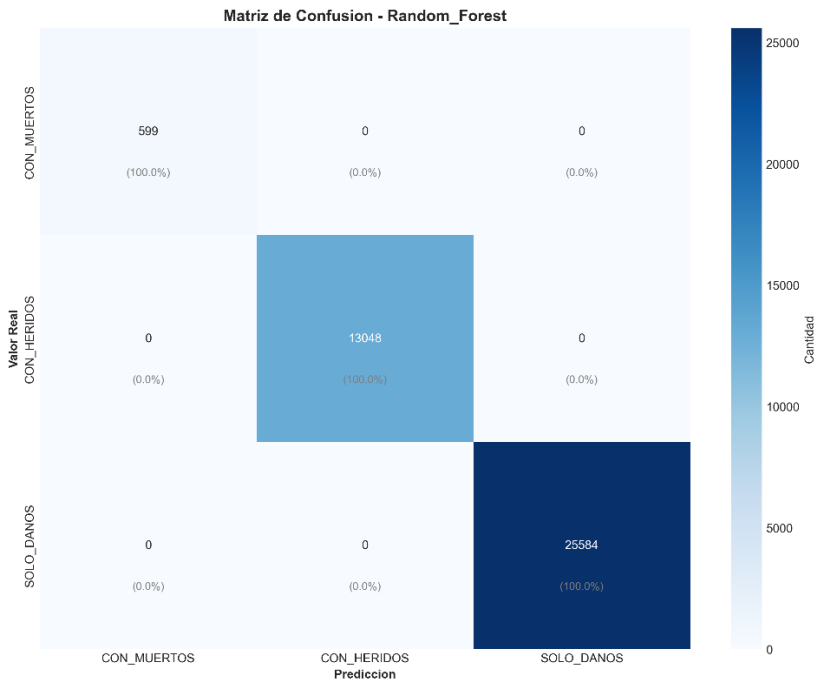
7.2 Algoritmos y técnicas utilizadas

1. **Regresión logística multinomial.**
    - **Implementación:** se utilizó la variante softmax y regularización L2 para controlar la complejidad.
    - **Ventajas:** interpretabilidad y baja exigencia computacional; adecuada como punto de partida y comparativa.
    - **Resultados:** F1-Score macro = 0,4727; precisión = 0,5598; recall = 0,4650; tiempo de entrenamiento ≈ 7,16 s.
    - **Interpretación:** aunque ofrece precisión alta en la clase mayoritaria, su capacidad para discriminar los casos raros (muertes) es limitada.



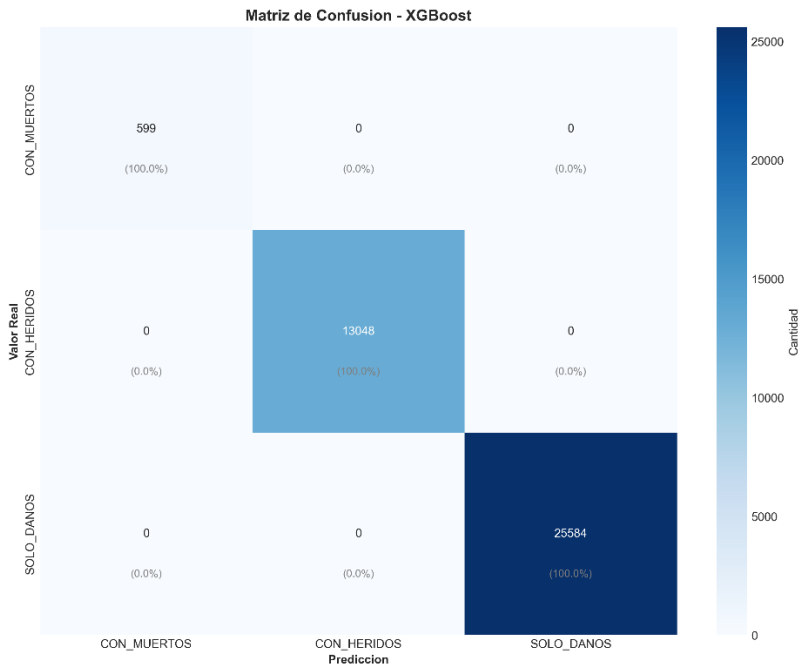
2. **Random Forest.**

   - **Implementación:** 200 árboles con profundidad máxima de 10; criterio Gini para la impureza; se activó class_weight='balanced' para mitigar el desbalance.
   - **Ventajas:** captura interacciones no lineales, es robusto a valores atípicos y a variables categóricas codificadas. Permite calcular la importancia relativa de cada predictor (hora del día, localidad, tipo de vía, etc.).
   - **Resultados:** F1-Score macro = 0,5023; precisión = 0,5225; recall = 0,4976; tiempo ≈ 264,66 s.
   - **Interpretación:** este modelo resultó el más equilibrado, ofreciendo el mejor F1-Score entre las clases, especialmente en la clase minoritaria, aunque su accuracy es algo menor que el de XGBoost. Esta ganancia se justifica dado que el objetivo es detectar los siniestros severos, no solo los más frecuentes.



