Pontificia Universidad
JAVERIANA
Cali

MODELO PREDICTIVO BASADO EN *MACHINE LEARNING* PARA PREDECIR LA CANTIDAD DE CASOS DE VIOLENCIA CONTRA LA MUJER EN LA CIUDAD DE BOGOTÁ

Estudiantes:
Mallory Yulieth Meneses Molina. Código 9026588
Samuel Melo Balcázar. Código 9025966

Proyecto Aplicado para optar al título de
Magíster en Ciencia de Datos

Director:
Dr. Diego Fernando Mosquera Valencia

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, MAYO DE 2026

FICHA RESUMEN

TÍTULO DEL PROYECTO: Modelo predictivo basado en *machine learning* para predecir la cantidad de casos de violencia contra la mujer en la ciudad de Bogotá

1. ÁREA DE TRABAJO:
2. TIPO DE PROYECTO (Aplicado, Innovación, Investigación): aplicado
3. ESTUDIANTE(S): y Mallory Yulieth Meneses Molina y Samuel Melo Balcázar
4. CORREO ELECTRÓNICO: mallorymeneses@javerianacali.edu.co y samuelmeloba@javerianacali.edu.co
5. DIRECCIÓN Y TELEFONO:
6. DIRECTOR: Diego Fernando Mosquera Valencia
7. VINCULACIÓN DEL DIRECTOR: Docente cátedra.
8. CORREO ELECTRÓNICO DEL DIRECTOR: dfmosquera@javerianacali.edu.co
9. CO-DIRECTOR (Si aplica):
10. GRUPO O EMPRESA QUE LO AVALA (Si aplica):
11. OTROS GRUPOS O EMPRESAS:
12. PALABRAS CLAVE (al menos 5): violencia, mujeres, homicidio, violencia intrafamiliar
13. FECHA DE INICIO: 2025
14. DURACIÓN ESTIMADA (En meses): 12 meses
15. RESUMEN:

El presente proyecto aplicado aborda la problemática de la violencia contra la mujer en Bogotá, enfocándose específicamente en los registros de violencia intrafamiliar y homicidios reportados por el Instituto Nacional de Medicina Legal y Ciencias Forenses (INMLCF). Este fenómeno constituye un problema social de carácter estructural y una violación sistemática de los derechos humanos, cuya persistencia vulnera la seguridad e integridad de las mujeres, lo que hace imperativo fortalecer la capacidad institucional de respuesta. La investigación tuvo como objetivo desarrollar un modelo basado en técnicas de aprendizaje automático que permitiera identificar patrones sociodemográficos, espaciales y de comportamiento, además de predecir la cantidad de casos en las diferentes localidades de la ciudad. Para su ejecución se implementó la metodología CRISP-DM, la cual permitió estructurar el proceso iniciando con la fase de comprensión del negocio, garantizando que el desarrollo técnico estuviera alineado con los objetivos estratégicos y las necesidades reales del entorno. Como principales resultados, se identificaron perfiles de victimización representativos y territorios con mayor incidencia, logrando modelos de segmentación y predicción capaces de anticipar tendencias futuras del fenómeno con alta consistencia técnica. Estos hallazgos demuestran la importancia de transitar de una gestión de datos puramente descriptiva hacia una capacidad de anticipación territorial fundamentada en la ciencia de datos. Las aplicaciones del proyecto se orientan al fortalecimiento de los sistemas de monitoreo, la priorización estratégica de recursos y la formulación de políticas públicas que mejoren la prevención activa de la violencia de género. Finalmente, se concluye que la integración

de modelos analíticos avanzados y herramientas de visualización permite transformar los registros históricos en activos de información crítica para la toma de decisiones, facilitando intervenciones estatales más oportunas, precisas y eficaces para salvaguardar la vida e integridad de las mujeres en Bogotá.

Contenido

Lista de tablas

Lista de imágenes

Lista de gráficos

Lista de anexos

Introducción

Este proyecto desarrolla modelos de aprendizaje automático aplicados al análisis de la violencia contra las mujeres en Bogotá, con el fin de disponer herramientas de apoyo en la toma de decisiones políticas y sociales fundamentadas en datos. A diferencia de los esfuerzos institucionales actuales, limitados a análisis descriptivos históricos, esta propuesta genera una herramienta proactiva orientada a la predicción del número de casos de violencia contra la mujer y la identificación de patrones sociodemográficos y de *modus operandi* de las formas de violencia, permitiendo intervenciones territorializadas efectivas. Bajo este propósito, la investigación resuelve: ¿de qué manera la ciencia de datos y los modelos de *machine learning* permiten identificar patrones relevantes en los casos de violencia contra las mujeres en Bogotá y generar capacidades predictivas sobre el comportamiento de estos eventos en las diferentes localidades?

Con el propósito de resolver este interrogante, se utilizan los datos abiertos del Instituto Nacional de Medicina Legal y Ciencias Forenses (INMLCF), integrando cerca de 40 variables que detallan el perfil de las víctimas, el agresor y el contexto de los hechos. El tratamiento de estos datos garantiza la sostenibilidad técnica del modelo, pues el uso de fuentes oficiales y actualizables asegura que la herramienta sea replicable y se mantenga vigente con una mínima inversión de recursos. Bajo un principio de rigor ético, el enfoque prioriza la predicción agregada sobre la estimación de riesgos individuales. En consecuencia, el desarrollo se articula en dos fases complementarias: primero, una etapa de caracterización y segmentación mediante modelos no supervisados (K-means) para clasificar patrones sociodemográficos e identificar perfiles de las víctimas de violencia; y segundo, un modelo predictivo basado en Random Forest Regressor, donde se procesa variables temporales y espaciales para anticipar el volumen de eventos de violencia contra la mujer, proporcionando un insumo técnico para el diseño de estrategias de prevención en la capital.

Capítulo I.
Contextualización del proyecto

1.1. Definición del problema

1.1.1. Planteamiento del problema

Este proyecto se enfoca en el análisis de la violencia intrafamiliar (VIF) hacia las mujeres en Bogotá, con el propósito de desarrollar modelos de *machine learning* (ML) orientados a la predicción de casos de violencia contra la mujer. El insumo principal consiste en los registros publicados por el Instituto Nacional de Medicina Legal y Ciencias Forenses (INMLCF) para el periodo 2015-2023.

Si bien este conjuntos de datos ofrece una visión exhaustiva sobre las víctimas y los contextos, su aprovechamiento institucional suele limitarse a la generación de estadísticas descriptivas y reportes periódicos. Esta situación genera una brecha crítica entre la disponibilidad de información y su uso estratégico para anticipar situaciones de riesgo, dificultando la orientación de políticas de prevención y la priorización de recursos en zonas o grupos poblacionales con mayor vulnerabilidad.

Este rezago representa una desventaja en Colombia, mientras que la literatura internacional documenta experiencias donde el uso de modelos de ML permite identificar factores de riesgo asociados a la violencia de género y focalizar acciones preventivas, en Bogotá persiste un vacío en la aplicación de estas metodologías a los registros del INMLCF. Existe, por tanto, una discrepancia entre el uso convencional de la información disponible sobre la violencia contra la mujer y el potencial que ofrece la ciencia de datos.

En la actualidad, el INMLCF se limita a un enfoque descriptivo que emplea tasas por cada 100.000 habitantes para facilitar la interpretación de la magnitud del fenómeno. Bajo este esquema, los reportes de 2021 para Bogotá evidencian disparidades críticas: en la violencia intrafamiliar (VIF) contra la infancia y adolescencia, la tasa en niños fue de 91,48, mientras que en niñas fue de 101,79. En cuanto a la población de adultos

mayores, se registraron tasas de 51,65 para hombres y 40,44 para mujeres. No obstante, la brecha más profunda se observa en la violencia de pareja, donde hubo una tasa de 45,84 para hombres frente a un alarmante 196,97 para mujeres durante el mismo periodo [1].

Para el año 2023, las tasas en la categoría de VIF en Bogotá alcanzaron 158,67 para niños y 167,02 para niñas. En la población de adultos mayores, se situaron en 50,24 para hombres y 45,03 para mujeres, mientras que la violencia de pareja registró tasas de 50,65 para hombres y 216,83 para mujeres.[1] Estos datos confirman una incidencia constante y, en categorías específicas, un incremento sostenido en el tiempo.

Resulta indispensable subrayar que, si bien se hace referencia a tasas por ser el estándar de los informes descriptivos del INMLCF, el objetivo central de este proyecto es predecir el volumen absoluto de casos. Este enfoque proporcionará una métrica de magnitud directa, optimizando así el diseño de estrategias para la planeación preventiva.

Pese a la persistencia de estas violencias y a la disponibilidad de registros oficiales, el país presenta un déficit en el desarrollo de herramientas predictivas que permitan a los tomadores de decisiones anticipar escenarios futuros. Los enfoques predominantes se limitan a análisis descriptivos y comparaciones históricas, [1] debido en gran medida a que el uso de datos "crudos" estatales impone barreras técnicas que exigen recursos significativos. Este escenario demanda un proceso riguroso de limpieza, transformación, integración y estandarización para asegurar la calidad de los datos de entrada en los algoritmos de *machine learning,* complejidad que explicaría la limitada proliferación de modelos analíticos avanzados en el contexto nacional.

El presente proyecto aplicado surge como respuesta al déficit de herramientas predictivas que permitan a los tomadores de decisiones anticipar escenarios futuros. Por ello, se plantea el diseño, entrenamiento y evaluación de modelos de aprendizaje automático capaces de generar predicciones robustas. El desafío central radica en

transformar los datos en conocimiento estratégico para comprender los factores asociados a la violencia contra la mujer y anticipar su ocurrencia. Mediante algoritmos de ML [2] es posible identificar variables críticas y predecir el volumen de casos de violencia contra la mujer en las localidades de Bogotá, basándose en dimensiones sociodemográficas y geográficas presentes en los registros.

Bajo este enfoque, resulta imperativo explorar patrones complejos para predecir el volumen casos de violencia intrafamiliar hacia las mujeres en Bogotá. Esto reviste una relevancia estratégica, ya que facilita la creación de herramientas de apoyo que optimicen la asignación de recursos y permitan focalizar acciones preventivas en las zonas y perfiles de mayor riesgo.

En última instancia, esta investigación trasciende el componente técnico para convertirse en un habilitador de políticas públicas con enfoque de género y precisión territorial. Al integrar la ciencia de datos con enfoque en la prevención, se establece un precedente metodológico para que Bogotá deje de limitarse al conteo de víctimas de violencia contra la mujer y comience a actuar con una capacidad de anticipación científica, orientada a salvaguardar la integridad de las mujeres y a transformar la gestión de la seguridad ciudadana en un modelo inteligente y proactivo.

1.1.2. Formulación del problema

Como se ha mencionado, la violencia contra la mujer en Bogotá exige un análisis profundo sustentado en los datos del INMLCF. Estos registros permiten caracterizar a las víctimas mediante variables clave como edad, localidad y ciclo vital; sin embargo, esta riqueza informativa se queda en un plano mayoritariamente descriptivo.

El problema central reside en que, a pesar de la disponibilidad de estos datos, existe una carencia de herramientas analíticas avanzadas que transformen el registro histórico en modelos de prevención proactiva. Esta brecha técnica limita la capacidad institucional para anticipar escenarios y optimizar recursos, lo que plantea el siguiente interrogante

¿de qué manera el uso de técnicas de análisis de datos, segmentación y modelos de *machine learning* permiten identificar patrones relevantes en los casos de violencia contra las mujeres en Bogotá y generar capacidades predictivas sobre el comportamiento de estos eventos en las diferentes localidades?

Esta pregunta delimita el objeto del proyecto en dos aspectos: por un lado, el conocimiento de los conjuntos de datos mediante una exploración técnica exhaustiva; y por el otro, la aplicación de modelos de ML y estadísticos para resolver, desde la ciencia de datos, el gran problema de la predicción de casos de violencia contra la mujer a partir de las variables disponibles.

Finalmente, es preciso subrayar que es técnicamente imposible, y jurídica y filosóficamente problemático, predecir delitos con detalles individuales para fines judiciales, por lo que en este proyecto no estimará el riesgo individual de las mujeres a sufrir violencias, sino la cantidad de casos a futuro a partir de los datos existentes, que son los registros de denuncias de posibles hechos ante una entidad estatal, la cual está facultada para esta función.

1.1.3. Sistematización

Para el desarrollo de esta actividad es necesario plantear una serie de preguntas de sistematización, las cuales ayudarán a orientar este proyecto. Estas son:

1. ¿Qué características sociodemográficas presentan las mujeres víctimas de violencia en Bogotá y cómo se relacionan con los patrones de *modus operandi*, ubicación geográfica y periodos de mayor incidencia?

2. ¿Qué variables son más relevantes para predecir la ocurrencia de nuevos casos de violencia hacia las mujeres y qué tan precisos son los modelos desarrollados?

3. ¿Cuáles modelos de aprendizaje automático son más adecuados para segmentar y perfilar a las mujeres víctimas de violencia en Bogotá, y por otro, predecir la cantidad de nuevos hechos de violencia hacia las mujeres en Bogotá?

4. ¿Qué tan efectivos son los modelos predictivos en términos de precisión, sensibilidad y aplicabilidad práctica para apoyar acciones de prevención de estos hechos?

5. ¿Cómo pueden presentarse los resultados del análisis a través de visualizaciones interactivas que faciliten su comprensión y aplicación en la toma de decisiones?

La primera pregunta permite conocer las características que tienen las presuntas víctimas en el periodo relacionado en el *dataset*. Esto es particularmente interesante, pues puede brindar variables dependientes para los modelos, además que es parte vital de los análisis que se puedan gestar. Lo anterior es así, puesto que se puede argüir que una persona que tenga características similares a las presentadas en el conjunto de datos tiene determinada probabilidad de sufrir determinada forma de violencia.

La segunda pregunta se orienta al análisis de las variables presentes en el conjunto de datos, con el fin de determinar cuáles aportan una mayor capacidad predictiva en la predicción de nuevos casos de violencia hacia las mujeres. Este análisis permite identificar relaciones entre variables sociodemográficas, espaciales, temporales y de contexto asociadas a los eventos de violencia registrados.

La tercera pregunta permite profundizar sobre los modelos de aprendizaje automático útiles para el fin propuesto en el presente proyecto aplicado. Esto quiere decir que se determinan los modelos a usar para realizar la segmentación y organización en grupos según las características de las víctimas. Además, se determinan aquellos modelos que permiten predecir la cantidad de casos, a partir de las localidades.

La cuarta pregunta permite realizar análisis geoespaciales, con los que se logra entender

la magnitud de la problemática en distintas localidades, en este caso, según el tipo de violencia; así, se pueden ver zonas que presentan más hechos en unos periodos de tiempo que otras, lo que enriquece la exploración de los datos.

La quinta pregunta permite ajustar y modelar, si es el caso, los datos obtenidos para la ejecución de los modelos a realizar en el marco de este proyecto aplicado. De este modo, se puede dar cierre al ciclo de los proyectos de ciencia de datos, que van desde el entendimiento de los datos hasta la realización e implementación de modelos para evaluarlos. Con este paso se estructura el factor diferencial de este proyecto: la estimación de los casos de violencia hacia la mujer en Bogotá.

1.2. Objetivos

1.2.1. Objetivo general

Desarrollar un modelo basado en técnicas de segmentación y *machine learning* que permita identificar patrones sociodemográficos, espaciales y de modus operandi en los casos de violencia contra las mujeres en Bogotá, así como analizar y predecir la cantidad de los casos en las diferentes localidades de la ciudad.

1.2.2. Objetivos específicos

- Realizar un análisis exploratorio de datos para caracterizar a las mujeres víctimas de violencia intrafamiliar y homicidios en Bogotá que ayude a la detección de patrones asociados al *modus operandi* y el análisis de distribuciones geográficas y temporales de los casos.

- Determinar los modelos de aprendizaje automático más adecuados para la segmentación y perfilamiento de las mujeres víctimas de violencia en Bogotá y, a la vez, la predicción de nuevos hechos de violencia hacia las mujeres en la ciudad.

- Implementar modelos de aprendizaje automático que permitan estimar el número de casos de hechos de violencia hacia mujeres para las localidades de Bogotá.

- Evaluar el desempeño de los modelos predictivos mediante métricas estadísticas que permitan la validación de la precisión, capacidad de generalización y utilidad para la toma de decisiones.

- Presentar los resultados del análisis a través de visualizaciones interactivas que permitan una interpretación accesible y apoyen la toma de decisiones informadas.

1.3. Marco de referencia

1.3.1. Marco teórico

El presente marco teórico reúne los fundamentos conceptuales y técnicos que sustentan el desarrollo de la investigación. En primer lugar, se abordan los conceptos relacionados con violencia intrafamiliar y violencia basada en género, como concepto transversal, tomando como referencia las definiciones empleadas por el Instituto Nacional de Medicina Legal y Ciencias Forenses (INMLCF), entidad responsable de la generación de los registros utilizados en el estudio. Posteriormente, se presentan los principales fundamentos de ciencia de datos y *machine learning* asociados a los modelos implementados en la investigación, incluyendo técnicas supervisadas y no supervisadas orientadas a procesos de predicción y segmentación. De esta manera, el marco teórico integra tanto la comprensión conceptual del fenómeno analizado como los elementos técnicos necesarios para el desarrollo de los modelos analíticos del proyecto.

