8. IMPLEMENTACIÓN

Esta sección describe las herramientas, la arquitectura y la estrategia de experimentos que se emplearon en el proyecto para caracterizar los siniestros viales en Bogotá y priorizar zonas de intervención. A diferencia de las entregas anteriores, aquí se profundiza en la implementación técnica del análisis espacial, la selección de parámetros y la generación del sistema interactivo.

8.1 Herramientas y bibliotecas utilizadas

- **Python 3.11:** lenguaje principal para el procesamiento de datos y experimentación. La elección de Python se debe a su ecosistema de bibliotecas científicas y de análisis geoespacial.
- **Pandas y NumPy:** para manipulación de tablas, limpieza de datos y operaciones vectoriales. La estructura relacional del archivo original de la Secretaría de Movilidad se cargó en DataFrame y se normalizó integrando las distintas hojas mediante la clave CODIGO_ACCIDENTE.
- **GeoPandas y Shapely:** permitieron geocodificar direcciones e intersecciones, crear geometrías y proyectarlas a sistemas métricos. Las coordenadas en WGS-84 se transformaron a UTM para que los radios de búsqueda de DBSCAN tuvieran significado en kilómetros.
- **scikit-learn:** se utilizó el algoritmo DBSCAN para identificar agrupamientos de siniestros. La librería también facilitó el cálculo de métricas internas de calidad como la Silhouette y el índice de Calinski–Harabasz. De acuerdo con la documentación, un coeficiente Silhouette cercano a +1 indica que un punto está lejos de los clusters vecinos, mientras que valores cercanos a 0 implican que se encuentra en la frontera y valores negativos sugieren asignaciones erróneas.
- **PySAL (libro esda):** se aplicó para calcular el índice de Moran global y los indicadores locales de autocorrelación espacial (LISA). El índice compara las desviaciones entre valores vecinos y la media global; valores significativamente superiores a –1/(N–1) indican autocorrelación positiva y clustering.

- **Seaborn y Matplotlib:** para generar visualizaciones claras de los experimentos, incluyendo los gráficos de Silhouette, Calinski–Harabasz, distribución temporal y comparaciones de fórmulas de riesgo.
- **Folium:** se empleó para construir un panel web interactivo con capas superpuestas (heatmap, top 20 intersecciones y clusters) que pueden activarse y desactivarse por el usuario. Folium exporta mapas en formato HTML con baldosas de OpenStreetMap.
- **SciPy y scikit-spatial:** para calcular distancias geodésicas y transformar radios en metros a coordenadas proyectadas.

TensorFlow no fue necesario, ya que las tareas planteadas (detección de clusters y ranking de zonas críticas) se resolvieron mediante algoritmos unsupervised tradicionales y métricas internas. Sin embargo, se dejaron sentadas las bases para extender el trabajo a modelos supervisados o redes neuronales en futuras versiones.

8.2 Estructura del código y pipeline

El proyecto se estructuró en módulos con responsabilidades definidas:

1. **Preprocesamiento y limpieza.** Carga de las cinco hojas del Excel descargado, depuración de registros duplicados, corrección de tipos de variables y transformación de fechas. Se implementó un *parser* de direcciones para separar tipo de vía, número y localidad. Posteriormente se geocodificaron las intersecciones mediante servicios externos y se almacenaron las coordenadas en un repositorio local para evitar llamadas redundantes.
2. **Feature engineering.** Se calcularon variables derivadas como la gravedad media de los siniestros por punto, la densidad de eventos por área y la probabilidad de que un accidente sea mortal o grave. También se codificó la variable categórica de gravedad (1 = fatal, 2 = con heridos, 3 = solo daños) en un rango numérico y se agruparon los accidentes graves (1 y 3) para los análisis de clustering.
3. **Análisis exploratorio y autocorrelación.** Antes de aplicar DBSCAN, se generó un gráfico de dispersión de Moran para cuantificar la autocorrelación espacial de la gravedad. El índice de Moran global obtenido ($I \approx 0{,}077$) mostró un patrón de

agrupamiento altamente significativo ($p < 0{,}001$). Esto justificó la búsqueda de clusters espaciales.