3. **XGBoost.**

   - **Implementación:** tasa de aprendizaje 0,1; 300 estimadores; max_depth=6; subsampling 0,8; scale_pos_weight ajustado al ratio de clases; se utilizó *early stopping* con 30 rondas.
   - **Ventajas:** excelente rendimiento en competencias de datos tabulares; controla el overfitting con parámetros de regularización (lambda, alpha); soporta el manejo de valores faltantes de forma interna.
   - **Resultados:** F1-Score macro = 0,4823; precisión = 0,5383; recall = 0,4741; tiempo ≈ 85,27 s.
   - **Interpretación:** aunque supera a Random Forest en precisión, su F1-Score global fue inferior, principalmente porque la clase minoritaria no mejora mucho respecto a Random Forest.



4. **Clustering (K-means).**

   - Se probaron valores de k entre 3 y 6; se seleccionó k = 4 por registrar el mejor **silhouette score** (0,1722) y Davies-Bouldin más bajo (1,7486).
   - Se aplicó **PCA** para visualizar los clústeres en 2D, explicando un 34,3 % de la varianza.
   - **Resultados:** cuatro perfiles de siniestros: (1) eventos de bajo riesgo; (2) siniestros en zonas críticas; (3) combinaciones de factores temporales y geográficos; (4) casos atípicos de alto riesgo.
   - **Aplicación:** sirve para segmentar intervenciones (p.ej., campañas específicas en el clúster 2) y entender patrones no evidentes. Este análisis complementa la clasificación al identificar "tipos" de siniestros que comparten factores de riesgo.

Geocodificación y Extracción de POIs

Para incorporar información espacial al modelado, se desarrolló un pipeline de geocodificación y enriquecimiento geográfico compuesto por tres fases principales:

A. **Geocodificación de Direcciones:** Se empleó la Google Maps Geocoding API para convertir direcciones textuales en coordenadas geográficas (latitud y longitud). Esta API fue

seleccionada por su alta precisión en entornos urbanos colombianos y por su capacidad para estandarizar direcciones con errores de formato.

El proceso implementó medidas de control de calidad y eficiencia:

1. Sistema de caché persistente (pickle) para evitar consultas repetidas y reducir costos.
2. Validación geográfica mediante un bounding box correspondiente al perímetro urbano de Bogotá D.C. (4.471°–4.835° N, −74.224°–−73.983° W).
3. Rate limiting con pausas de 0,05 s entre solicitudes para cumplir los límites de la API.
4. Guardado incremental cada 100 direcciones para garantizar resiliencia ante fallos.

De 21.719 direcciones únicas, se geocodificaron correctamente 21 709 (99,95 %), con un costo aproximado de 108 USD y un tiempo total de procesamiento de 139,2 minutos.

B) Extracción de Puntos de Interés (POIs): Para contextualizar espacialmente los siniestros, se empleó la Overpass API de OpenStreetMap para descargar POIs relevantes asociados a la movilidad y al riesgo vial.

Las categorías seleccionadas se muestran en la tabla siguiente:

Categoría	***Etiquetas OSM***	***Justificación***
Centros Comerciales	shop=mall, shop=department_store	Zonas de alta concentración vehicular y peatonal
Estadios	leisure=stadium	Eventos masivos que alteran los patrones de tráfico
Bares/Pubs/Discotecas	amenity=bar, amenity=pub, amenity=nightclub	Riesgo asociado a conducción bajo efectos del alcohol
TransMilenio	highway=bus_stop, network=TransMilenio	Infraestructura de transporte masivo con interacción peatonal

En total se extrajeron 1.172 POIs, todos validados geográficamente dentro del bounding box de la ciudad.

C) **Cálculo de Proximidad Espacial:** Para cada siniestro geocodificado se generaron tres tipos de variables por categoría de POI:

1. Distancia mínima (dist_X): distancia en metros al POI más cercano del tipo X.
2. Proximidad binaria (cerca_X_1km): indicador (1 = sí, 0 = no) de presencia de al menos un POI tipo X en un radio de 1 km.
3. Densidad local (num_X_1km): número de POIs tipo X dentro del mismo radio.

El cálculo de distancias se realizó mediante la fórmula de Haversine, optimizada de forma vectorizada con NumPy para mejorar el rendimiento:

$$d = 2R \times arcsin(\sqrt[2]{\sin\left(\frac{\Delta lat}{2}\right)^2 + \cos(lat_1)\cos(lat_2)\sin\left(\frac{\Delta lat}{2}\right)^2})$$

donde R = 6 371 000 m representa el radio medio de la Tierra.