1.3.1.1. Violencia intrafamiliar

La violencia intrafamiliar se define como “Todo daño o maltrato físico, psicológico, sexual, económico o patrimonial, producido entre miembros de la familia, aunque no convivan bajo el mismo techo”.[2] Este tipo de violencia comprende múltiples manifestaciones que afectan de manera directa la integridad y bienestar de las víctimas, incluyendo agresiones físicas, control económico, afectaciones emocionales, violencia sexual y conductas de intimidación o dominación dentro del entorno familiar.

Dentro de estas manifestaciones, la violencia física hace referencia a acciones que producen lesiones o afectaciones corporales, mientras que la violencia patrimonial y económica involucra conductas orientadas al control o limitación de recursos, bienes, documentos o ingresos económicos de la víctima[2]. Por su parte, la violencia sexual comprende cualquier acto o conducta de carácter sexual ejercida mediante coerción, intimidación, manipulación o amenaza, vulnerando los derechos sexuales y reproductivos de las personas afectadas [2]. Finalmente, la violencia psicológica se relaciona con acciones de desvalorización, autoritarismo, control y agresión verbal que impactan negativamente la estabilidad emocional y mental de los integrantes del núcleo familiar [2].

1.3.1.2. Violencia basada en género

Estas dinámicas de violencia suelen desarrollarse dentro de relaciones asimétricas de poder construidas social y culturalmente, particularmente alrededor de los roles de género.[2] En este sentido, la violencia basada en género puede entenderse como "cualquier acción u omisión ejercida contra una persona debido a su identidad, expresión de género u orientación sexual, generando afectaciones físicas, psicológicas, económicas, patrimoniales, sexuales o incluso la muerte".[3] Desde esta perspectiva, la violencia contra la mujer constituye una problemática estructural y multicausal que trasciende el ámbito individual, involucrando factores sociales, culturales y territoriales que influyen en su ocurrencia y comportamiento.

Si bien la violencia contra la mujer se enmarca en las violencias basadas en género, para este caso se toma como género el sexo de las personas, pues los datos no permiten hacer una desagregación en la orientación sexual de las víctimas de violencia. Por esta razón, se menciona que es víctima hacia la mujer, lo que involucra a niñas y adolescentes mujeres, sin considerar a personas trans.

Las definiciones anteriormente descritas constituyen el marco conceptual bajo el cual el

Instituto Nacional de Medicina Legal y Ciencias Forenses (INMLCF) realiza la clasificación, registro y consolidación de los casos utilizados en la presente investigación. En este sentido, dichos conceptos no solo permiten comprender la naturaleza del fenómeno estudiado, sino también interpretar adecuadamente la estructura y características de los conjuntos de datos analizados.

En consecuencia, el análisis de fenómenos complejos y multicausales como la violencia contra la mujer requiere herramientas capaces de identificar patrones, relaciones y tendencias a partir de grandes volúmenes de información. En este contexto, las técnicas de ciencia de datos y *machine learning* han adquirido relevancia como herramientas de apoyo para el análisis de comportamientos sociales complejos, permitiendo extraer conocimiento a partir de datos estructurados y no estructurados.

1.3.1.3. Machine learning

Dentro del campo del *machine learning*, los modelos pueden clasificarse principalmente en técnicas supervisadas y no supervisadas. Los modelos supervisados utilizan variables objetivo conocidas para realizar procesos de predicción o estimación, mientras que los modelos no supervisados buscan identificar estructuras, agrupamientos o patrones ocultos dentro de los datos sin depender de una variable de salida previamente definida.

1.3.1.4. Series temporales e ingeniería de características temporales

Una serie temporal corresponde a un conjunto de observaciones organizadas cronológicamente, en el cual el orden de los datos resulta fundamental para identificar patrones como tendencia, estacionalidad y dependencia entre periodos[5]. En el presente proyecto, la información se estructuró longitudinalmente mediante combinaciones localidad-mes, permitiendo representar la evolución temporal de los casos registrados y formular horizontes de predicción de uno, dos y tres meses.Modelos de aprendizaje supervisado

Para incorporar la dependencia histórica dentro de los modelos supervisados se construyeron variables rezagadas o *lag features*, las cuales utilizan valores observados en periodos anteriores como predictores del comportamiento futuro [5], [6] En particular, se emplearon rezagos de 1, 2, 3, 6 y 12 meses por localidad, con el propósito de capturar patrones de persistencia en diferentes escalas temporales.

Asimismo, se incorporaron medias y sumas móviles calculadas sobre ventanas históricas, orientadas a resumir el nivel promedio y el comportamiento acumulado reciente de la serie. Estas transformaciones permiten reducir el efecto de fluctuaciones puntuales y representar tendencias locales dentro de periodos definidos [5]. De manera complementaria, se utilizaron diferencias mensuales y variaciones porcentuales para caracterizar cambios absolutos y relativos en la evolución de los casos.

La estacionalidad hace referencia a patrones que se repiten de manera sistemática en determinados intervalos temporales[5]. Para representar estas dinámicas se incluyeron variables calendario, como mes y trimestre, además de transformaciones seno y coseno que conservan la naturaleza cíclica del tiempo y permiten representar adecuadamente la proximidad entre periodos consecutivos dentro del ciclo anual [7]. En conjunto, estas características permitieron incorporar información temporal en los modelos Random Forest y XGBoost sin recurrir exclusivamente a modelos clásicos de series temporales.

Los modelos de aprendizaje supervisado planteados son *random forest* y *XGBoost*. Estos algoritmos permiten estimar el número de casos según el proyecto en el cual se usen. Para el presente caso, se plantean para estimar el número de casos de violencia hacia la mujer en Bogotá.

1.3.1.4.1. Random forest

El *Random Forest* (Bosques Aleatorios) se define como una técnica de aprendizaje supervisado basada en el concepto de ensamble (*ensemble learning*), la cual busca incrementar la precisión y estabilidad de las predicciones mediante la combinación de

múltiples modelos individuales denominados árboles de decisión[8]. Esta metodología se fundamenta específicamente en el *Bagging* (*Bootstrap Aggregating*), un proceso donde cada árbol del "bosque" se ajusta de manera independiente utilizando un remuestreo aleatorio con reemplazo conocido como *Bootstrap[8]*. En esta etapa de simulación, se generan diversos subconjuntos de entrenamiento del mismo tamaño que el original, permitiendo que algunas observaciones se repitan en diferentes muestras mientras que otras quedan excluidas, lo que facilita la estimación de parámetros estadísticos y la robustez del modelo.[9]

La arquitectura de este modelo se desarrolla en etapas técnicas críticas: primero, se restringe la selección de variables en cada partición o nodo del árbol, cumpliendo la condición de que el número de predictoras evaluadas (*m*) sea menor al total de predictoras disponibles (M), lo que garantiza que los árboles crezcan de forma descorrelacionada.[9] Segundo, los árboles crecen hasta su máxima extensión para capturar relaciones complejas y no lineales dentro de los datos.[9] Un aspecto fundamental de esta técnica es que las observaciones no seleccionadas durante el *bootstrap*, denominadas *Out-of-Bag* (OOB), se utilizan como un conjunto de validación interno para medir la capacidad de generalización del modelo sin recurrir a datos externos adicionales en la fase inicial.[9]

Finalmente, el resultado del modelo o ensamble se obtiene mediante la agregación de las predicciones de todos los árboles generados (frecuentemente entre 500 y 1000 iteraciones) [10], [11]. Esta consolidación se realiza a través de dos mecanismos: por promedio, en casos de problemas numéricos o de regresión, o mediante el conteo de votos por mayoría, cuando se trata de variables categóricas o de clasificación. Gracias a esta estructura, el Random Forest logra mitigar eficazmente el riesgo de sobreajuste (*overfitting*) y se posiciona como una herramienta versátil para el análisis de fenómenos sociales complejos, permitiendo transformar grandes volúmenes de registros históricos en activos de información con alta capacidad predictiva.

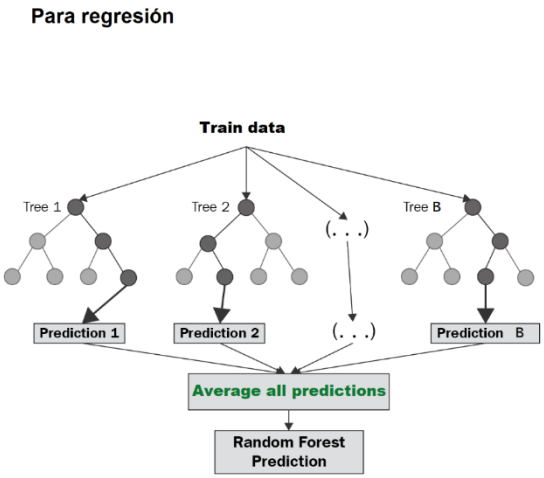


Imagen 1. Predicción de un random forest
Fuente:[12]

1.3.1.4.2. XGBoost

Aunque el Random Forest ofrece una base sólida mediante el uso de árboles independientes en paralelo para reducir la varianza, el campo del aprendizaje supervisado dispone de otros enfoques de ensamble orientados a mejorar la precisión mediante procesos secuenciales. Esta transición se encuentra en los algoritmos basados en Boosting, los cuales construyen modelos de manera iterativa donde cada nuevo árbol se especializa en corregir los errores o residuos generados por las etapas previas. Dentro de este grupo, el XGBoost (eXtreme Gradient Boosting) destaca como una implementación optimizada del aumento de gradiente, diseñado para manejar estructuras complejas de datos con alta eficiencia computacional.[13]

El fundamento teórico de este algoritmo reside en su función objetivo específica, la cual busca un equilibrio entre el poder predictivo y la simplicidad del modelo. A diferencia de otros métodos de boosting, XGBoost integra un término de regularización (Ω) directamente en el proceso de minimización de la función de pérdida (L), formulándose matemáticamente como:

$$Obj^{(t)} = \sum_{i=1}^{n} L(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t)$$

Imagen 2. Fórmula de XGBoost
Fuente: [14]

Esta fórmula muestra que el modelo se construye de forma secuencial, pues en cada iteración se agrega un árbol nuevo f_t para corregir los errores. En la fórmula se ve que n es el número de observaciones del conjunto de entrenamiento, i es el índice de cada observación, y_i es el valor real de la observación i, $\hat{y}_i^{(t-1)}$ es la predicción o corrección producida por el nuevo árbol para la observación x_i. $\hat{y}_i^{(t-1)}+f_t(x_i)$ es la nueva predicción, del modelo luego de añadir el árbol t. $L(\cdot)$ es la función de pérdida, que mide qué tan diferente es la predicción respecto al valor real. Ω^(f_t) término de regularización que penaliza la complejidad del árbol nuevo.

Donde el término de regularización penaliza la complejidad de los árboles para prevenir el sobreajuste (overfitting), permitiendo que el sistema alcance un comportamiento técnico óptimo sobre datos tabulares [14]. Convencionalmente, modelos basados en árboles potenciados por gradiente son considerados altamente eficientes en conjuntos de datos estructurados, especialmente cuando el número de observaciones supera significativamente al número de variables [15].

Un componente técnico fundamental de XGBoost es el uso de la expansión de Taylor de segundo orden para optimizar la función de pérdida. Mientras que el Gradient Boosting tradicional utiliza únicamente la primera derivada del error (gradiente), XGBoost incorpora tanto la primera como la segunda derivada (Hessian), permitiendo calcular de manera más precisa la ganancia asociada a cada partición dentro de los nodos del árbol y acelerando la convergencia hacia el mínimo global del error [14], [16].

No obstante, debido a su estructura de ensamble secuencial y a la complejidad interna de sus procesos de optimización, este tipo de modelos suele clasificarse dentro del

concepto de “caja negra” (*black box*) [17], [18]. Desde una perspectiva conceptual, esto implica que gran parte del procesamiento interno que conduce a una predicción que ocurre de manera autónoma y con alta complejidad matemática, dificultando la interpretación directa de cómo cada variable influye sobre el resultado final [19]. Esta característica introduce desafíos asociados a la explicabilidad, entendida como la capacidad de comprender, justificar y auditar las decisiones generadas por sistemas de inteligencia artificial [20].

1.3.1.5. Modelos de aprendizaje no supervisado

Los modelos de aprendizaje no supervisado usados se orientaron a la identificación de patrones y segmentaciones dentro de los registros analizados. Concretamente se usaron los modelos k-means y Birch, los cuales son de agrupamiento. A diferencia de los modelos supervisados antes vistos, cuyo objetivo principal consiste en estimar valores futuros o clasificar observaciones a partir de una variable objetivo previamente definida, los modelos de agrupamiento permiten descubrir estructuras ocultas, similitudes y comportamientos compartidos entre los datos sin necesidad de etiquetas previas.[21] Ese uso permite realizar cruces de datos para conocer relaciones entre los datos de manera mucho más eficiente que con técnicas de estadística descriptiva, como el análisis de pareto u otra forma.

1.3.1.5.1. K-means

Este algoritmo es ampliamente utilizado en aplicaciones de *machine learning* debido a su capacidad para segmentar conjuntos de datos en grupos homogéneos a partir de medidas de distancia entre observaciones [22]. Matemáticamente, el modelo busca particionar un conjunto de datos compuesto por (n) vectores ((x_1, x_2, ..., x_n)) en (k) grupos ((G_1, G_2, ..., G_k)), minimizando la distancia entre cada observación y el centroide asociado a su grupo. Su función objetivo se expresa de la siguiente manera:

$$\min_{s} \sum_{i=1}^{k} \sum_{x_j \in S_i} ||x_j - \bar{x}_i||^2$$

Imagen 3. Fórmula del k-means
Fuente:[23]

En la fórmula, *k* es el número de clústers que a formar. S_i (que es S_1, S_2, etc.) es el conjunto de observaciones asignadas al clúster *i*. x_i es la observación o punto de los datos. $\bar{x}_i$ es el centroide del clúster *i*, es decir, el promedio de los grupos de los puntos pertenecientes a dicho grupo. $||x_i - \bar{x}_i||^2$ es la distancia euclidiana al cuadrado entre la observación y el centroide del clúster. Min_s indica que el algoritmo busca la asignación a los clústers, que produzca el menor valor posible de la función.

El funcionamiento del algoritmo se basa en la asignación iterativa de cada observación al centroide más cercano, recalculando posteriormente los centroides hasta alcanzar estabilidad en las agrupaciones o minimizar el costo total intra-clúster [44]. Debido a esta lógica de funcionamiento, K-Means resulta especialmente útil para identificar perfiles o tipologías dentro de fenómenos sociales complejos, permitiendo reconocer agrupaciones con características sociodemográficas, espaciales o contextuales similares.

La siguiente es una representación gráfica del funcionamiento general del algoritmo:

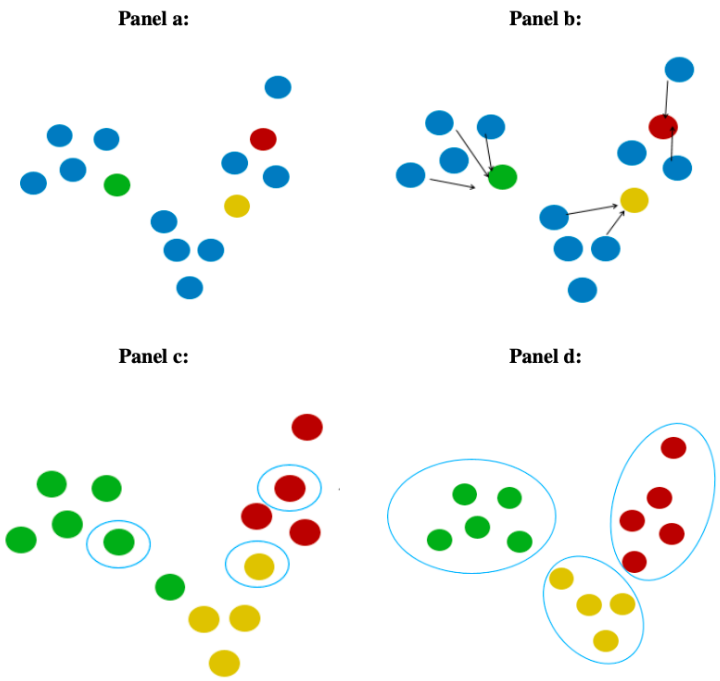


Imagen 4. Representación de un k-means
Fuente: [23].

1.3.1.5.2. BIRCH

Birch quiere decir Balanced Iterative Reducing and Clustering using Hierarchies. Es diseñado especialmente para trabajar con grandes volúmenes de datos y estructuras de alta dimensionalidad [24]. Su objetivo es esumir progresivamente la información mediante estructuras jerárquicas conocidas como Clustering Feature Trees (CF Tree), lo que permite reducir la complejidad computacional antes de realizar el agrupamiento final.