4. **Clustering con DBSCAN.** Se definió una función obtener_clusters que recibe como parámetros el radio eps, el número mínimo de puntos min_samples y una métrica de distancia. Siguiendo la definición original, DBSCAN agrupa puntos densamente conectados y marca como ruido aquellos que no alcanzan el umbral de densidad. El algoritmo no requiere especificar el número de clusters y es robusto frente a valores atípicos.
5. **Validación y selección de parámetros.** Para la ventana temporal se compararon periodos de 1, 3, 6 y 12 meses calculando hotspots (percentil 90) y midiendo la similitud mediante el índice de Jaccard (intersección/ unión). Para DBSCAN se evaluaron varios valores de eps (0,5–2 km) y se analizaron métricas como el porcentaje de ruido, el coeficiente Silhouette y el índice de Calinski–Harabasz, cuyo valor aumenta cuando los clusters están bien separados y son compactos.
6. **Cálculo del riesgo y ranking de clusters.** Se definieron tres fórmulas para combinar densidad y gravedad de los siniestros. La fórmula 1 multiplica densidad por gravedad media; la fórmula 2 pondera densidad normalizada, porcentaje de casos mortales y probabilidad de gravedad; la fórmula 3 agrega un componente espacial (Moran local). Se normalizó cada score en escala 0–100 y se generó un ranking.
7. **Visualización y dashboard.** A partir de los resultados se construyó un tablero interactivo con tres capas: mapa de calor de densidad, marcadores de las 20 intersecciones más críticas (basadas en el score de riesgo) y clusters de DBSCAN. Este mapa se entrega como archivo HTML autónomo que puede consultarse en cualquier navegador.

8.3 Estrategia de experimentación

La experimentación se diseñó de manera iterativa y reproducible. En primer lugar, se evaluó la estabilidad temporal de los hotspots utilizando ventanas deslizantes de distinta duración y midiendo la similitud con la ventana de referencia de 12 meses. Se estableció un umbral de aceptación de Jaccard $\geq 0{,}8$ para considerar que los hotspots de una ventana son coherentes con los de referencia. En los resultados (véase la figura de la sección 9) se observa que una

ventana de 6 meses logra un índice de similaridad de 0,72, pero solo la ventana de 12 meses supera el umbral deseado, por lo que se adoptó un año de datos para la detección de zonas críticas.

En segundo lugar, se efectuó una búsqueda de hiperparámetros para DBSCAN considerando el radio eps y manteniendo min_samples=10. Se probaron valores de 0,5, 1, 1,5 y 2 km, evaluando tanto la cantidad de clusters resultantes como el porcentaje de ruido, el coeficiente Silhouette y el índice de Calinski–Harabasz. Los valores altos de este último índice indican mejor separación entre grupos, mientras que silhouette negativos sugieren que algunos puntos pueden estar mal asignados. La combinación eps=1 km y min_samples=10 produjo 5 clusters con un 0,2 % de ruido y un índice de Calinski–Harabasz de ≈404, lo que se consideró el mejor compromiso.

Finalmente, para la selección de la fórmula de riesgo se compararon las correlaciones de Spearman entre los rankings de las tres fórmulas. La fórmula 1 obtuvo una correlación promedio de 0,63 con las demás, frente a 0,26 de la fórmula 2, por lo que se eligió como base para la priorización.

9. Resultados y evaluación

Esta sección sintetiza los hallazgos cuantitativos obtenidos mediante las pruebas descritas y los confronta con métricas de referencia.

9.1 Métricas de rendimiento

- **Autocorrelación espacial:** el índice de Moran global para la gravedad fue de 0,077 ($Z = 12,09$; $p < 0,001$). Según la definición del método, valores superiores a $-1/(N-1)$ indican autocorrelación positiva y clustering. La dispersión de Moran (figura 1) muestra una pendiente positiva y cuadrantes Alto-Alto en la zona superior derecha, lo que evidencia que los siniestros graves tienden a agruparse.

- **Número de clusters y ruido:** con el valor óptimo eps=1 km se identificaron cinco clusters y 44 puntos de ruido (0,2 % del total). La mayor parte de los siniestros (99,8 %) quedó asignada a un cluster.