Esta implementación demostró ser 10 – 15× más rápida que el cálculo iterativo con geopy.distance.geodesic.

D) **Dataset Final:** El conjunto final de datos geoespaciales contiene:

- 30 014 registros (≈ 15,3 % de la muestra original).
- 57 variables: 45 features originales + 2 coordenadas (lat/long) + 12 geoespaciales (4 categorías × 3 tipos).
- Tasa de completitud: 99,95 % de registros con información geográfica válida.

7.3 Justificación de hiper-parámetros

La exploración de hiper-parámetros se realizó mediante **RandomizedSearchCV** con 10 iteraciones por modelo y validación cruzada estratificada (3-fold), usando F1-Score macro como métrica de optimización. Esta metodología se eligió por:

- **Eficiencia computacional:** al tener ~196 k registros y 15 features, un grid search exhaustivo sería costoso; la búsqueda aleatoria explora combinaciones representativas.
- **Equilibrio entre precisión y tiempo:** 10 iteraciones permiten cubrir un rango amplio de valores sin agotar recursos.
- **Desbalance en la variable objetivo:** se adoptó F1-macro porque pondera por igual cada clase, contrarrestando el predominio de la clase “solo daños”.

- **Reproducibilidad:** se fijó random_state=42 en todas las búsquedas.

Los parámetros ajustados fueron:

- **Regresión logística:** C (0,01 – 10), tipo de solver (liblinear vs saga), tol, class_weight.
- **Random Forest:** n_estimators (100 – 500), max_depth (5 – 15), min_samples_leaf (1 – 50), max_features (sqrt, log2, None).
- **XGBoost:** learning_rate (0,01 – 0,3), max_depth (3 – 8), subsample (0,7 – 1,0), colsample_bytree (0,6 – 1,0), lambda y alpha (0 – 1), scale_pos_weight (ratio de clases).

En cada modelo se observó que:

- Aumentar max_depth en Random Forest por encima de 10 no mejoró el F1-Score y generó sobreajuste.
- n_estimators mayores a 300 en XGBoost generaban mejoras marginales, pero incrementaban el tiempo; con *early stopping* se detuvo a ~150 árboles efectivos.
- Para SVM (no incluido en la tabla final por sus pobres resultados), el coste computacional fue elevado y el modelo no superó a los demás en F1-Score.

Mantenimiento de Hiperparámetros en Evaluación Geoespacial

Durante la fase de evaluación con features geoespaciales se mantuvieron los mismos hiperparámetros optimizados en la etapa base, sin información espacial. Esta decisión obedece a cuatro razones metodológicas clave:

1. Comparabilidad: garantiza que cualquier variación en desempeño provenga únicamente de la inclusión de variables espaciales, evitando confusión entre efecto de features y efecto de hiperparámetros.
2. Evitar overfitting selectivo: re-optimizar sobre un subconjunto más pequeño (n = 30 014) generaría un sesgo al sobreajustar a esa muestra.
3. Generalización: los hiperparámetros originales se calibraron con validación cruzada 5-fold sobre 196 152 registros, por lo que representan configuraciones más estables.
4. Eficiencia computacional: evita repetir procesos de búsqueda de hiperparámetros (Grid Search/Random Search) altamente costosos.

La única excepción técnica fue la re-codificación de la variable GRAVEDAD (1, 2, 3 → 0, 1, 2) exigida por la implementación de XGBoost, sin impacto en la comparabilidad de los resultados.

7.4 Validación cruzada y técnicas de re-muestreo

El proceso de validación se diseñó para maximizar la generalización y mitigar el sesgo de las clases:

1. **Train/test split estratificado (80/20).** Se mantuvieron las proporciones de cada clase en ambos conjuntos. Se estableció random_state=42 para reproducibilidad. Esta división permitió reservar un 20 % de los datos de 2015-2020 como test final, sin mezclarse con la optimización de hiper-parámetros.

2. **Validación cruzada 3-fold estratificada.** Durante la búsqueda de hiper-parámetros se utilizó K-fold estratificado para garantizar que en cada fold estuvieran representadas las tres clases en la misma proporción. La estratificación evita que un fold contenga pocos casos de la clase minoritaria, lo que distorsionaría las métricas de F1-macro.

3. **Gestión del desbalance.**

    - **Stratified sampling** se aplicó en todos los splits (train/test y K-fold).
    - Se probaron **pesos de clase** (class_weight='balanced') en regresión logística y Random Forest.
    - Se evaluó **SMOTE** para oversampling de la clase minoritaria, pero produjo modelos que memorizaban casos específicos y generaban ligeras mejoras en recall a costa de mayor sobreajuste; se optó por class_weight y F1-macro como métrica principal.