La reducción mencionada se realiza mediante características de agrupamiento compuestas principalmente por tres elementos: (N), (LS) y (SS), donde (N) corresponde al número de observaciones contenidas en el grupo, (LS) representa la suma lineal de las características de dichas observaciones y (SS) corresponde a la suma cuadrática de sus dimensiones [24]. Gracias a esta estructura, el algoritmo puede resumir grandes cantidades de información preservando propiedades estadísticas relevantes de los datos originales.

En la siguiente imagen se muestran las principales fases del modelo BIRCH:

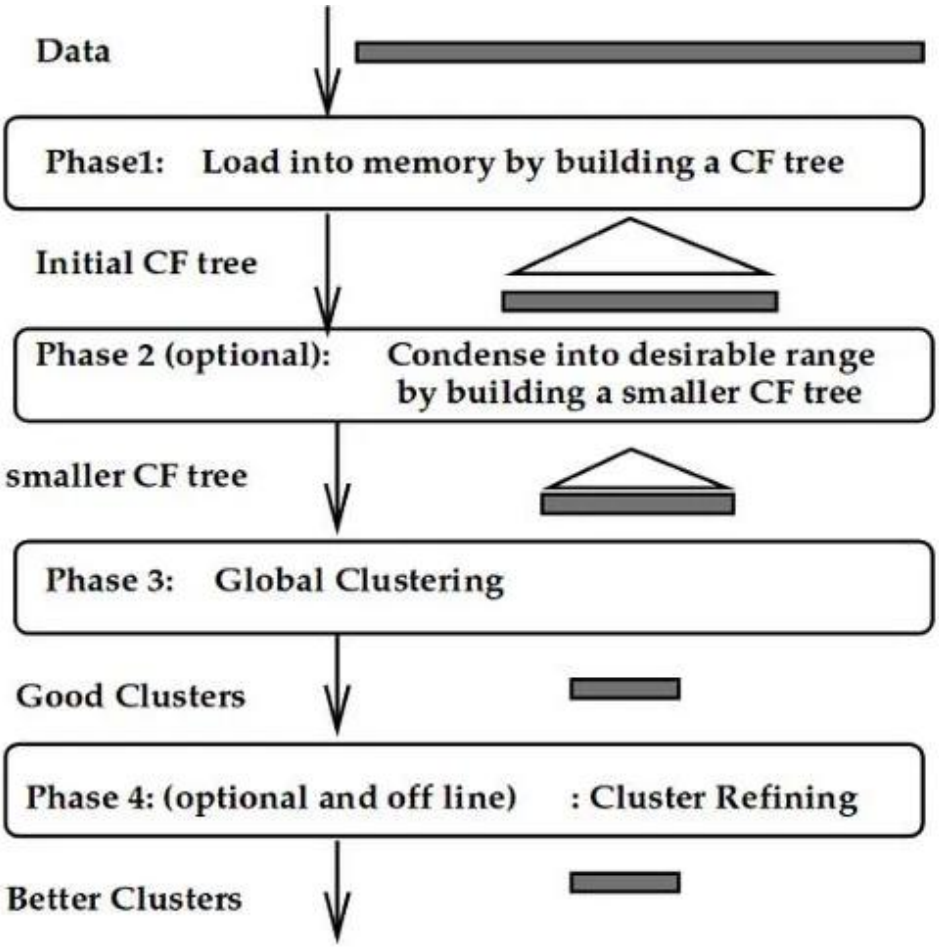


Imagen 5. Fases de Birch clustering
Fuente: [25]

En términos generales, la implementación de modelos de *machine learning* requiere mecanismos de validación orientados a medir su capacidad predictiva, estabilidad y calidad estructural [26], [27]. Debido a que cada tipo de algoritmo posee objetivos distintos, las métricas de evaluación varían dependiendo de si el modelo corresponde a técnicas supervisadas de predicción o a modelos no supervisados de segmentación[28].

En el caso de los modelos supervisados, las métricas se enfocan principalmente en cuantificar la diferencia entre los valores reales y las predicciones generadas por el sistema. Por otro lado, en los modelos de agrupamiento, la evaluación se centra en

analizar la cohesión interna de los clústeres y el nivel de separación existente entre los diferentes grupos identificados. Estas métricas constituyen herramientas fundamentales para determinar el comportamiento técnico de los algoritmos, permitiendo comparar modelos, validar resultados y analizar la capacidad de representación de los patrones presentes en los datos [29], [30], [31].

1.3.1.6. Métricas de evaluación

1.3.1.6.1. MAE

En modelos predictivos, las métricas de error son utilizadas para cuantificar la diferencia existente entre los valores reales y las estimaciones generadas por el algoritmo [27]. Una de las más utilizadas es el Mean Absolute Error (MAE), métrica que calcula el promedio absoluto de las diferencias entre los valores observados y los predichos, permitiendo interpretar el error medio del modelo en las mismas unidades originales de la variable analizada. Su formulación matemática se expresa de la siguiente manera:

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

Imagen 6. Fórmula del Mean Absolute Error
Fuente: [27]

La fórmula del *mean absolute error* tiene *n*, que el número total de observaciones; *i*, que es el índice de observación; y_i, que es el valor real de *i*; $\hat{y}_i$, que es el valor predicho por el modelo en la observación *i*; $|y_i - \hat{y}_i|$, que es el error absoluto de la predicción; la Σ con *n* en el extremo superior y *i*=1 en el extremo inferior, que es la suma de los errores absolutos de todas las observaciones; 1/*n*, que obtiene el promedio de los errores.

Esta métrica se interpreta, teniendo en cuenta que 0 son predicciones perfectas. El número que arroja son unidades en las que se equivoca en promedio, por lo que si su resultado es 5, quiere decir que se equivocó en promedio en 5 unidades. De igual

manera, hay que tener en cuenta que esta métrica conserva las mismas unidades de las unidades objetivo. En este caso, si se estiman casos mensuales de violencia hacia la mujer, el MAE significa que en promedio se desvía o equivoca en dicho valor las estimaciones mensuales.

1.3.1.6.2. RMSE

Otra métrica ampliamente utilizada es el Root Mean Squared Error (RMSE), el cual calcula la raíz cuadrada del promedio de los errores elevados al cuadrado [27]. A diferencia del MAE, esta métrica asigna una penalización mayor a errores de gran magnitud, razón por la cual resulta útil para identificar modelos sensibles a desviaciones extremas. Matemáticamente, el RMSE se representa así:

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

Imagen 7. Fórmula de Root Mean Squared Error
Fuente:[28]

La fórmula de esta métrica contiene un *n*, que el número total de las observaciones; un y_i, que es el valor real de las observaciones *i*; un $\hat{y}_i$, que es el valor predicho por el modelo; la resta y_i - $\hat{y}_i$, que son el error de predicción, el cual se eleva al cuadrado y un 1/*n* Σ, que es el promedio de los errores cuadrados.

Esta métrica, así como el MAE, indica que el valor 0 quiere decir predicciones perfectas y que el valor obtenido son las unidades en que se desvió el modelo. Es decir, si su valor es 7, quiere decir que se desvió en 7 unidades. También representa las mismas unidades de la variable objetivo.

1.3.1.6.3. SMAPE

De manera complementaria, el Symmetric Mean Absolute Percentage Error (SMAPE)

corresponde a una métrica porcentual diseñada para medir la precisión de modelos predictivos, especialmente en contextos de series temporales y forecasting [5]. Su principal característica consiste en normalizar el error respecto al promedio entre el valor real y el valor estimado, disminuyendo problemas asociados a escalas heterogéneas o valores cercanos a cero. La expresión matemática del SMAPE es:

$$SMAPE = \frac{100\%}{n} \sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{(|y_i| + |\hat{y}_i|)/2}$$

Imagen 8. Fórmula de Symmetric Mean Absolute Percentage Error
Fuente: [5]

Esta fórmula tiene un *n*, que es el número total de observaciones; un y_i, que es el valor real de las observaciones *i*; un $\hat{y}_i$, que es el valor predicho por el modelo; un $|y_i - \hat{y}_i|$, que es la diferencia absoluta entre el valor real y el predicho; un $(|y_i - \hat{y}_i,|)/2$, que es el promedio entre el valor predicho y el valor real. El resultado de esta métrica es un porcentaje.

La interpretación de esta métrica es, como en los casos anteriores, que entre menor sea el porcentaje, menor será la desviación de la predicción, siendo 0 % la predicción perfecta. Sin embargo, en este caso, el SMAPE siempre estará entre mayor o igual a 0 y menor o igual a 200%, puesto que se multiplica por 100% la sumatoria de las predicciones y valores reales.

Finalmente, hay que mencionar que si el resultado del SMAPE es 20 %, quiere decir que las predicciones se alejaron en un 20 % de la magnitud media entre los casos reales y los pronosticados.

1.3.1.6.4. Silhoueett Scorr

El Silhouette Score es una métrica usada para analizar la cohesión interna de los grupos y la separación existente entre ellos [30], para medir qué tan similar es una observación

respecto a los elementos de su mismo clúster, en comparación con observaciones pertenecientes a otros grupos. Valores cercanos a 1 representan agrupamientos compactos y bien diferenciados, mientras que valores negativos pueden indicar asignaciones incorrectas. Su representación matemática corresponde a:

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

Imagen 9. Fórmula Silhouette Score
Fuente: [30]

En esta fórmula se parte del $S(i)$, que es el coeficiente de silueta de la observación i; $a(i)$, que es la distancia promedio entre la observación i y las demás observaciones de su mismo clúster; $b(i)$, que es la menor distancia promedio entre la observación i y las observaciones de cualquier otro clúster; y $\max(a(i), b(i))$, que toma el mayor valor entre ambas distancias para normalizar el resultado.

$a(i)$ mide la cohesión interna. Por ello, si su valor es bajo, significa que el punto está más cerca de sus compañeros, por lo que si su valor es alto, el punto está más lejos de sus compañeros. En el caso de $b(i)$, este mide la separación con respecto al clúster vecino más cercano.

Si el valor de $S(i)$ es cercano a 1, la observación está bien asignada a su clúster; si es cercana a 0, quiere decir que la observación está en la frontera entre dos clústeres; si es negativo, quiere decir que probablemente está asignada al clúster equivocado.

1.3.1.6.5. Calinski-Harabasz

Asimismo, el índice de Calinski-Harabasz evalúa la relación entre la dispersión entre clústeres y la dispersión interna de cada agrupamiento [32]. Este índice busca identificar configuraciones donde los grupos presenten alta separación y baja variabilidad interna, características deseables en procesos de *clustering*. Matemáticamente, se expresa de la

siguiente forma:

$$CH = \frac{Tr(B_k)/(k-1)}{Tr(W_k)/(n-k)}$$

Imagen 10. Fórmula de índice Calinski-Harabasz
Fuente:[32].

El índice Calinski-Harabasz contiene un *n*, que es el número total de las observaciones, *k*, que es el número de clústeres; B_k, que es la matriz de dispersión entre clústeres; W_k, que es la matriz de dispersión dentro de los clústeres; *Tr*(), que es la traza de una matriz, o la suma de sus elementos diagonales; y tiene *k*-1 y *n*-*k*, que son los factores de ajuste relacionados con los grados de libertad.

En el caso de $Tr(B_k)$, representa la separación entre los clústeres, lo alejados que están los centroides de cada grupo, con respecto al centroide global del *dataset*. $Tr(W_k)$ representa la dispersión interna de los clústeres, mide qué tan alejadas están las observaciones respecto al centroide del grupo al que pertenecen.

Esta métrica no tiene un límite definido, como –1 a 1, sino que dice que entre mayor sea el número, mejor estructura de agrupamiento hay. Por ello, el número *k* que tenga el número más alto sería la mejor opción. Por ejemplo, si el resultado dice que de 6 clústeres, el número 3 tiene el número más alto, entonces *k*=3 será la mejor opción.

1.3.1.6.6. Davies-Boulin Index

Finalmente, el Davies-Bouldin Index corresponde a una métrica orientada a medir la similitud promedio entre cada clúster y el grupo más cercano [33]. Esta métrica evalúa simultáneamente la compacidad interna y la distancia entre agrupamientos, donde valores bajos indican una mejor calidad de segmentación. Su formulación matemática se representa así:

$$DB = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \left(\frac{S_i + S_j}{M_{ij}} \right)$$

Imagen 11. Fórmula de Davies-Boudin Index
Fuente: [33].

Este índice toma *k*, que es el total de clústeres; *i*, que es el clúster que se está evaluando; *j*, que es otro clúster distinto de *i*; S_i, que es la dispersión interna del clúster *i*; S_j, que es la dispersión interna del clúster *j*; M_{ij}, que es la distancia entre los centroides de los clústeres *i* y *j*; $\max_{j \neq i}$, que selecciona, para cada clúster *i* el clúster *j* con el que tiene la peor separación; i/k Σ calcula el promedio de las comparaciones para cada clúster.

El resultado de esta métrica es un número que indica que entre menor sea el valor, hay un mejor agrupamiento, pues hay poca dispersión entre los clústeres. Si el valor es grande, indica que la separación de estos es grande. Por lo general se interpreta como DB de 0 contiene clústeres compactos y bien separados, mientras que un DB alto quiere decir que son dispersos, cercanos o parcialmente superpuestos. Ahora bien, este indicador no tiene límite superior.

En términos generales, estas métricas constituyen herramientas fundamentales para analizar la calidad y comportamiento de modelos de Machine Learning, permitiendo comparar algoritmos, validar resultados y determinar el nivel de ajuste o separación alcanzado por cada técnica analítica implementada. En este sentido, la evaluación cuantitativa de los modelos no solo permite medir su desempeño técnico, sino también identificar aquellas metodologías con mayor capacidad para representar adecuadamente los patrones y dinámicas presentes en fenómenos sociales complejos.

Bajo este fundamento teórico y metodológico, los algoritmos descritos anteriormente fueron considerados como alternativas analíticas para el desarrollo del presente proyecto, teniendo en cuenta la naturaleza, dimensionalidad y tipología de los datos obtenidos.

1.3.2. Antecedentes

La ciencia de datos se caracteriza por su naturaleza multidisciplinaria, ya que sus métodos y técnicas pueden ser aplicados en diversos campos del conocimiento para apoyar procesos de análisis y toma de decisiones. Aunque gran parte de sus aplicaciones se han desarrollado en contextos empresariales y financieros, en los últimos años su uso en problemáticas sociales ha adquirido mayor relevancia debido al potencial que ofrece para comprender fenómenos complejos y apoyar la formulación de estrategias basadas en evidencia. Dentro de este contexto, la violencia contra la mujer constituye una problemática de alto impacto social que ha comenzado a ser abordada mediante herramientas de *machine learning*, minería de datos y análisis predictivo.

La integración de técnicas de ciencia de datos en el estudio de la violencia basada en género ha permitido transformar registros administrativos, expedientes judiciales, reportes médicos, llamadas de emergencia y publicaciones en redes sociales en fuentes de información útiles para el análisis de patrones de riesgo, identificación de poblaciones vulnerables y detección de zonas geográficas críticas[34], [35]. Estas tecnologías facilitan la extracción de conocimiento a partir de datos estructurados y no estructurados mediante el uso de algoritmos estadísticos, modelos de aprendizaje automático y técnicas de análisis exploratorio[35].

Diversos estudios han implementado modelos supervisados como árboles de decisión, bosques aleatorios (*Random Forest*), regresión logística, XGBoost y redes neuronales para analizar y predecir situaciones asociadas a violencia de género. Por ejemplo, Sáez *et al*. [36] desarrollaron un sistema basado en árboles de decisión para predecir la reincidencia de violencia de pareja en España, utilizando diez variables extraídas de expedientes judiciales. Los autores identificaron que variables asociadas a antecedentes de agresión, contexto familiar y condiciones socioeconómicas permitían mejorar la capacidad predictiva del modelo frente a enfoques estadísticos tradicionales. Estos resultados evidencian el potencial de las técnicas de *machine learning* para fortalecer el

análisis y comprensión de fenómenos asociados a violencia basada en género.

A nivel global, la transición hacia sistemas asistidos por inteligencia artificial ha sido liderada por referentes como el Sistema VioGén en España, que integra información institucional para predecir el riesgo mediante formularios técnicos.[37], [38] Este sistema ha incorporado recientemente herramientas de análisis automatizado basadas en inteligencia artificial orientadas a optimizar la priorización de recursos institucionales y fortalecer los procesos de valoración del riesgo de revictimización de estas formas de violencia.[37], [38] En una revisión sistemática, Wang *et al*.[39] concluyeron que los algoritmos basados en ensemble learning y redes neuronales profundas presentaban mejores resultados en términos de precisión y estabilidad, especialmente en conjuntos de datos con alta complejidad y desbalance de clases[10], [13], [39].

En el contexto latinoamericano, Borrero *et al*. [11] aplicaron técnicas de clasificación supervisada sobre registros de denuncias judiciales, identificando que factores como el desempleo, la dependencia económica y la convivencia prolongada se relacionaban significativamente con la ocurrencia de violencia doméstica en Bogotá. Los resultados evidenciaron el potencial de las técnicas de ciencia de datos para apoyar el análisis de fenómenos sociales complejos en contextos urbanos.