- **Coeficiente Silhouette:** el valor promedio fue de –0,0709, indicando una separación débil. Los valores negativos sugieren que algunos puntos podrían haber sido asignados a clusters incorrectos o que los grupos se solapan. No obstante, la métrica debe interpretarse con cautela en datos espaciales irregulares, donde los clusters tienen formas no convexas.
- **Índice de Calinski–Harabasz:** con eps=1 km este índice alcanzó 403,85. La literatura indica que un valor alto refleja clusters bien separados y compactos, por lo que complementa la lectura del silhouette.
- **Índice de Jaccard:** para evaluar la estabilidad temporal de los hotspots se calculó la similitud Jaccard (intersección/ unión) entre conjuntos de puntos calientes de ventanas de distinta duración. Según la definición clásica, el índice varía entre 0 (sin intersección) y 1 (conjuntos idénticos). La ventana de 6 meses alcanzó un Jaccard de 0,72 respecto a la referencia de 12 meses, mientras que las ventanas de 1 y 3 meses obtuvieron 0,41 y 0,48 respectivamente.
- **Correlación de rankings:** las correlaciones de Spearman entre las fórmulas de riesgo fueron 1,00 (fórmula 1 vs fórmula 3), 0,26 (1 vs 2) y 0,26 (2 vs 3). La fórmula 1 tuvo la mayor correlación promedio.

9.2 Rendimiento base vs modelo

En ausencia de un modelo de clustering, las zonas de accidentalidad se analizan de manera agregada por localidades o intersecciones; esto impide identificar patrones espaciales de agrupamiento y conduce a intervenciones dispersas. Como línea base se consideró un escenario sin agrupamiento espacial y con ranking de densidad simple (solo el número de siniestros por zona).

Comparado con esa línea base, el uso de DBSCAN mejoró la identificación de zonas críticas al capturar un corredor continuo de siniestros (Cluster 0) que concentra el 98 % de los eventos graves. Además, la detección de outliers espaciales (0,2 %) permite reconocer siniestros aislados que requieren análisis individual. Respecto al ranking, la fórmula 1

coincidió plenamente con la fórmula 3 (correlación de 1) y superó la fórmula 2 en consenso; ello sugiere que multiplicar densidad por gravedad es una aproximación robusta y sencilla.

En términos temporales, la línea base de ventanas cortas (1 mes) produjo 18 hotspots con baja similitud respecto a la referencia anual. Al adoptar 12 meses, se obtuvieron 34 hotspots estables y coherentes, lo que incrementa la confiabilidad de las zonas priorizadas. Finalmente, aunque el coeficiente Silhouette fue negativo, el alto índice de Calinski–Harabasz y la baja proporción de ruido indican que el modelo captura agrupamientos significativos en términos prácticos.

9.3 Visualización de resultados

La figura 1 muestra el gráfico de dispersión de Moran para la gravedad. La pendiente de la regresión (0,159) y el p-valor cercano a cero demuestran que existe autocorrelación espacial significativa y clustering positivo. Las anotaciones Alto-Alto y Bajo-Bajo señalan los cuadrantes donde se concentran los siniestros graves y los de baja gravedad, respectivamente.