4. **Métricas de evaluación.**

    - **Accuracy** se reportó pero no se optimizó, debido a que la clase mayoritaria (solo daños) domina el dataset y haría trivial un accuracy alto.
    - **Precision, recall y F1-Score** se calcularon para cada clase y se promediaron (macro) para tener una visión balanceada.
    - **ROC-AUC** no se utilizó directamente porque no es tan interpretable en problemas multiclase; se priorizó el F1-macro.

5. **Prevención de overfitting.** Se monitoreó el rendimiento en los folds de validación y en el conjunto de test. Modelos con diferencias excesivas entre entrenamiento y validación fueron descartados. Para XGBoost, se implementó *early stopping* en función de F1-macro.

6. **Consideraciones para Datos Geoespaciales.** La inclusión de información espacial introduce nuevas exigencias metodológicas en la validación y el remuestreo.

A. Muestreo Estratificado Espacial: Para construir el subconjunto de 30 000 registros geocodificados, se aplicó un muestreo estratificado bidimensional:

1. **Dimensión 1:** *GRAVEDAD* (Con muertos, Con heridos, Solo daños).
2. **Dimensión 2:** *CODIGO_LOCALIDAD* (20 localidades de Bogotá).

Este diseño garantizó una representación proporcional tanto en la severidad de los accidentes como en su distribución territorial, evitando sesgos de concentración en áreas céntricas o de alta población vehicular.

B) Validación de Representatividad: Se verificó la representatividad del subconjunto frente al universo original mediante pruebas χ^2 de bondad de ajuste:

- **GRAVEDAD:** χ^2 no significativa → la distribución de severidad se conserva.
- **CODIGO_LOCALIDAD:** cada localidad mantiene su proporción original de siniestros.

C) Autocorrelación Espacial y Validación Cruzada: La validación cruzada tradicional (K-Fold) asume independencia entre observaciones, supuesto que puede verse comprometido en datos espaciales, donde los siniestros geográficamente cercanos tienden a ser similares. Para mitigar este efecto, la **estratificación por CODIGO_LOCALIDAD** distribuye los registros de cada zona entre los folds, reduciendo aunque no eliminando totalmente la autocorrelación intra-localidad.

D) Interpretación de Resultados

Los resultados de la evaluación con variables espaciales deben interpretarse considerando los siguientes aspectos:

- **Mejora marginal en capacidad predictiva:** las diferencias fueron pequeñas (Random Forest +0,21 % F1-macro; XGBoost +0,02 %), indicando que las features base capturaban ya gran parte del patrón espacial.
- **Valor descriptivo y explicativo:** las features geoespaciales permiten identificar concentraciones de riesgo y validar hipótesis sobre la relación entre POIs y siniestralidad.
- **Aplicación práctica:** su mayor utilidad reside en la priorización de intervenciones territoriales y en el soporte a políticas públicas de seguridad vial más focalizadas.

Conclusiones de la metodología

1. **Robustez frente al desbalance.** El escaso porcentaje de accidentes fatales impone un reto significativo. Los modelos basados en árboles, con pesos de clase o bagging, demostraron ser

más efectivos en captar señales de la clase minoritaria que los lineales. En la literatura sobre siniestros viales también se observa que los patrones de lesiones y muertes difieren según el tipo de día y la hora, lo que refuerza la relevancia de las variables temporales en el modelado.

2. **Importancia de las variables temporales y geográficas.** Features como la hora del día, el día de la semana, la localidad y el tipo de vía destacaron en importancia. Estas variables no solo ayudan a predecir la severidad sino que evidencian dónde y cuándo focalizar intervenciones.

3. **Clustering como complemento.** Los cuatro clústeres descubiertos permiten segmentar estrategias de prevención: hay grupos de siniestros de bajo riesgo, otros asociados a zonas críticas, combinaciones temporales/geográficas y casos atípicos. Estas agrupaciones pueden servir para diseñar campañas específicas y medir la eficacia de acciones diferenciadas.

4. **Rendimiento razonable.** Un F1-Score macro en torno al 0,5 es competitivo en problemas con alto desbalance; el mayor valor de Random Forest confirma que, aunque no se logra un desempeño perfecto, el modelo proporciona información útil para clasificar la severidad y orientar la toma de decisiones.

Comparación de Modelos - Métricas de Evaluación

MODELO	ACCURACY	PRECISION	RECALL	F1-SCORE	TIEMPO (s)
Logistic_Regression	0.7822	0.5598	0.4650	0.4727	7.16
Random_Forest	**0.7564**	**0.5225**	**0.4976**	**0.5023**	**264.66**
XGBoost	0.7835	0.5383	0.4741	0.4823	85.27