Asimismo, organismos internacionales han promovido el uso de herramientas analíticas para el estudio de la violencia basada en género. El programa UN Global Pulse, [40] por ejemplo, ha desarrollado iniciativas orientadas al análisis de narrativas de riesgo mediante procesamiento de lenguaje natural (PLN) aplicado a redes sociales, permitiendo identificar patrones territoriales y monitorear reportes asociados a posibles eventos de violencia en tiempo real. En este tipo de enfoques, la validación cruzada (cross validation) y métricas de evaluación orientadas a medir precisión, estabilidad y capacidad de generalización han sido ampliamente utilizadas para evaluar el desempeño de los modelos desarrollados, permitiendo reducir problemas asociados al sobreajuste y fortalecer la capacidad de generalización de los algoritmos en diferentes contextos de

análisis.

La investigación de Goin *el al*., llamada Predictor of firearm violence in urban communities, publicada en el 2018, [41] tuvo como objetivo identificar las variables comunitarias que mejor predicen los niveles de violencia con arma de fuego en zonas urbanas del estado de California. Para ello, los autores utilizaron datos provenientes de hospitales, registros oficiales de muertes violentas y censos poblacionales, integrando variables sociodemográficas relacionadas con pobreza, nivel educativo, estado civil y características demográficas. Los modelos implementados fueron LASSO, utilizado para la selección de variables relevantes, y Random Forest, empleado para evaluar la importancia predictiva de los atributos seleccionados. Como resultado, los autores lograron explicar el 77,8 % de la variabilidad en la tasa de violencia, identificando como variables más relevantes los índices de aislamiento, el estado civil, el nivel educativo y la ubicación geográfica.

La investigación de Rodríguez-Rodríguez, *et al*., llamada Modeling and forecasting gender-based violence through machine learning techniques, publicada en 2020,[13] tuvo como objetivo predecir el número de denuncias judiciales por violencia de género en España utilizando registros históricos de más de una década. Para ello, integraron variables demográficas, indicadores de salud pública y factores socioeconómicos asociados a diferentes regiones del país. Los algoritmos evaluados incluyeron Random Forest, regresión logística, k-NN y árboles de decisión. Los resultados mostraron que el modelo Random Forest presentó el mejor desempeño predictivo frente a los demás algoritmos analizados, evidenciando el potencial del machine learning para apoyar procesos de planificación y priorización de políticas públicas a nivel territorial [13].

La investigación de Pinto-Muñoz *et al*., llamada Machine learning applied to gender violence: A systematic mapping study, publicada en 2023,[10] tuvo como propósito realizar un mapeo sistemático de estudios relacionados con la aplicación de *machine learning* en contextos de violencia de género. Para ello, los autores revisaron

publicaciones científicas indexadas en bases de datos especializadas como Scopus y ScienceDirect, identificando los algoritmos más utilizados, las principales líneas de investigación y los contextos territoriales donde estas tecnologías han sido implementadas. Como resultado, encontraron un uso recurrente de algoritmos como *Random Forest*, redes neuronales y regresión logística en investigaciones relacionadas con predicción y análisis de violencia. Asimismo, identificaron que la mayor parte de estos estudios se concentra en países como España, Estados Unidos y México, evidenciando una limitada producción investigativa aplicada específicamente al contexto colombiano.

Finalmente, el artículo de Thomson Lee *et al*., llamado Harnessing the power of machine learning to prevent gender-based violence, publicado en 2025,[42] exploró el uso de *big data* y *machine learning* en la prevención y análisis de la violencia contra las mujeres, destacando tanto su potencial analítico como los desafíos éticos asociados a estas tecnologías. Los autores analizaron el uso de registros médicos, reportes policiales, redes sociales, encuestas de salud y expedientes judiciales para construir modelos predictivos orientados a la identificación de patrones de riesgo. Entre las técnicas utilizadas se encuentran redes neuronales, Random Forest, procesamiento de lenguaje natural y modelos de *forecasting*. Los hallazgos evidenciaron que los algoritmos de *machine learning* pueden mejorar la capacidad predictiva frente a modelos estadísticos tradicionales, además de facilitar la identificación de fenómenos de violencia no reportados o subregistrados. No obstante, también resaltaron riesgos relacionados con privacidad, sesgos algorítmicos y manejo ético de los datos.

La revisión documental realizada evidencia que las técnicas de *machine learning* han sido utilizadas en diferentes contextos para el análisis, clasificación y predicción de fenómenos asociados a la violencia basada en género. Particularmente, algoritmos como Random Forest, árboles de decisión, redes neuronales y modelos de agrupamiento han mostrado resultados relevantes en la identificación de patrones y estimación de comportamientos futuros. Sin embargo, a partir de la literatura consultada, se observa

una limitada cantidad de investigaciones aplicadas específicamente al contexto colombiano y, en particular, a la ciudad de Bogotá. En este sentido, el presente proyecto aplicado busca aportar evidencia analítica mediante la implementación de modelos de segmentación y predicción orientados al estudio de la violencia contra la mujer en la ciudad.

1.3.3. Marco metodológico

Este proyecto aplicado se desarrolla con la metodología CRISP-DM. Este es un estándar usado en los proyectos de ciencia de datos, pues tiene la ventaja de abarcar todo el proceso del proyecto, incluso cuando hay inconvenientes. Para llevar a cabo este proyecto de predicción del número de casos de violencia contra la mujer en Bogotá, se usa esta metodología como se explica a continuación.

En la Imagen 12 se puede ver todo el ciclo del que se habla previamente. Esta ilustración es útil para proyectar los pasos del proyecto:

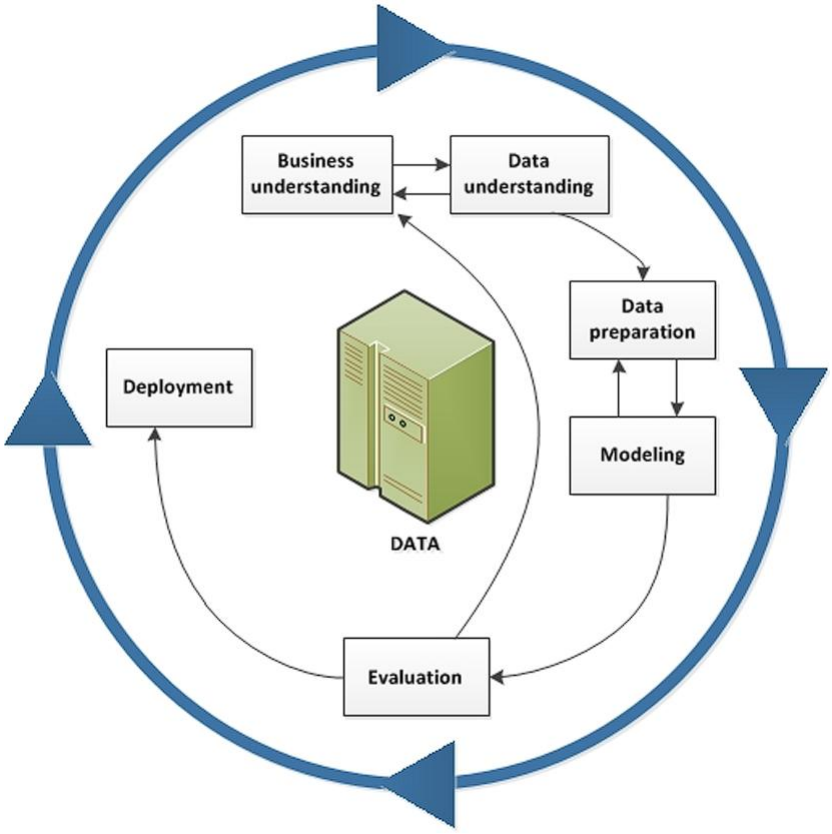


Imagen 12. Ciclo CRISP-DM
Fuente: [43]

Esta metodología, como se puede observar, se compone de seis fases cíclicas que se retroalimentan continuamente: comprensión del negocio, comprensión de los datos, preparación de los datos, modelación, evaluación y despliegue[43], [44], [45], [46]. Existe retroalimentación en estos pasos porque se permite regresar para tomar correctivos, cuando es necesario. Por ejemplo, en el modelado se descubre que hay información que es necesario eliminar (que en la etapa de preparación no se eliminó) porque el modelo no puede funcionar con esos datos o es necesario hacer otro paso. En tal caso, se puede dar un paso atrás y volver a la preparación. También puede ocurrir que en la evaluación se observe que para mejorar la exactitud del modelo sea perentorio volver a la preparación de los datos.

A continuación, se explican las fases de la metodología CRISP-DM.

1.3.3.1. Comprensión del negocio

La fase de Comprensión del Negocio constituye la etapa inicial de la metodología CRISP-DM y tiene como propósito fundamental traducir la problemática social de la violencia contra la mujer en un desafío técnico de analítica avanzada[47]. En el marco de este proyecto aplicado, esta fase se orientó al análisis situacional de la violencia contra la mujer en la ciudad de Bogotá, enfocándose específicamente en los registros de violencia intrafamiliar reportados por el Instituto Nacional de Medicina Legal y Ciencias Forenses.

A nivel social, la violencia contra la mujer es reconocida como un problema de salud pública a nivel global, que afecta a una de cada tres mujeres en el mundo[48]. En el contexto colombiano, se estima que el 33,3% de las mujeres ha sufrido violencia física o sexual a manos de su pareja en algún momento de su vida [49]. "En Bogotá, aunque existen múltiples canales de reporte, persiste un fenómeno de subregistro y fragmentación de la información. Esta situación se deriva de barreras estructurales como el temor a la denuncia y la falta de interoperabilidad entre los sistemas de las entidades, lo que impide que el alto volumen de reportes ciudadanos se traduzca en una respuesta institucional preventiva, coordinada y basada en evidencia[47].

La revisión de reportes oficiales de organismos distritales, como la Secretaría Distrital de Seguridad, Convivencia y Justicia (SDSCJ), permitió identificar la magnitud del fenómeno y una brecha analítica crítica: la actual gestión de datos es predominantemente descriptiva e histórica [50]. En consecuencia, existe una insuficiencia en la toma de decisiones basada en evidencia, ya que la capacidad humana para procesar simultáneamente múltiples variables de riesgo es limitada frente a la velocidad que requiere la prevención activa. Por ello, surgió la necesidad imperante de desarrollar herramientas analíticas que actúen como sistemas de apoyo, permitiendo a las autoridades transitar de la descripción de hechos cumplidos hacia una capacidad de

anticipación territorial.

Asimismo, esta etapa permitió comprender las limitaciones institucionales en términos de análisis preventivo. A partir de esto, el fenómeno fue traducido en objetivos analíticos concretos: la identificación de patrones sociodemográficos y espaciales mediante segmentación, y la predicción del comportamiento futuro del fenómeno en las diferentes localidades de Bogotá

Finalmente, la comprensión del negocio permitió establecer el alcance del proyecto, asegurando que el modelo desarrollado no solo sea técnicamente consistente, sino que se alinee con las metas del Plan Distrital de Desarrollo para salvaguardar la vida de las mujeres[50].

1.3.3.2. Comprensión de los datos

La fase de comprensión de datos se centra en la exploración de las fuentes de información del INMLCF, con el propósito de determinar su estructura, calidad y limitaciones técnicas. En esta etapa, se realiza un análisis exploratorio de datos (EDA) detallado, integrando dimensiones sociodemográficas, temporales y espaciales para caracterizar la violencia intrafamiliar hacia mujeres en Bogotá. Este proceso permite detectar vulnerabilidades críticas, como registros duplicados o valores faltantes, a la vez que se generan visualizaciones estadísticas que facilitan la comprensión del fenómeno. Al tener un carácter descriptivo, esta fase establece el rigor necesario para las etapas subsecuentes de preprocesamiento y modelado analítico.

La obtención de la información se realizó mediante el Portal de Datos Abiertos (datos.gov.co), el repositorio centralizado de transparencia del Gobierno nacional. Se seleccionó un conjunto de datos que integra registros históricos de violencia intrafamiliar, abarcando un periodo longitudinal desde 2015 hasta 2023. Este conjunto permite un análisis exhaustivo de las víctimas a nivel nacional, garantizando la trazabilidad y procedencia oficial de los insumos.

El conjunto de datos sobre violencia intrafamiliar consolida información relativa a las víctimas, la caracterización de los sucesos y el perfil de los presuntos victimarios. Desde una perspectiva técnica, el *dataset* se compone predominantemente de variables categóricas, tanto nominales como ordinales. Las únicas excepciones corresponden a los códigos de la División Político-Administrativa (DIVIPOLA) del DANE para municipios y departamentos; aunque son numéricos, funcionan como identificadores geográficos dentro de la estructura de datos.

A continuación, se presenta la *metadata* correspondiente al *dataset*:

Tabla 1. *Metadata* conjunto de datos de violencia intrafamiliar

Variable	Tipo de variable	Descripción de la variable
ID	Discreta	Identificador único del registro del hecho de violencia.
Año del hecho	Discreta	Año en el que ocurrió el hecho registrado.
Sexo de la víctima	Categórica nominal	Sexo biológico de la víctima reportada.
Grupo de edad de la víctima	Categórica ordinal	Clasificación etaria de la víctima según rangos de edad establecidos.
Grupo Mayor Menor de Edad	Categórica ordinal	Clasificación de la víctima según condición de mayoría o minoría de edad.
Edad judicial	Categórica ordinal	Categoría legal de edad utilizada en procesos judiciales.
Ciclo Vital	Categórica ordinal	Etapa del ciclo de vida de la víctima (infancia, adolescencia, adultez, vejez, etc.).
País de Nacimiento de la Víctima	Categórica nominal	País de origen o nacimiento de la víctima.
Escolaridad	Categórica ordinal	Nivel educativo alcanzado por la víctima.
Estado Civil	Categórica nominal	Estado civil reportado de la víctima.

Tipo de Discapacidad	Categórica nominal	Tipo de discapacidad registrada para la víctima, si aplica.
Pertenencia Étnica	Categórica nominal	Grupo étnico al que pertenece la víctima.
Pertenencia Grupal	Categórica nominal	Grupo poblacional diferencial al que pertenece la víctima.
Mes del hecho	Categórica ordinal	Mes del año en el que ocurrió el hecho.
Día del hecho	Categórica ordinal	Día de la semana en el que ocurrió el hecho.
Rango de Hora del Hecho X 3 Horas	Categórica ordinal	Intervalo horario agrupado cada tres horas para identificar patrones temporales.
Código Dane Municipio	Discreta	Código oficial DANE asociado al municipio del hecho.
Municipio del hecho DANE	Categórica nominal	Municipio donde ocurrió el hecho según clasificación DANE.
Departamento del hecho DANE	Categórica nominal	Departamento donde ocurrió el hecho según clasificación DANE.
Código Dane Departamento	Discreta	Código oficial DANE asociado al departamento.
Localidad del Hecho	Categórica nominal	Localidad de Bogotá donde ocurrió el hecho de violencia.
Zona del Hecho	Categórica nominal	Tipo de zona donde ocurrió el hecho (urbana, rural, etc.).
Escenario del Hecho	Categórica nominal	Lugar o escenario específico donde ocurrió el hecho de violencia.
Actividad Durante el Hecho	Categórica nominal	Actividad que realizaba la víctima al momento del hecho.
Circunstancia del Hecho	Categórica nominal	Contexto o situación en la que ocurrió el hecho de violencia.
Contexto del hecho	Categórica nominal	Tipo de contexto asociado al evento violento.

Mecanismo Causal	Categórica nominal	Medio o mecanismo mediante el cual se ocasionó la lesión o agresión.
Diagnóstico Topográfico de la Lesión	Categórica nominal	Zona anatómica del cuerpo afectada por la lesión.
Sexo del Presunto Agresor	Categórica nominal	Sexo reportado del presunto agresor.
Presunto Agresor	Categórica nominal	Relación o perfil del presunto agresor respecto a la víctima.
Condición de la Víctima	Categórica nominal	Estado o condición particular de la víctima al momento del hecho.
Medio de Desplazamiento o Transporte	Categórica nominal	Medio de transporte utilizado por la víctima o involucrado en el hecho.
Servicio del Vehículo	Categórica nominal	Clasificación del servicio del vehículo involucrado.
Clase o Tipo de Accidente	Categórica nominal	Tipo de accidente asociado al hecho, cuando aplica.
Objeto de Colisión	Categórica nominal	Elemento u objeto involucrado en la colisión o impacto.
Servicio del Objeto de Colisión	Categórica nominal	Tipo de servicio relacionado con el objeto de colisión.
Días de Incapacidad Medicolegal	Categórica ordinal	Rango o clasificación de días de incapacidad medicolegal asignados a la víctima.

Fuente: Elaboración propia.

El *dataset* inicial de violencia intrafamiliar comprende 615.709 registros y 38 variables a nivel nacional entre el periodo 2015 al 2023. Con el propósito de cumplir con el alcance geográfico definido en esta investigación, se realizó un filtro para extraer exclusivamente los datos de Bogotá, obteniendo un subconjunto final de 153.597 registros y 38 columnas. Esta muestra consolidada aporta más de 5.8 millones de datos.