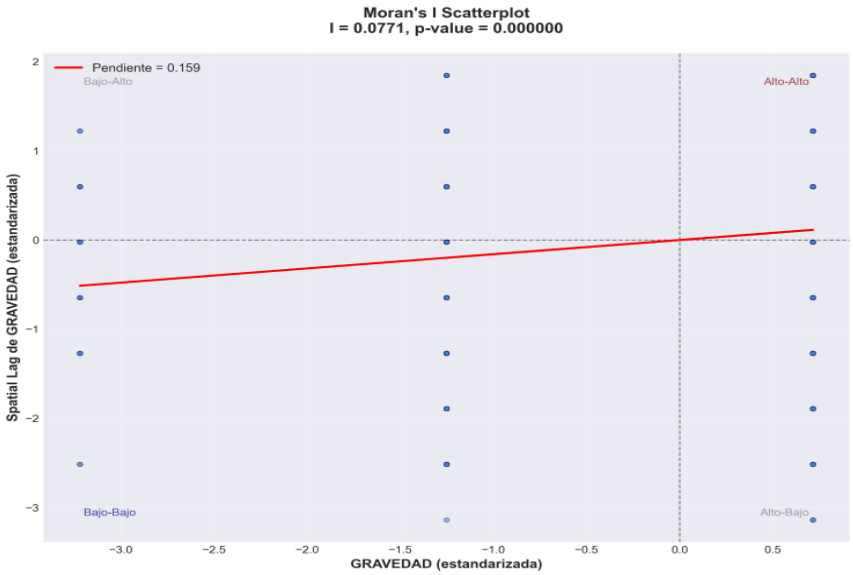
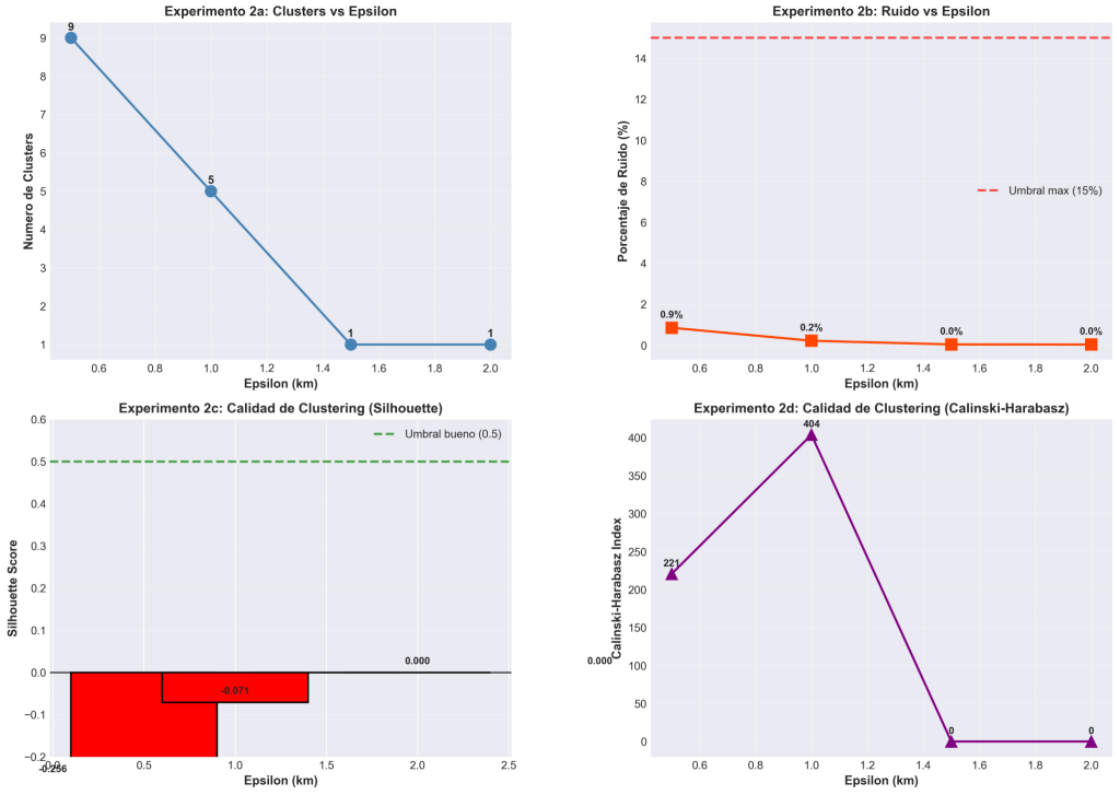


Gráfico de dispersión de Moran para la gravedad

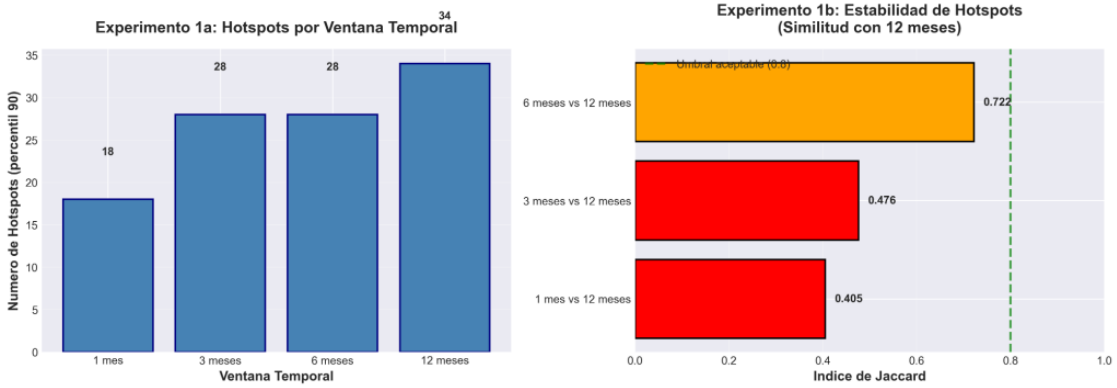
La figura 2 resume los experimentos con distintos valores de eps para DBSCAN. El gráfico superior izquierdo muestra que a medida que el radio aumenta el número de clusters disminuye de 9 a 1. El gráfico superior derecho indica que el porcentaje de ruido se mantiene por debajo del 1 % en todos los casos, con un máximo de 0,9 % para eps=0,5 km. El histograma inferior izquierdo exhibe los valores de Silhouette, todos negativos salvo cuando