Es preciso señalar que no la totalidad de las 38 variables originales aportan valor analítico al proyecto. Se identificó una alta prevalencia de etiquetas como “no aplica”, “sin información” o “por determinar”, lo que compromete la utilidad de ciertas dimensiones.

Por tanto, estas inconsistencias serán gestionadas en la fase de preparación de datos mediante técnicas de limpieza e imputación, con el fin de garantizar la integridad y calidad del modelo final.

1.3.3.3. Preparación de los datos

La fase de preparación de datos comprende las transformaciones técnicas necesarias para convertir los datos originales en una estructura apta para el aprendizaje automático[51]. Este proceso garantiza que la información pueda ser procesada adecuadamente por los modelos y facilita la identificación de patrones relevantes dentro del fenómeno analizado. En esta etapa se desarrollaron actividades de depuración, recategorización de variables e ingeniería de características (*feature engineering*). Aunque no se identificaron valores nulos en el conjunto de datos, fue necesario implementar protocolos de normalización, transformación e imputación para fortalecer la consistencia y calidad analítica de la información. Como resultado, se obtuvo un *dataset* consolidado y optimizado para la fase de modelado.

Tras aplicar los filtros por ciudad (Bogotá) y sexo de la víctima (Mujer), se consolidó una muestra de 153.597 registros, sometidos a un protocolo de preparación exhaustivo para garantizar su idoneidad en los modelos analíticos. Debido a la naturaleza de los datos, se identificó una alta prevalencia de variables categóricas con elevada cardinalidad, exigiendo un proceso riguroso de depuración y transformación de atributos [52]. Este procedimiento inició con una exploración descriptiva para caracterizar la tipología de las variables y asegurar la integridad ante duplicidades [28]. Posteriormente, se efectuó la normalización textual mediante la estandarización de formatos y corrección de inconsistencias léxicas, mitigando el ruido semántico para evitar que categorías idénticas fueran interpretadas como disímiles por los algoritmos [53]. Finalmente, se seleccionaron las variables clave, fundamentando esta elección en el conocimiento del dominio para robustecer el análisis multivariado de la violencia contra la mujer.

Criterios de selección y exclusión de variables del *dataset* final:

Tabla 2. Criterios de selección y exclusión de variables

Variable	¿Se mantiene en el *dataset* final?	Justificación metodológica
ID	No	Corresponde únicamente a un identificador único del registro y no aporta valor analítico ni predictivo al modelo.
Sexo de la víctima	No	La variable presenta una única categoría ("Mujer"), debido al enfoque del estudio sobre víctimas femeninas, por lo que no aporta variabilidad al análisis.
Grupo de edad de la víctima	No	Variable redundante respecto a "Edad judicial", por lo que se conserva únicamente la variable con mayor utilidad analítica y consistencia categórica.
Grupo Mayor Menor de Edad	No	Variable redundante derivada de la clasificación etaria ya contenida en "Edad judicial".
Edad judicial	Sí	Se conserva debido a su relevancia para identificar patrones asociados a rangos etarios y dinámicas de violencia según grupos de edad.
Ciclo Vital	Sí	Permite analizar comportamientos diferenciales según etapas del ciclo de vida de las víctimas.
País de Nacimiento de la Víctima	Sí	Variable útil para identificar posibles patrones asociados a población migrante o extranjera.
Escolaridad	Sí	Aporta información sociodemográfica relevante para el análisis de vulnerabilidad y segmentación de casos.
Estado Civil	Sí	Permite identificar relaciones entre vínculos afectivos y tipos de violencia registrados.
Tipo de Discapacidad	Sí	Variable relevante para analizar posibles condiciones de vulnerabilidad diferencial.
Pertenencia Étnica	No	El 94% de los registros presentan categorías sin información o sin

		pertenencia étnica, limitando significativamente su capacidad analítica.
Pertenencia Grupal	Sí	Permite identificar condiciones poblacionales diferenciales relevantes para la caracterización de víctimas.
Mes del hecho	Sí	Variable temporal relevante para identificar patrones estacionales y tendencias de ocurrencia.
Día del hecho	Sí	Permite analizar comportamientos asociados a la ocurrencia de hechos según el día de la semana.
Rango de Hora del Hecho X 3 Horas	Sí	Facilita la identificación de patrones horarios asociados a los eventos de violencia.
Municipio del hecho DANE	No	La investigación se restringe a Bogotá, por lo que la variable pierde variabilidad y capacidad discriminativa.
Código Dane Municipio	No	Variable eliminada debido a que todos los registros corresponden a Bogotá, generando una única categoría efectiva.
Departamento del hecho DANE	No	Se elimina por falta de variabilidad al trabajar exclusivamente con registros de Bogotá.
Código Dane Departamento	No	Variable redundante y sin aporte analítico debido al filtrado geográfico aplicado.
Localidad del Hecho	Sí	Variable espacial fundamental para el análisis territorial y la predicción de casos por zona geográfica.
Zona del Hecho	Sí	Permite diferenciar dinámicas de violencia entre zonas urbanas y rurales.
Escenario del Hecho	Sí	Aporta contexto espacial y situacional sobre los lugares donde ocurren los hechos.
Actividad Durante el Hecho	Sí	Permite analizar las circunstancias y actividades asociadas a los eventos registrados.
Circunstancia del Hecho	Sí	Variable clave para caracterizar las dinámicas de violencia y relaciones entre víctima y agresor.
Contexto del Hecho	No	Complementa la caracterización contextual y tipológica de los casos

		analizados, sin embargo, la caracterización es muy similar a circunstancia del hecho.
Mecanismo Causal	Sí	Variable relevante para identificar patrones de modus operandi y severidad de los hechos.
Diagnóstico Topográfico de la Lesión	Sí	Permite analizar patrones relacionados con el tipo y localización de las lesiones.
Sexo del Presunto Agresor	Sí	Variable relevante para caracterizar perfiles asociados a los agresores.
Presunto Agresor	Sí	Permite identificar relaciones recurrentes entre víctima y agresor dentro de los casos analizados.
Condición de la Víctima	No	Presenta una única categoría (“No aplica”), por lo que no aporta información útil para el análisis.
Medio de Desplazamiento o Transporte	No	Presenta una única categoría (“No aplica”), por lo que no aporta información útil para el análisis.
Servicio del Vehículo	No	Presenta una única categoría (“No aplica”), por lo que no aporta información útil para el análisis.
Clase o Tipo de Accidente	No	Presenta una única categoría (“No aplica”), por lo que no aporta información útil para el análisis.
Objeto de Colisión	No	Presenta una única categoría (“No aplica”), por lo que no aporta información útil para el análisis.
Servicio del Objeto de Colisión	No	Presenta una única categoría (“No aplica”), por lo que no aporta información útil para el análisis.
Días de Incapacidad Medicolegal	No	La variable presenta una alta concentración en pocas categorías (2 categorías con un 80%) y baja capacidad discriminativa para el análisis predictivo.

Fuente: elaboración propia.

En esta etapa, se ejecutó la reconstrucción de categorías no informativas mediante un proceso de imputación de datos, transformando etiquetas ambiguas como "Sin información" o "Por determinar" en registros con valor analítico. Para variables críticas como el mecanismo causal, el contexto de violencia y el diagnóstico de la lesión, se aplicó la técnica de imputación por moda, reemplazando los valores faltantes por la

categoría con mayor frecuencia observada en cada variable.[54] Esta estrategia permitió mejorar la completitud de la base de datos y garantizar una mayor consistencia en el análisis posterior, minimizando la pérdida de información y facilitando la aplicación de los algoritmos de segmentación y predicción desarrollados en el estudio.[55]

Una vez finalizada esta fase, para dar continuidad al procesamiento de datos para la entrada de los modelos de segmentación, se realizó la optimización de categorías a través de una reducción de cardinalidad en atributos con excesiva dispersión, tales como el escenario del hecho y el perfil del agresor. Mediante la aplicación del principio de Pareto, se priorizó la representatividad estadística al conservar solo las categorías que acumulaban el 90% de la frecuencia total, consolidando el volumen residual bajo la etiqueta "Otros".[56] Esta simplificación permitió limpiar los datos de variaciones irrelevantes que dificultaban la identificación de patrones claros. Al agrupar las categorías poco frecuentes, se evitó que los algoritmos crearan grupos minúsculos y sin valor práctico, los cuales suelen ser producto del azar y no de una tendencia real del fenómeno.[57]

Dando continuidad al procesamiento, se ejecutó la transformación del espacio dimensional mediante el Análisis de Correspondencias Múltiples (MCA). Esta acción permitió convertir los atributos cualitativos en dimensiones numéricas que preservan la estructura original de los datos, evitando las distorsiones que introduciría un PCA tradicional en variables nominales.[58] El resultado fue una representación compacta y metodológicamente óptima, diseñada para potenciar la identificación de patrones en los modelos de segmentación no supervisada.[59]

En cuanto al componente predictivo, se consolidó una estructura longitudinal mediante la agregación temporal y geográfica de los registros. Este proceso permitió la integración de la dependencia histórica del fenómeno a través de variables rezagadas (lag features), como casos_total_lag_1, dotando al análisis de un componente temporal crítico.[5] Finalmente, se definieron los marcos de entrenamiento para algoritmos de ensamble

como Random Forest y XGBoost, asegurando la compatibilidad técnica de las variables. Esta preparación garantizó un conjunto de datos robusto, capaz de capturar tanto la variabilidad espacial como la evolución futura de la violencia en la ciudad.[60]

Como resultado integral de este proceso, se consolidó un conjunto de datos depurado y técnicamente optimizado, el cual constituye el insumo fundamental para las etapas de aprendizaje automático. El refinamiento de los registros y la creación de estructuras específicas, tanto dimensionales para la segmentación como longitudinales para la predicción, garantizan que los datos posean la estabilidad y el valor explicativo requeridos para el estudio de la violencia contra la mujer en Bogotá. Con esta base sólida y coherente, se da paso a la fase de Modelado, donde se describe la implementación de los algoritmos encargados de extraer los patrones, agrupamientos y tendencias futuras del fenómeno.

1.3.3.4. Modelado

La fase de modelado tuvo como propósito desarrollar técnicas de aprendizaje automático capaces de predecir el comportamiento futuro de los casos de violencia contra la mujer en Bogotá, así como identificar patrones sociodemográficos, espaciales y de ocurrencia asociados al fenómeno. Para ello, se implementaron modelos de aprendizaje supervisado y no supervisado, permitiendo abordar tanto la predicción de casos como la segmentación de perfiles dentro de la población analizada.

Ahora bien, haciendo un paréntesis en este punto, es necesario advertir que, aunque los datos tienen registros con años, meses y días de la semana en que se reportaron los hechos, en este ejercicio no se consideraron modelos de ML de series de tiempo o que tienen un desempeño considerable en esos casos, como las redes neuronales. Un criterio para esto es que se cuenta con un poco más de 112.000 registros, lo que puede representar un limitante importante para la cantidad de datos que se requieren con estos modelos. Además, los datos obtenidos son categóricos en su mayoría y no numéricos

Esto se hizo así porque el proyecto está enfocado al desarrollo de una herramienta que permita estimar la cantidad de casos de violencia hacia la mujer en Bogotá como mecanismo de ayuda a la hora de mitigar problemáticas o de dirigir la acción estatal con dicho fin. Por consiguiente, se enfocó en modelos que permitan lograr este fin, considerando los límites de un proyecto aplicado.

Por ende, en el componente predictivo, se implementaron modelos basados en técnicas de ensamble, específicamente *Random Forest* y XGBoost, debido a su capacidad para modelar relaciones no lineales, manejar estructuras complejas de datos y capturar interacciones entre variables temporales y geográficas. El entrenamiento de los modelos se realizó utilizando estructuras longitudinales construidas a partir de agregaciones temporales y territoriales, incorporando variables históricas y rezagadas (*lag features*) para representar la evolución temporal del fenómeno.

La evaluación del desempeño predictivo se realizó mediante métricas de error como MAE (*Mean Absolute Error*), RMSE (*Root Mean Squared Error*) y SMAPE (*Symmetric Mean Absolute Percentage Error*), considerando horizontes de predicción de uno, dos y tres periodos futuros. A partir de esta comparación, el modelo Random Forest presentó un comportamiento más estable y consistente frente a XGBoost, razón por la cual fue seleccionado como el modelo predictivo principal de la investigación.

Adicionalmente, se implementaron modelos de segmentación orientados a identificar agrupamientos homogéneos dentro de la población de mujeres víctimas. Para ello, se evaluaron algoritmos de *clustering* como K-Means y BIRCH (*Balanced Iterative Reducing and Clustering using Hierarchies*), seleccionados por su capacidad para detectar estructuras y patrones en conjuntos de datos con múltiples atributos categóricos transformados.

La evaluación de los modelos de segmentación se realizó mediante métricas de validación interna como Silhouette Score, Davies-Bouldin Index y Calinski-Harabasz

Index, permitiendo analizar la cohesión y separación entre los grupos generados. Como resultado de este proceso, el modelo BIRCH presentó una mejor capacidad para identificar agrupamientos representativos y consistentes dentro de la población analizada, siendo seleccionado como el modelo de segmentación del estudio.

Finalmente, los modelos seleccionados constituyen la base analítica de la investigación y permiten tanto estimar la cantidad de casos futuros de violencia contra la mujer en Bogotá como identificar patrones relevantes asociados al comportamiento del fenómeno en las diferentes localidades de la ciudad.

1.3.3.5. Evaluación

La fase de evaluación tuvo como propósito analizar el desempeño y la capacidad analítica de los modelos implementados, verificando su utilidad para identificar patrones y predecir el comportamiento futuro de los casos de violencia contra la mujer en Bogotá. Esta etapa permitió validar tanto la estabilidad de los agrupamientos generados en los modelos de segmentación como la precisión de las predicciones obtenidas en los modelos predictivos.

Para el componente de segmentación, la evaluación se fundamentó en métricas de validación interna como Silhouette Score, Davies-Bouldin Index y Calinski-Harabasz Index, dada la naturaleza no supervisada del análisis y la ausencia de una variable objetivo. Este rigor estadístico se integró con una valoración cualitativa de la interpretabilidad de los grupos, asegurando su coherencia con el comportamiento histórico del fenómeno de la violencia contra la mujer. Bajo este enfoque, el modelo BIRCH demostró un desempeño superior frente a K-Means en términos de separación y consistencia, permitiendo consolidar estructuras analíticas más representativas para la caracterización de los perfiles sociodemográficos, espaciales y de modus operandi en la ciudad.

En lo que respecta al componente predictivo, la evaluación se efectuó mediante métricas

de error de regresión, específicamente MAE (Mean Absolute Error), RMSE (Root Mean Squared Error) y SMAPE (Symmetric Mean Absolute Percentage Error), las cuales permitieron cuantificar las discrepancias entre los valores reales y las proyecciones realizadas en diferentes horizontes temporales. La selección de estas métricas obedeció a la naturaleza del problema abordado, dado que el objetivo consistió en predecir una variable numérica continua correspondiente a la cantidad de casos de violencia contra la mujer por localidad. En consecuencia, no se utilizaron métricas propias de problemas de clasificación, tales como Accuracy, Sensibilidad (Recall), Especificidad o Precisión, debido a que estas se emplean cuando la variable objetivo corresponde a categorías discretas y no a valores cuantitativos. A través del análisis realizado, se determinó que el modelo Random Forest exhibió un comportamiento más estable y consistente en comparación con XGBoost, demostrando una mayor capacidad para estimar el número de casos futuros en las localidades de Bogotá. Este desempeño se vio favorecido por la incorporación de variables temporales y rezagadas (*lag features*), las cuales fortalecieron la capacidad de generalización de los algoritmos al capturar con mayor fidelidad la evolución histórica, las tendencias y la dinámica temporal del fenómeno analizado.

Finalmente, esta fase permitió validar la idoneidad de los modelos seleccionados para la segmentación y predicción de la violencia contra la mujer. Al garantizar que los resultados obtenidos fueran técnicamente consistentes y estuvieran alineados con los objetivos de la investigación, se consolidó el marco de confianza necesario para transitar hacia la etapa de Despliegue, orientada a la integración y utilidad práctica de estos modelos como herramientas de apoyo para la toma de decisiones en la ciudad.

1.3.3.6. Despliegue

La fase de despliegue corresponde a la etapa en la cual los resultados obtenidos a través de los modelos analíticos son presentados de manera comprensible y útil para los potenciales usuarios de la investigación.[46] En el contexto del presente proyecto aplicado, esta fase estuvo orientada a transformar los hallazgos derivados de los

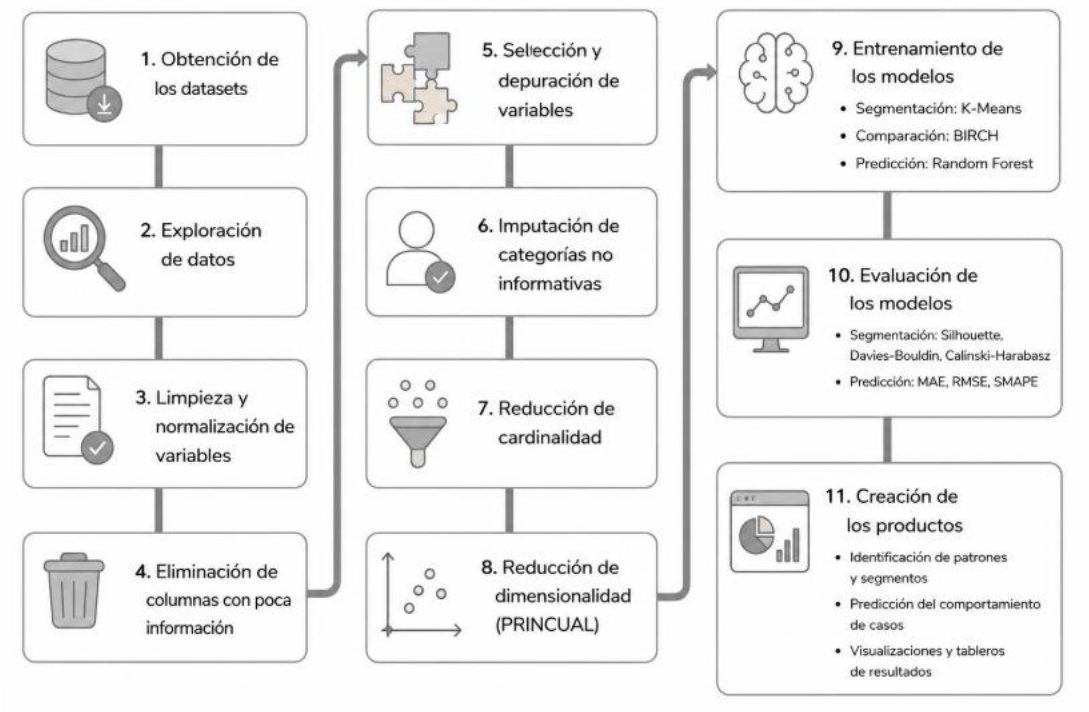
modelos de segmentación y predicción en herramientas de apoyo para el análisis y la toma de decisiones relacionadas con la violencia contra la mujer en Bogotá.

El despliegue de los resultados se materializó mediante la construcción de visualizaciones y *dashboards* interactivos, diseñados para facilitar la exploración de la información por localidad, comportamiento temporal y características sociodemográficas asociadas a los casos analizados. Estas herramientas permiten representar de forma gráfica los patrones identificados, los territorios con mayor concentración de casos y las tendencias estimadas por los modelos predictivos.

Si bien el alcance del proyecto no contempla la implementación de un sistema operativo en producción, los resultados obtenidos constituyen un insumo analítico con potencial de integración en entornos institucionales orientados a la prevención, monitoreo y atención de la violencia contra las mujeres. Asimismo, los modelos desarrollados y las estructuras analíticas construidas pueden servir como base para futuras estrategias de vigilancia, priorización territorial y formulación de políticas públicas sustentadas en evidencia.

Una vez descritas las fases de la metodología CRISP-DM implementadas en el presente proyecto aplicado, se procede a exponer el desarrollo técnico y analítico de la investigación, así como los resultados obtenidos durante la implementación de los modelos de *machine learning*.

El proceso metodológico desarrollado puede resumirse en el siguiente esquema:



Infografía 1. Proceso de elaboración de los modelos
Fuente: elaboración propia.

Capítulo II.

Contextualización de la violencia intrafamiliar y homicidio en Bogotá

En Colombia hay distintas entidades gubernamentales que compilan datos sobre hechos de violencias. La Policía Nacional tiene el SIEDCO –Sistema de Información Estadístico, Delincuencial, Contravencional y Operativo de la Policía Nacional de Colombia– y recopila información sobre delitos, convenciones y demás temas que tienen que ver con esta entidad y que son útiles para la toma de decisiones en esta materia. Otra fuente de datos es Medicina Legal, que tiene un observatorio en el que usa el sistema Forensis, en donde publica información periódicamente de los casos de violencias mortales y no mortales a nivel nacional. Dentro de este sistema de Medicina Legal están los archivos publicados en Datos Abiertos, que es un portal web que tiene el Gobierno colombiano para aumentar la transparencia en sus instituciones a nivel nacional.

Cada año, el INMLCF publica en Datos Abiertos un conjunto de datos con información de las violencias, el cual es tomado en este proyecto aplicado. Para este ejercicio, se tomó el periodo 2019-2023.

En este apartado, se muestran los resultados del análisis exploratorio de los conjuntos de datos mencionado. El primer lugar, se explica el *dataset* para dar contexto. Posteriormente, se explica a detalle el proceso de filtrado, imputación y creación del conjunto de datos final usado para los modelos de ML.

El conjunto de datos contiene datos de forma estructuradas, llenado en forma de formulario, lo que implica que cada columna contiene solo datos tipificados como categóricos, incluso en los que corresponden a edades. Este conjunto de datos tiene un peso de 326 megabytes.

2.1. Exploración de los datos

El conjunto de datos es producto de formularios, lo que significa que la información es

estándar y no varía en los años contenidos en el *dataset*, por lo que la estructura de los datos es similar. Esto facilita el proceso del análisis exploratorio y los siguientes pasos.

Este conjunto corresponde al periodo 2015-2023 y tiene 37 columnas para intrafamiliar, y 35, para homicidios. El contenido de este es nacional durante el periodo y abarca víctimas tanto hombres como mujeres; por esta razón, se filtran por mujer y por Bogotá en este ejercicio. Los primeros resultados de este ejercicio son los siguientes: Los datos ocurridos en Bogotá fueron 138.038 casos de violencia intrafamiliar. Por otra parte, los casos que se registraron como víctimas mujeres correspondieron a 112.688 de violencia intrafamiliar.

A continuación, en las tablas 4 y 5 se muestran las variables de cada conjunto de datos y cuántos valores sin información tenían.

Tabla 3. Columnas del conjunto de violencia intrafamiliar

Columna	Tipo de variable	Sin información y nulos
ID	Discreta	0 nulos 0 sin información
Año del hecho	Discreta	0 nulos 0 sin información
Sexo de la víctima	Categórica nominal	0 nulos 0 sin información
Grupo de edad de la víctima	Categórica ordinal	0 nulos 0 sin información
Grupo Mayor Menor de Edad	Categórica ordinal	0 nulos 0 sin información
Edad judicial	Categórica ordinal	0 nulos 0 sin información
Ciclo Vital	Categórica ordinal	0 nulos 0 sin información
País de Nacimiento de la Víctima	Categórica nominal	0 nulos 134 sin información
Escolaridad	Categórica ordinal	0 nulos 785 sin información
Estado Civil	Categórica nominal	0 nulos 455 sin información
Tipo de Discapacidad	Categórica nominal	0 nulos 0 sin información

Pertenencia Étnica	Categórica nominal	0 nulos 17322 sin información
Pertenencia Grupal	Categórica nominal	0 nulos 13960 sin información
Mes del hecho	Categórica ordinal	0 nulos 0 sin información
Dia del hecho	Categórica ordinal	0 nulos 0 sin información
Rango de Hora del Hecho X 3 Horas	Categórica ordinal	0 nulos 10883 sin información
Código Dane Municipio	Discreta	0 nulos 0 sin información
Municipio del hecho DANE	Categórica nominal	0 nulos 0 sin información
Departamento del hecho DANE	Categórica nominal	0 nulos 0 sin información
Código Dane Departamento	Discreta	0 nulos 0 sin información
Localidad del Hecho	Categórica nominal	0 nulos 19019 sin información
Zona del Hecho	Categórica nominal	0 nulos 3090 sin información
Escenario del Hecho	Categórica nominal	0 nulos 566 sin información
Actividad Durante el Hecho	Categórica nominal	0 nulos 240 sin información
Circunstancia del Hecho	Categórica nominal	0 nulos 0 sin información
Contexto de violencia	Categórica nominal	0 nulos 0 sin información
Mecanismo Causal	Categórica nominal	0 nulos 0 sin información
Diagnostico Topográfico de la Lesión	Categórica nominal	0 nulos 10526 sin información
Sexo del Presunto Agresor	Categórica nominal	0 nulos 359 sin información
Presunto Agresor	Categórica	0 nulos 254 sin información
Condición de la Víctima	Categórica nominal	0 nulos 0 sin información
Medio de Desplazamiento o Transporte	Categórica nominal	0 nulos 0 sin información
Servicio del Vehículo	Categórica nominal	0 nulos 0 sin información
Clase o Tipo de Accidente	Categórica nominal	0 nulos 0 sin información

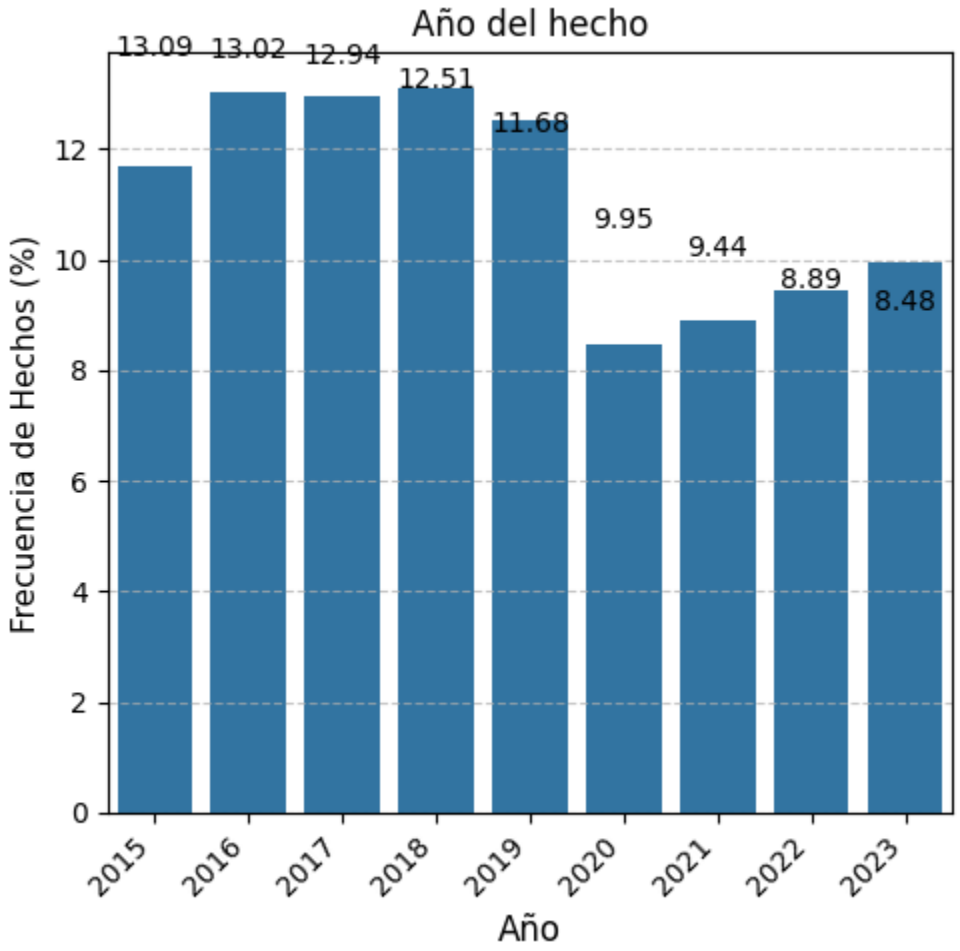
Objeto de Colisión	Categórica nominal	0 nulos 0 sin información
Servicio del Objeto de Colisión	Categórica nominal	0 nulos 0 sin información
Días de Incapacidad Medicolegal	Categórica ordinal	0 nulos 4534 sin información

Fuente: elaboración propia.

El conjunto de datos contiene variables redundantes. Por ejemplo, las relacionadas con la edad: Grupo de edad de la víctima, Grupo Mayor Menor de Edad, Edad judicial y Ciclo vital. Estas variables contienen un categórico (pues, aunque es un rango, lo reconoce como un dato categórico), que es un grupo de años cumplidos y hay otra que indica si son mayores o menores de edad (Grupo Mayor Menor de Edad). En ese caso, la columna que interesa es Grupo de edad de la víctima por contener un grupo de edad que va hasta los 17 años y separar entre mayores y menores de edad.

El análisis exploratorio se hizo teniendo dos propósitos: comprender los conjuntos de datos y organizar los conjuntos de datos para facilitar los pasos de creación de los modelos de ML. Esta exploración se hizo en un jupyter notebook en lenguaje Python, con las librerías Pandas, Numpy, Matplotlib y Seaborn.

Gráfico 1. Año del hecho



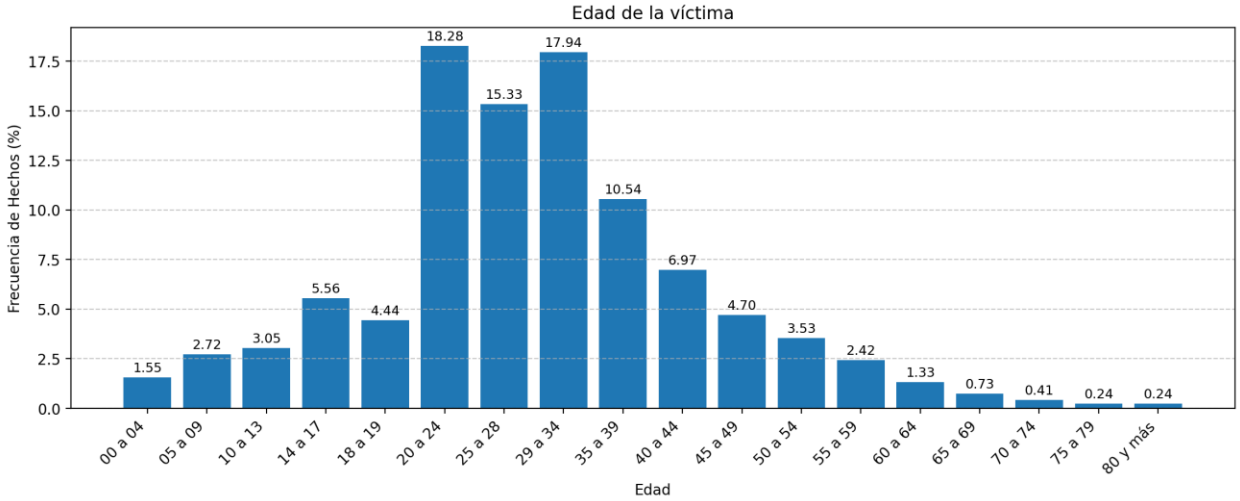
Fuente: elaboración propia.

El gráfico 1 muestra la distribución porcentual de los hechos registrados según el año del hecho entre 2015 y 2023. Se observa que los mayores porcentajes se concentran entre 2016 y 2017, con valores cercanos al 13%: 2016 presenta el porcentaje más alto con 13,02 %, seguido de 2017 (12,94 %), 2018 (12,51 %) y 2019 (11,68 %). En contraste, a partir de 2020 se evidencia una disminución importante en la proporción de registros, especialmente en 2020, con 9,95 %, y en los años posteriores, que se mantienen por debajo del 10%.

Esta disminución en el año 2020 y posterior a este puede ser efecto del covid-19 que vivió el país o algún otro cambio que pudo afectar la cantidad de casos reportados en

esos años. Sin embargo, es notable que año tras año se presentan más datos, sin que se presenten cantidades similares niveles previos al 2020.

Gráfico 2. Edad de las víctimas



Fuente: elaboración propia.

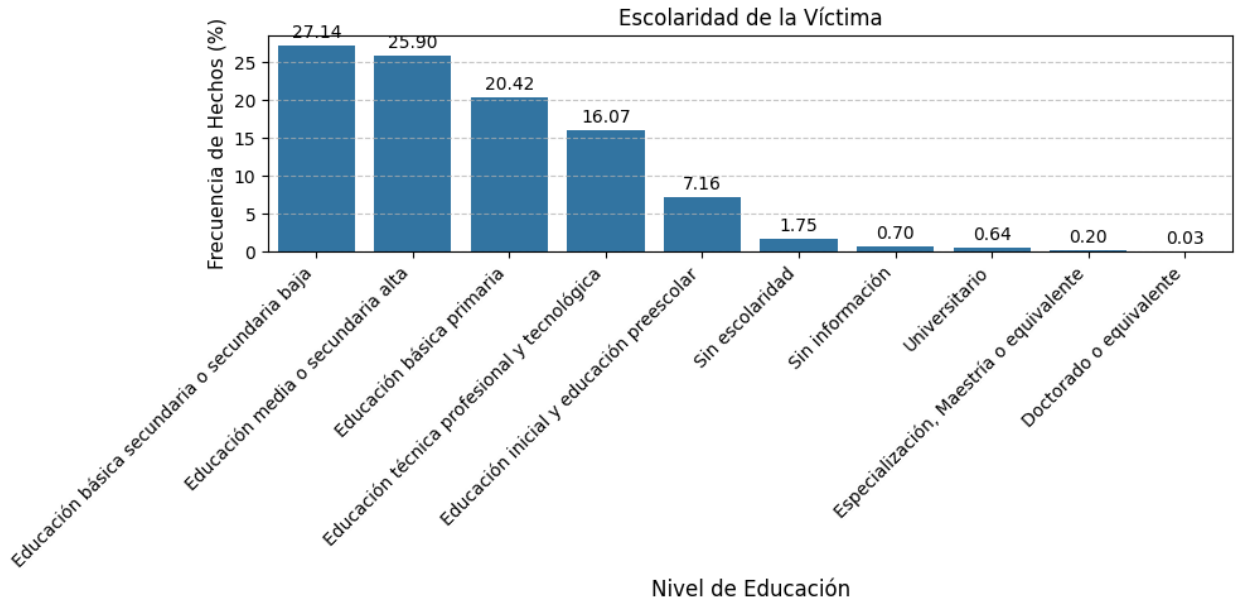
La distribución de los hechos registrados según la edad de la víctima evidencia una concentración en mujeres jóvenes y adultas jóvenes. El grupo de 20 a 24 años presenta la mayor proporción de registros, con 18,28 %, seguido por los rangos de 25 a 28 años (15,33 %), 29 a 34 (17,94 %) y 35 a 39 años (10,54 %). En conjunto, las víctimas de 20 a 39 años representan aproximadamente el 62,09 % de los hechos registrados, lo cual indica que la mayor parte de los casos reportados se concentra en este tramo de edad. A partir de los cuarenta años, la proporción de registros disminuye progresivamente: el rango de 40 a 44 años representa el 6,97 %, mientras que en los grupos de 60 años o más los porcentajes son inferiores al 1,5 %. Asimismo, se registran hechos en niñas y adolescentes, aunque con menores proporciones que en la población adulta joven.