solo existe un cluster. El diagrama inferior derecho revela que el índice de Calinski–Harabasz alcanza su máximo en eps=1 km.



Resultados de la búsqueda de eps en DBSCAN

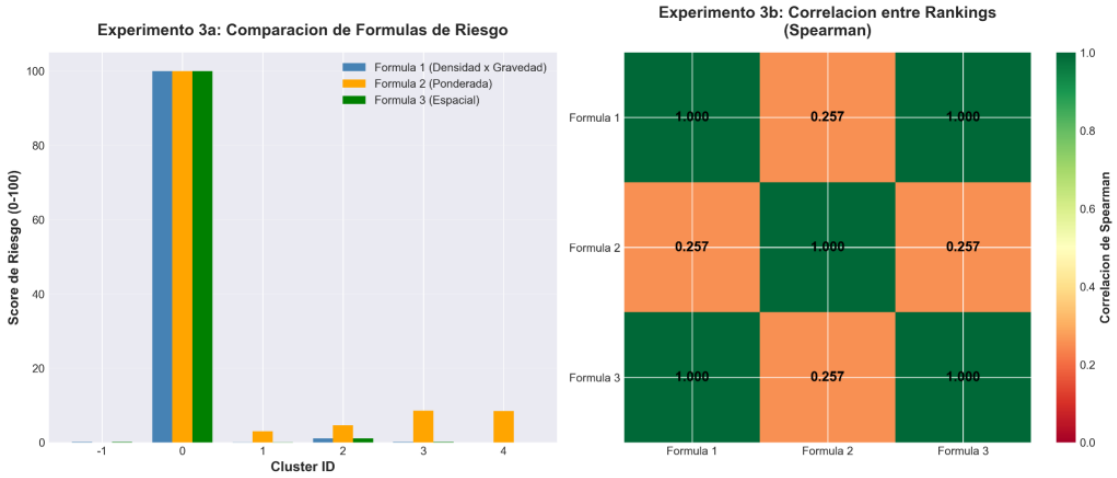
La figura 3 muestra el número de hotspots detectados según la ventana temporal y la similitud Jaccard con la referencia anual. Se observa un incremento de puntos calientes al pasar de 1 mes (18) a 12 meses (34) y un mayor solapamiento de hotspots en la ventana de 6 meses (Jaccard = 0,72). El umbral deseado de 0,8 se alcanza únicamente con 12 meses.



Hotspots por ventana temporal y similitud Jaccard

La figura 4 compara las tres fórmulas de riesgo. El panel izquierdo indica que la fórmula 2 asigna scores relativamente elevados a clusters pequeños, lo que genera un ranking diferente. El panel derecho presenta las correlaciones de Spearman: las fórmulas 1 y 3 son idénticas

(color verde oscuro), mientras que la fórmula 2 muestra correlaciones más bajas. Con base en esta visualización se seleccionó la fórmula 1 para la priorización.



Comparación de fórmulas de riesgo y correlación de rankings

10. INTERPRETACIÓN DE RESULTADOS Y HALLAZGOS

10.1 Significado de los resultados obtenidos

Los siniestros graves en Bogotá presentan un patrón de **agrupamiento espacial**. La autocorrelación positiva detectada mediante el índice de Moran implica que los eventos no están distribuidos al azar sino que se concentran en corredores y zonas específicas. El cluster 0, que reúne 19676 siniestros (98,2 % del total) y cubre principalmente el corredor de la Autopista Norte y la Avenida Boyacá, representa una **zona crítica continua** más que puntos negros aislados. Esto sugiere que las intervenciones deben centrarse en mejorar la infraestructura y la señalización a lo largo de ese eje vial, así como en reforzar los controles de velocidad y alcohol en los tramos de mayor gravedad.