Gráfico 3. País de nacimiento de la víctima

Fuente: elaboración propia.

El Gráfico 3 evidencia que la gran mayoría de los hechos registrados corresponde a víctimas nacidas en Colombia, categoría que concentra el 97,97 % de los casos, lo que sugiere que se muestra una problemática que afecta mayoritariamente a las connacionales. En segundo lugar se encuentra Venezuela, con aproximadamente 1,81 %, mientras que las demás categorías (incluyendo “Sin información” y otros países de nacimiento) presentan participaciones muy reducidas, cercanas o inferiores al 0,1%. No obstante, es necesario mencionar que de presentarse realmente mayores casos donde las mujeres venezolanas son víctimas de estas formas de violencia, y que se cuenten con denuncias, puede haber cierto nivel de subregistro por temas de documentación legal para permanecer en el país. Este subregistro no es tratado en el presente proyecto.

Gráfico 4. Escolaridad de las víctimas

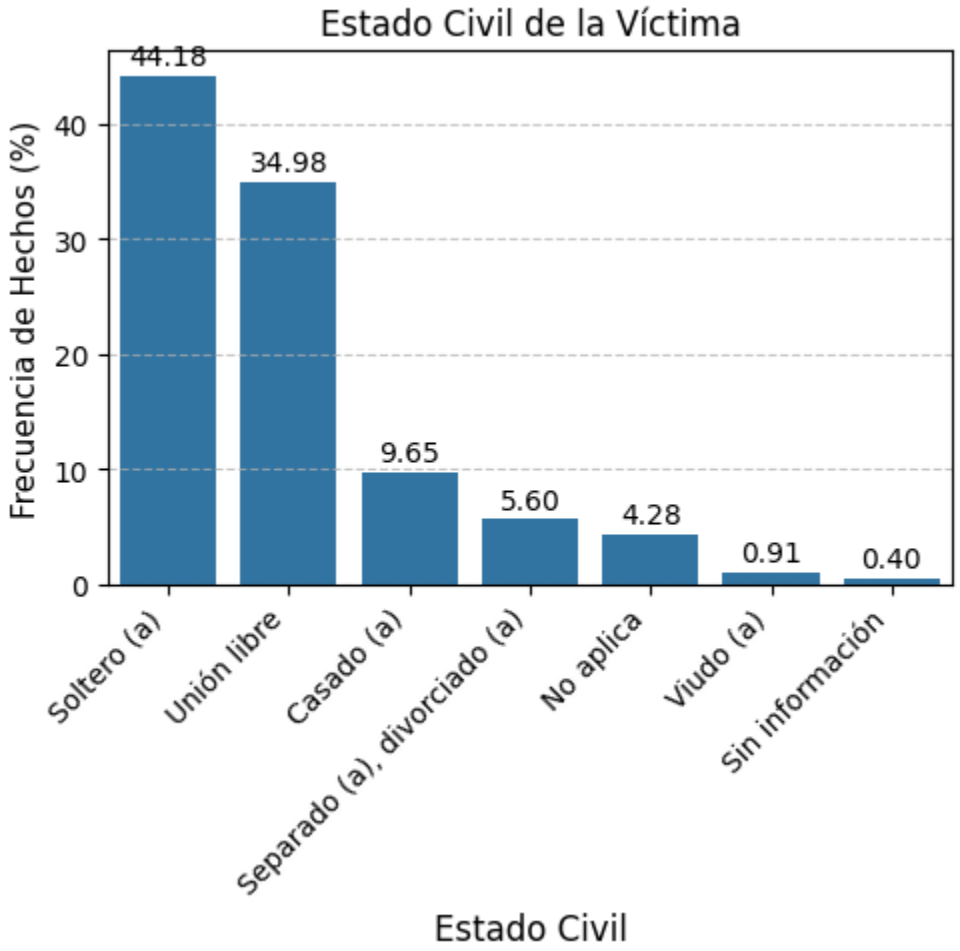


Fuente: elaboración propia.

La Gráfica 4 muestra el nivel de educación registradas de las víctimas de VIF. Las categorías son desde “Sin escolaridad” hasta “doctorado o equivalente”, abarcando todas las etapas del modelo educativo pre y posgradual.

Los resultados están ordenados de mayor a menor peso porcentual. Con esto se puede ver que las personas con educación baja –primaria, secundaria y media– son quienes representaron más del 70 % de los casos. Esto sugiere que entre mayor es la formación culminada es menor la cantidad de casos reportados.

Gráfico 5. Estado civil de la víctima

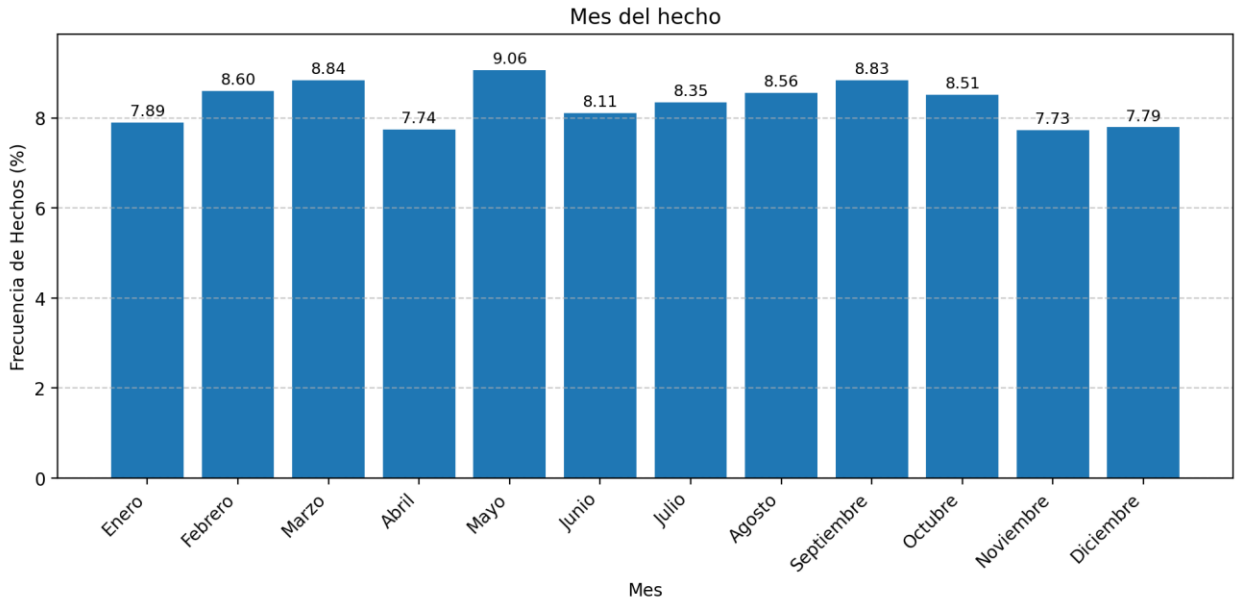


Fuente: elaboración propia.

La Gráfica 5 muestra el estado civil de las mujeres víctimas. En este, mujeres solteras y en unión libre son las que representan la mayor cantidad de los casos reportados, con el 79 % del total. En este sentido, las personas que son viudas o están separadas o divorciadas son menos del 7 % de los casos reportados.

Esta variable no necesariamente representa un factor determinante en general para sufrir de violencia, puesto que la violencia intrafamiliar también hacia niñas y adolescentes, para quienes no aplican esas categorías; otro ejemplo es que la violencia puede ser propiciada por alguien de la misma familia, aspecto que no guarda relación con el estado civil de la víctima.

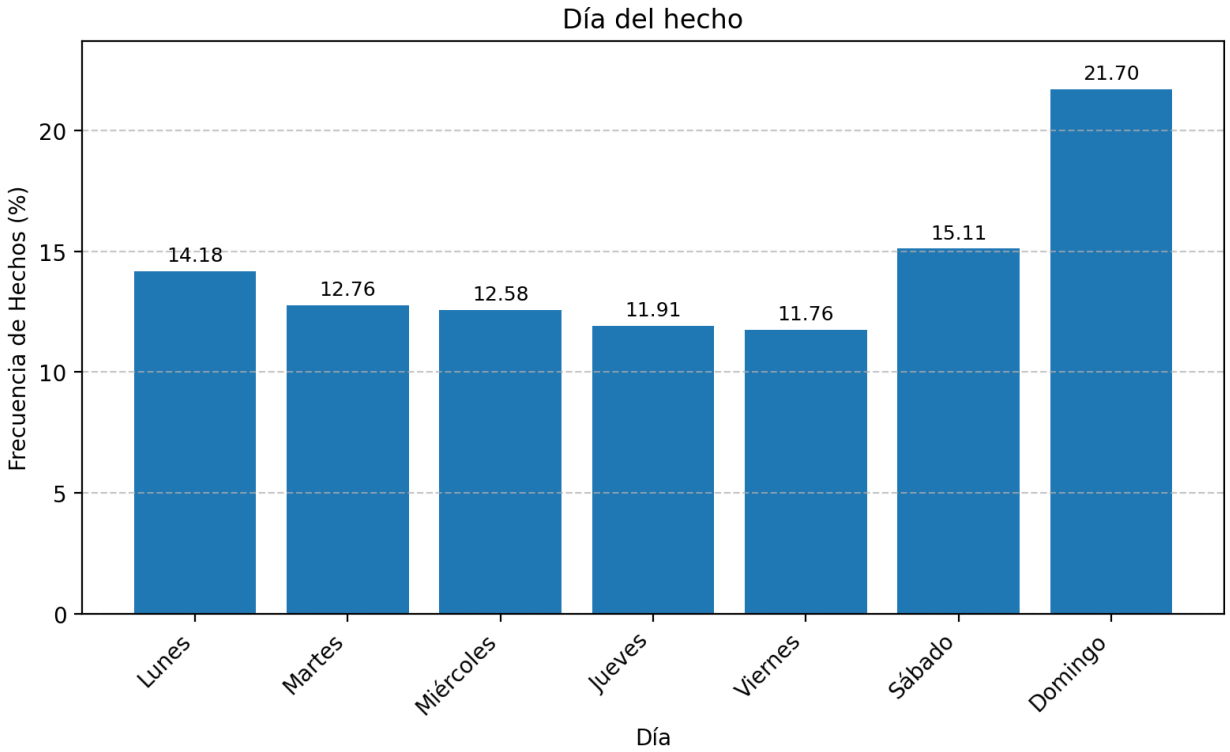
Gráfico 6. Mes del hecho



Fuente: elaboración propia.

El histograma revela que la violencia contra la mujer en Bogotá es un fenómeno estructural y persistente, con una distribución anual relativamente homogénea. No obstante, destaca un pico crítico en mayo (9,06%), atribuido a la celebración del Día de la Madre, considerada históricamente la fecha más violenta del año en Colombia debido a factores como el consumo excesivo de alcohol y la intolerancia [61]. A este periodo le siguen marzo y septiembre (8,84 y 8,83 %, respectivamente), mientras que las menores frecuencias se registran en noviembre, abril y diciembre. Esta estabilidad temporal del dato confirma que la violencia no es estacional y valida la implementación de analítica predictiva para la anticipación territorial.

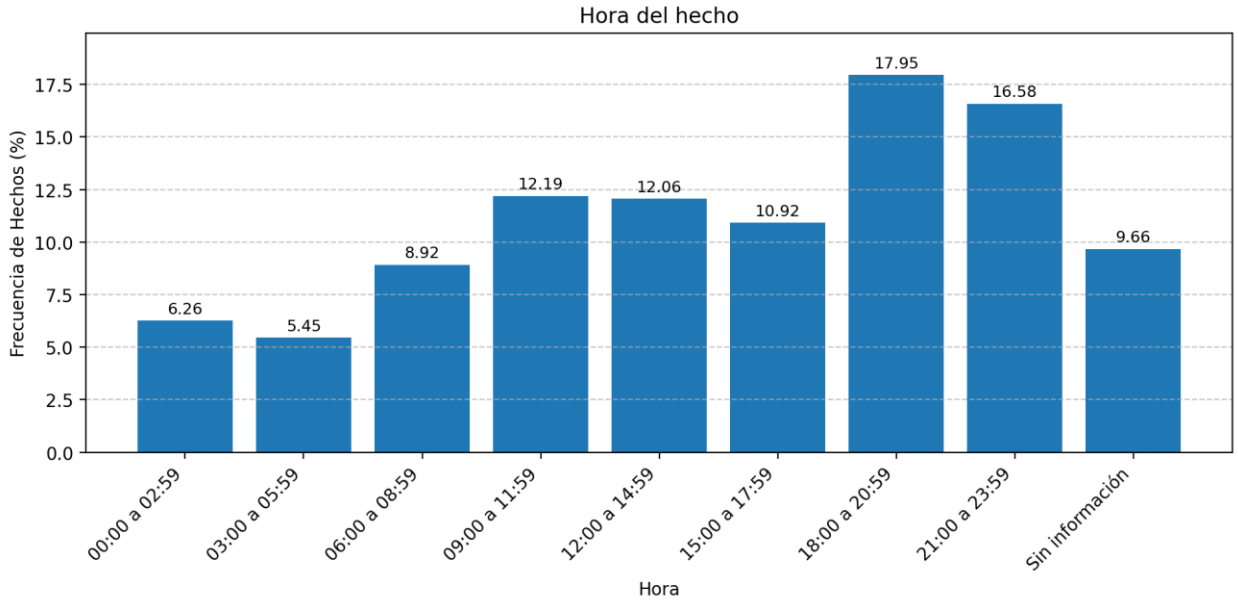
Gráfico 7. Día del hecho



Fuente: elaboración propia.

La gráfica muestra que los casos de violencia intrafamiliar contra la mujer presentan una mayor frecuencia durante los fines de semana, especialmente los domingos, día que concentra el porcentaje más alto de hechos registrados (21,70 %), seguido de los sábados (15,11 %). En contraste, los días jueves y viernes presentan las menores frecuencias, con 11,91 % y 11,76 %, respectivamente. Estos resultados evidencian un comportamiento temporal relevante, donde los eventos tienden a incrementarse durante los días de descanso y convivencia social. En términos generales, este patrón sugiere que determinados contextos sociales y familiares podrían influir en la ocurrencia de los hechos, convirtiendo la variable “día del hecho” en un elemento importante dentro del análisis exploratorio de los datos.

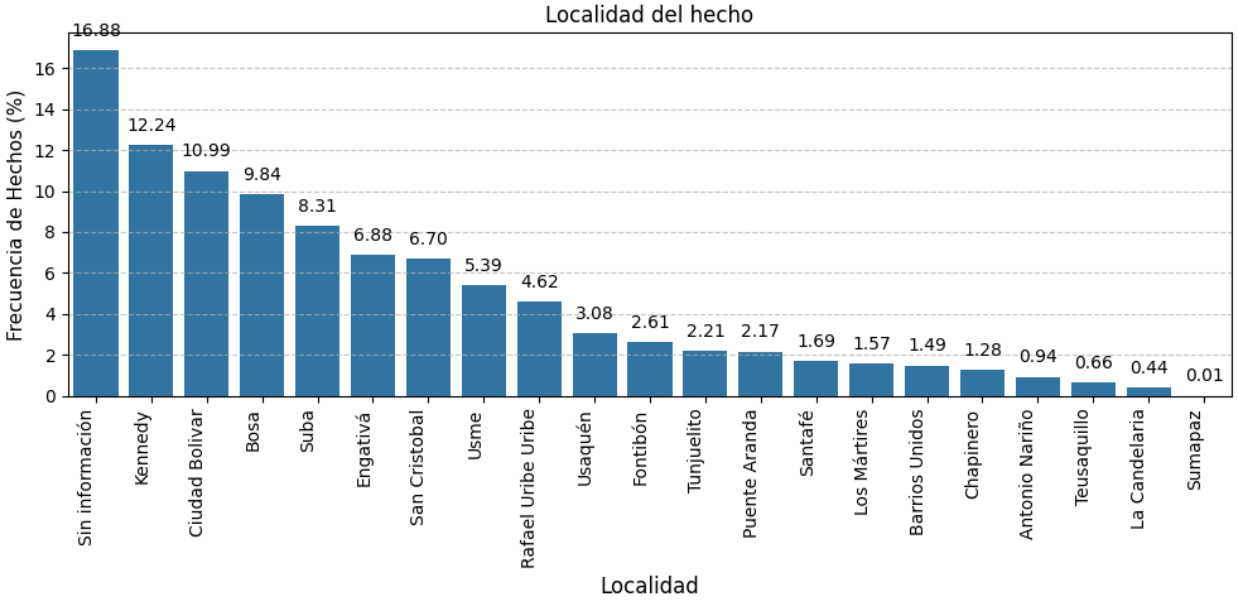
Gráfico 8. Rango de hora del día del hecho



Fuente: elaboración propia.