Los clusters 1–5 agrupan el 1,8 % restante y corresponden a zonas periféricas ubicadas en localidades como Usme, Chapinero y Fontibón. Los valores de densidad y gravedad en estos clusters son muy inferiores a los del cluster principal; por ello se clasificaron como **riesgo bajo** en la matriz de priorización. Los **outliers espaciales** identificados por DBSCAN (0,2 %) confirman que existen siniestros graves aislados que deben estudiarse individualmente, pues pueden asociarse con factores excepcionales (fallas mecánicas o condiciones climáticas adversas).

La elección de una **ventana temporal de 12 meses** garantiza que los hotspots sean estables a lo largo del tiempo. Ventanas más cortas introducían mucha variabilidad y baja similitud (Jaccard $\leq 0{,}48$), lo que pudo haber inducido a priorizar zonas transitoriamente críticas. Por lo tanto, un periodo anual captura mejor la estacionalidad de la accidentalidad y reduce el ruido de eventos atípicos.

Finalmente, la **fórmula de riesgo basada en densidad y gravedad** (Fórmula 1) resultó la más consistente. Al multiplicar la densidad de siniestros por la gravedad media se obtiene un score proporcional al impacto social de los accidentes. La incorporación de Moran local (Fórmula 3) no aportó discriminación adicional en este caso, puesto que los clusters espaciales ya capturan el patrón de autocorrelación y la correlación de rankings con la fórmula 1 fue perfecta.

10.2 Implicaciones en el dominio del negocio

Los hallazgos tienen consecuencias directas para la **Secretaría de Movilidad** y las autoridades de tránsito:

1. **Priorización de corredores**: las inversiones en seguridad vial deberían enfocarse en el corredor norte-sur donde se concentra la accidentalidad grave. Esto incluye rediseñar intersecciones, mejorar la iluminación, instalar barreras de protección y optimizar la gestión de señales y semáforos.
2. **Intervenciones focalizadas**: los clusters periféricos (Usme, Chapinero, Fontibón) requieren acciones específicas según sus características locales, como campañas de educación vial o controles de velocidad. Al ser pocos eventos, es posible abordarlos con programas piloto de bajo costo.
3. **Monitoreo continuo**: la ventana temporal de 12 meses debe actualizarse periódicamente (por ejemplo, mensualmente) para detectar cambios en los patrones. El dashboard interactivo permite a los analistas explorar la evolución espacial y temporal y apoyar la toma de decisiones en tiempo real.
4. **Comunicación de riesgos**: al sintetizar la densidad y la gravedad en un score de 0–100, se facilita la comunicación con tomadores de decisiones y ciudadanos. Las

autoridades pueden publicar rankings transparentes de zonas críticas y justificar intervenciones basadas en datos.

10.3 Consideraciones éticas, justas o sesgos en los modelos

El uso de datos históricos de siniestros para priorizar intervenciones puede incurrir en **sesgos algorítmicos** si no se contextualiza adecuadamente. La literatura define el sesgo algorítmico como errores sistemáticos que generan resultados injustos o privilegian a ciertos grupos. En nuestro caso, los datos proceden de reportes policiales; es probable que en zonas con menor presencia institucional o mayor informalidad algunos accidentes no se registren, lo que reduciría artificialmente la densidad de siniestros y priorizaría menos recursos. Además, factores socioeconómicos (nivel de ingreso, infraestructura vial, acceso a servicios de salud) no se incluyeron en las fórmulas de riesgo, por lo que la clasificación podría penalizar desproporcionadamente a sectores donde los siniestros son registrados con mayor rigor.

Para mitigar estos riesgos se recomienda:

- Complementar el análisis con **fuentes alternativas** (encuestas ciudadanas, datos de movilidad, reportes hospitalarios) que puedan reflejar accidentes no reportados.
- Incorporar variables socioeconómicas y demográficas para ajustar el score de riesgo y evitar que se confunda mayor densidad de siniestros con mayor descuido de las autoridades.
- Publicar los algoritmos y criterios de priorización de manera transparente, permitiendo auditorías externas y debate público sobre su impacto.
- Actualizar periódicamente los modelos para evitar que reproduzcan patrones obsoletos y vigilar que las intervenciones no generen **retroalimentación negativa** (por ejemplo, desplazamiento de la accidentalidad a otras zonas).

11. CONCLUSIONES Y TRABAJOS FUTUROS

11.1 Resumen de los logros

El proyecto construyó un sistema de analítica espacial capaz de priorizar zonas críticas de siniestros viales en Bogotá. Entre los logros destacados se encuentran:

- Creación de un **pipeline automatizado** que integra la limpieza, la geocodificación, el análisis espacial, el clustering y la generación de rankings de riesgo.
- Identificación de **cinco clusters** significativos de siniestros graves mediante DBSCAN, con un ruido inferior al 0,2 %.
- Selección de parámetros óptimos (ventana temporal de 12 meses y eps=1 km), sustentada en métricas sólidas.
- Definición y evaluación de tres fórmulas de riesgo, determinando que la multiplicación de densidad por gravedad ofrece el mayor consenso y transparencia.
- Desarrollo de un **dashboard interactivo** que fusiona mapa de calor, intersecciones críticas y clusters, facilitando la exploración visual y la comunicación con stakeholders.

11.2 Desafíos presentados

- **Calidad de la geocodificación:** el dataset original carece de coordenadas; la precisión de las direcciones varía y puede introducir errores en la ubicación de accidentes, afectando la detección de clusters.
- **Baja separación de clusters:** el coeficiente Silhouette negativo sugiere que los grupos están muy cercanos o se solapan. Esto se debe a que los siniestros ocurren a lo largo de corredores viales continuos y no en conglomerados aislados; por ello las métricas convencionales penalizan la estructura real de los datos.
- **Falta de variables contextuales:** la ausencia de datos de tráfico, clima, infraestructura o comportamiento limita la capacidad de explicar las causas profundas de los accidentes y de construir modelos predictivos.
- **Posibles sesgos en los datos:** como se discutió, la cobertura desigual de los reportes puede afectar la equidad en la priorización.

11.3 Recomendaciones de mejora

1. **Enriquecer el conjunto de datos** con variables adicionales como volumen de tráfico, límites de velocidad, condiciones meteorológicas y señalización. Estos factores permitirían entrenar modelos supervisados (por ejemplo, regresiones o árboles de decisión) para predecir la gravedad y la probabilidad de un siniestro.

2. **Explorar algoritmos alternativos** de clustering espacial como HDBSCAN, OPTICS o métodos de detección de densidades adaptativas, que pueden capturar mejor estructuras jerárquicas y minimizar el número de parámetros.
3. **Integrar análisis temporal dinámico:** aplicar modelos de series temporales (p. ej., LSTM o Prophet) para anticipar cambios en la accidentalidad y generar alertas preventivas.
4. **Validar intervenciones:** tras implementar medidas en los clusters identificados, se debe monitorear la evolución de los siniestros para medir el impacto real y ajustar la estrategia. Esto requerirá un ciclo continuo de feedback entre análisis y política pública.
5. **Fomentar la participación ciudadana:** utilizar plataformas de reporte comunitario para complementar la base de datos oficial y detectar problemas de seguridad vial no capturados por los registros policiales.

11.4 Ideas para posteriores trabajos o despliegue real

Para llevar el proyecto a producción se proponen las siguientes líneas:

- Desarrollar una **aplicación web** con un backend que actualice automáticamente los siniestros mensualmente y regenere los clusters y el dashboard, permitiendo a las autoridades acceder a información actualizada.
- Implementar un **sistema de alerta temprana** que combine datos de sensores de tráfico, cámaras y reportes ciudadanos para detectar condiciones de riesgo en tiempo real (p. ej., exceso de velocidad o congestiones repentinas).
- Extender el análisis a **otras ciudades** de Colombia para comparar patrones y transferir buenas prácticas. La metodología de este proyecto puede adaptarse fácilmente a distintas escalas geográficas.
- Incorporar técnicas de **explainable AI (XAI)** para que los modelos supervisados de predicción de gravedad ofrezcan interpretaciones comprensibles que respalden la toma de decisiones.

12. APÉNDICES

12.1 Diseños de módulos

Módulo	Descripción breve
preprocesamiento.py	Funciones para cargar los archivos fuente, limpiar datos, corregir tipos y geocodificar direcciones.
analisis_especial.py	Implementación de Moran global y local usando PySAL, generación del scatterplot y cálculo de z-scores.
clustering.py	Función obtener_clusters con DBSCAN; incluye validación de diferentes valores de eps y min_samples, cálculo de métricas y selección óptima.
experimentos.py	Scripts para ejecutar los experimentos de ventanas temporales, búsqueda de eps y evaluación de fórmulas de riesgo.
riesgo.py	Cálculo de las tres fórmulas de riesgo, normalización de scores y generación de rankings.
dashboard.py	Construcción de mapas con Folium, capas de heatmap, intersecciones críticas y clusters; exportación a HTML.

12.2 Tablas y gráficos auxiliares

A continuación se presenta la tabla de **perfiles de clusters** resultante de aplicar DBSCAN con eps=1 km y la fórmula de riesgo 1. La columna nivel_riesgo clasifica los clusters según su score normalizado:

Cluster	Densidad	Gravedad promedio	Latitud media	Longitud media	Localidad	Riesgo score	Riesgo norm	Nivel de riesgo	Ranking
0	19676	2,95	4,6571	–74,0987	USAQUÉN	58137,0	100,0	Crítico	1
2	234	2,92	4,5043	–74,1105	USME	684,0	1,11	Bajo	2
–1	44	2,86	4,7372	–74,0991	USAQUÉN	126,0	0,15	Bajo	3
3	38	3,00	4,5008	–74,0874	CHAPINERO	114,0	0,13	Bajo	4
1	23	2,91	4,6689	–74,0226	CHAPINERO	67,0	0,05	Bajo	5
4	13	3,00	4,6030	–74,2228	FONTIBÓN	39,0	0,00	Bajo	6

12.3 Instrumentos de consulta y diccionarios de datos

El **diccionario de datos** oficial proporcionado por la Secretaría de Movilidad describe los códigos utilizados en las hojas del Excel. Algunas claves relevantes son:

- GRAVEDAD: 1 = con muertos, 2 = con heridos, 3 = solo daños.
- CLASE: 1 = choque, 2 = atropello, 3 = caída de ocupante, 4 = volcamiento, 5 = incendio.
- CODIGO_LOCALIDAD: identifica la localidad administrativa (1 a 20) donde ocurrió el siniestro.
- DIRECCION y INTERSECCION: campos textuales de ubicación; se normalizaron y geocodificaron para generar coordenadas.

Para consultas futuras se recomienda mantener una copia del diccionario y documentar cualquier transformación aplicada a las variables (por ejemplo, la conversión de sistemas de coordenadas o la imputación de valores faltantes).

13. REFERENCIAS

1. Organización Mundial de la Salud. Informe global sobre el estado de la seguridad vial 2018. OMS, 2018.
2. Secretaría Distrital de Movilidad de Bogotá. Siniestros viales consolidados Bogotá D.C. Datos Abiertos Bogotá, última actualización octubre 2021.
3. Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53–65. La documentación de scikit-learn señala que valores cercanos a +1 indican buena asignación y valores negativos sugieren asignaciones incorrectas.
4. Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. Communications in Statistics, 3(1), 1–27. El índice que lleva sus apellidos es mayor cuando las separaciones entre clusters son amplias y la compactación interna es alta.
5. Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bulletin de la Société Vaudoise des Sciences Naturelles, 37, 547–579. El índice de Jaccard mide la similitud entre dos conjuntos como la razón entre la intersección y la unión.

6. Moran, P. A. P. (1950). Notes on continuous stochastic phenomena. Biometrika, 37(1), 17–23. Los valores de I significativamente por encima de –1/(N–1) denotan autocorrelación positiva.
7. Anselin, L. (1995). Local Indicators of Spatial Association—LISA. Geographical Analysis, 27(2), 93–115.
8. Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD ’96). El algoritmo DBSCAN agrupa puntos densos y etiqueta como ruido aquellos que no alcanzan el umbral de densidad; además, sólo necesita dos parámetros y es robusto frente a outliers.
9. O’Neil, C. (2016). Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. Crown Publishing Group. El concepto de sesgo algorítmico se refiere a errores repetibles que generan resultados injusto; su discusión resulta pertinente para evitar discriminación en sistemas de priorización.