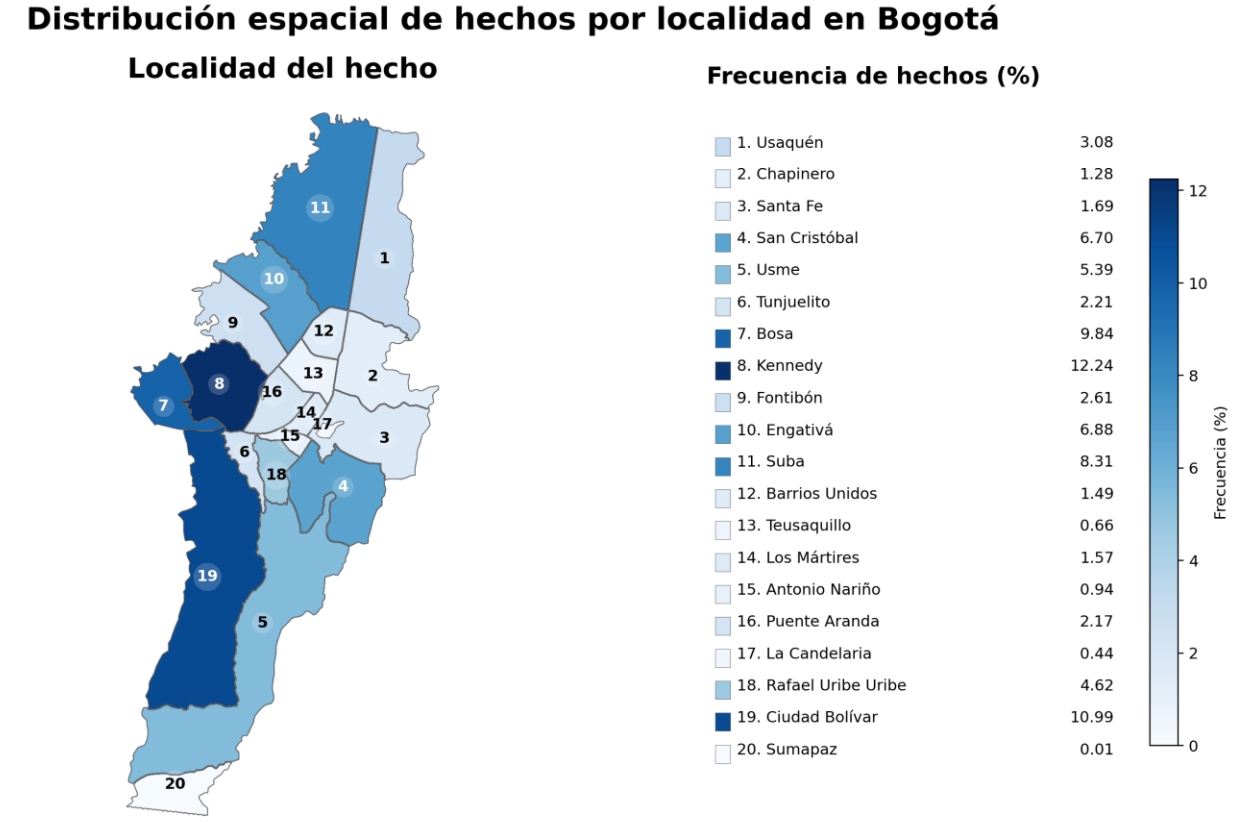
La gráfica evidencia que los casos de violencia intrafamiliarcontra la mujer presentan una mayor frecuencia durante las horas de la noche. El intervalo comprendido entre las 18:00 y las 20:59 concentra el porcentaje más alto de hechos registrados (17.95%), seguido del rango entre las 21:00 y las 23:59 (16,58 %). En contraste, las menores frecuencias se presentan durante la madrugada, especialmente entre las 03:00 y las 05:59 (5,45 %). Estos resultados permiten identificar un patrón temporal asociado al incremento de los hechos en horarios nocturnos, lo que sugiere una posible relación con dinámicas de convivencia social y familiar. En términos generales, la variable “hora del hecho” aporta información relevante para comprender el comportamiento temporal de los eventos dentro del análisis exploratorio de los datos.

Gráfico 9. Localidad del hecho



Fuente: elaboración propia.

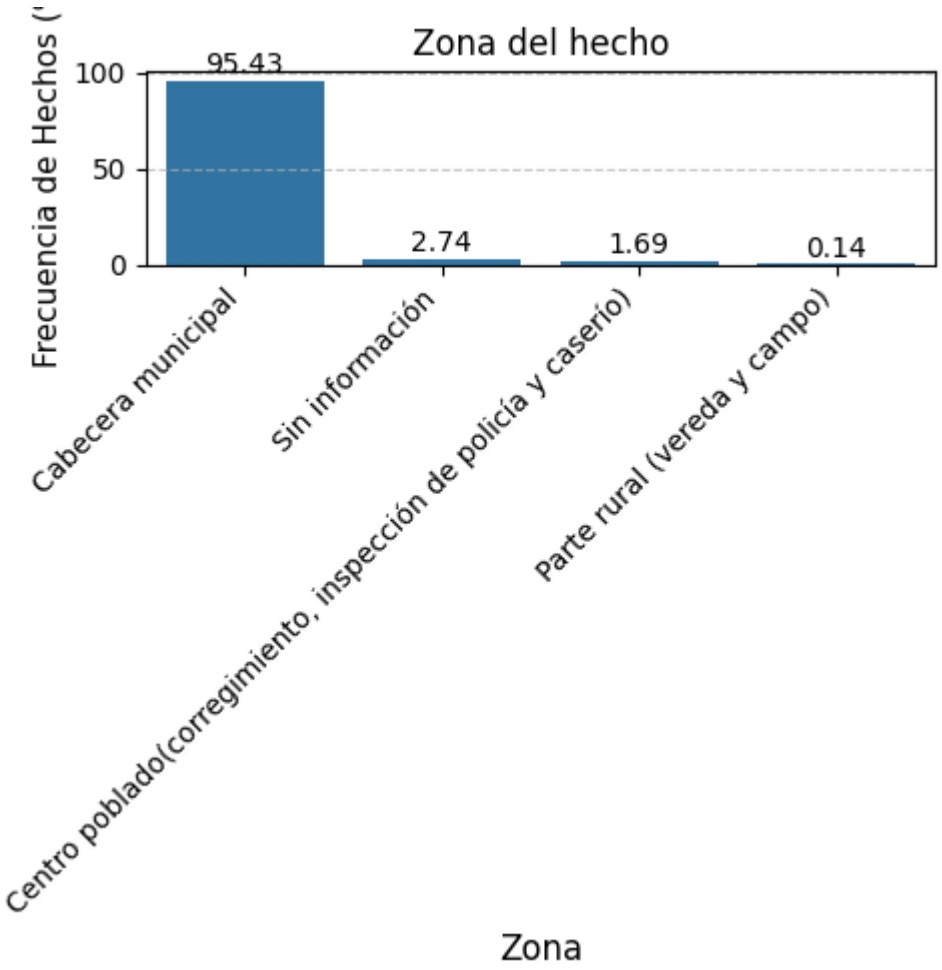
Gráfico 10. Localidad del hecho mapa



Fuente: elaboración propia.

Las gráficas anteriores muestran que las localidades con mayor frecuencia de casos de violencia intrafamiliar contra la mujer corresponden a Kennedy (12,24 %), Ciudad Bolívar (10,99 %), Bosa (9,84 %) y Suba (8,31 %). En contraste, localidades como Sumapaz, La Candelaria y Teusaquillo presentan las menores participaciones dentro del conjunto de datos (< 1 %). . Asimismo, se identifica una categoría denominada "Sin información", que representa el 16,75 % de los registros, evidenciando la existencia de valores faltantes en la variable localidad del hecho. En la fase de preparación de los datos, estos valores fueron imputados mediante la moda, sustituyendo los registros faltantes por la categoría más frecuente de la variable. Este procedimiento permitió reducir la pérdida de información y garantizar la consistencia del conjunto de datos utilizado en los análisis descriptivos y en los modelos de aprendizaje automático.

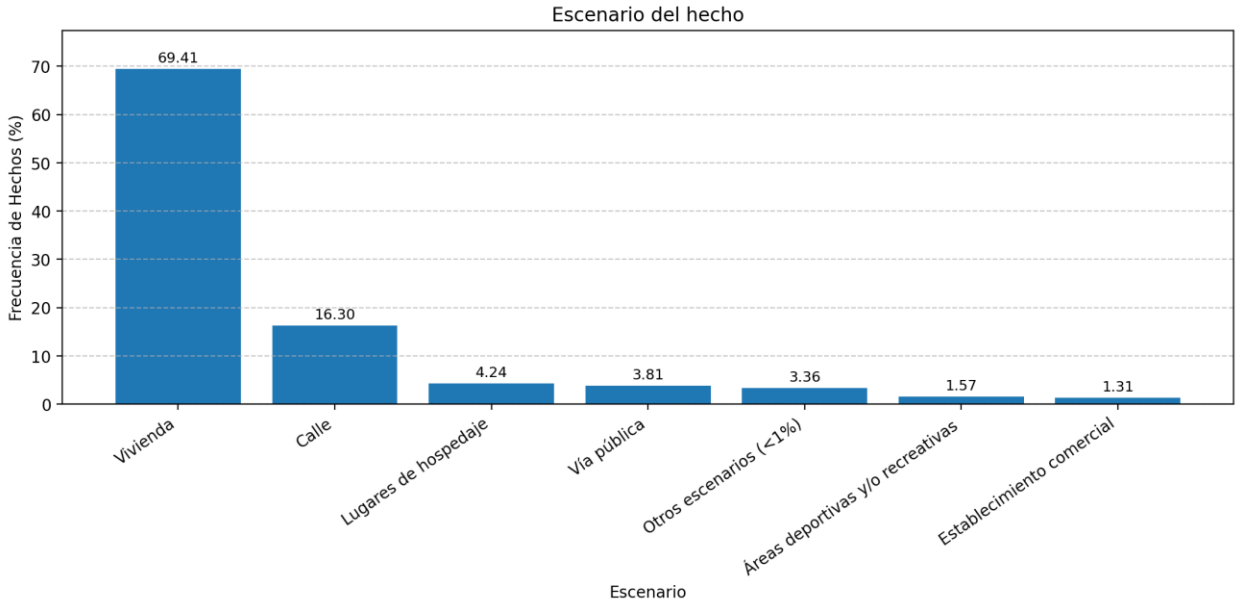
Gráfico 11. Zona del hecho



Fuente: elaboración propia.

El Gráfico 11 evidencia que la gran mayoría de los casos de violencia intrafamiliar contra la mujer ocurrieron en cabeceras municipales, las cuales concentran el 95,43 % de los hechos registrados. En contraste, las zonas rurales presentan una participación significativamente menor, con porcentajes inferiores al 1 %. Estos resultados sugieren una alta concentración del fenómeno en contextos urbanos, posiblemente asociada a factores como la densidad poblacional, mayores niveles de interacción social y una mayor capacidad de reporte institucional en las ciudades. Asimismo, se identifica una pequeña proporción de registros clasificados como “Sin información” (2,74 %), los cuales son tratados posteriormente durante la etapa de preparación de los datos para garantizar la calidad e integridad del análisis.

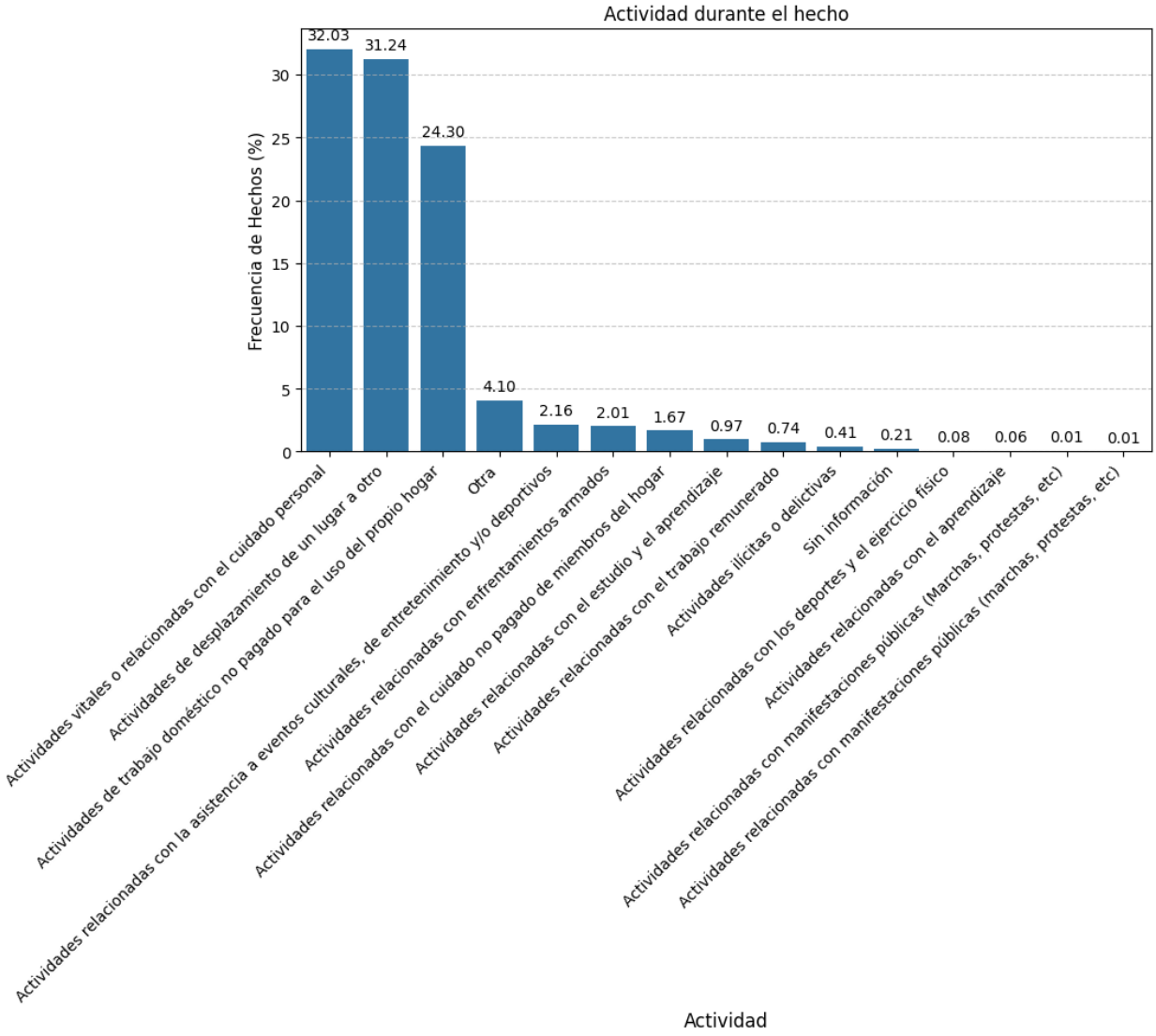
Gráfico 12. Escenario del hecho



Fuente: elaboración propia.

La gráfica evidencia que la mayoría de los casos de violencia intrafamiliar y homicidios contra la mujer ocurrieron en viviendas, categoría que concentra el 69,41 % de los hechos registrados. En segundo lugar, se encuentran las calles (16,30 %), seguidas de lugares de hospedaje y la vía pública, aunque con porcentajes considerablemente menores. Estos resultados sugieren que el entorno doméstico representa el principal escenario de ocurrencia de los hechos analizados, lo que refleja una fuerte relación entre la violencia y los espacios de convivencia familiar o cercana. Asimismo, se observa la presencia de múltiples escenarios con participaciones bajas, lo que evidencia la diversidad de contextos donde pueden presentarse este tipo de eventos. En general, la variable “escenario del hecho” aporta información relevante para comprender las dinámicas espaciales y contextuales asociadas a la violencia contra la mujer.

Gráfico 13. Actividad durante el hecho

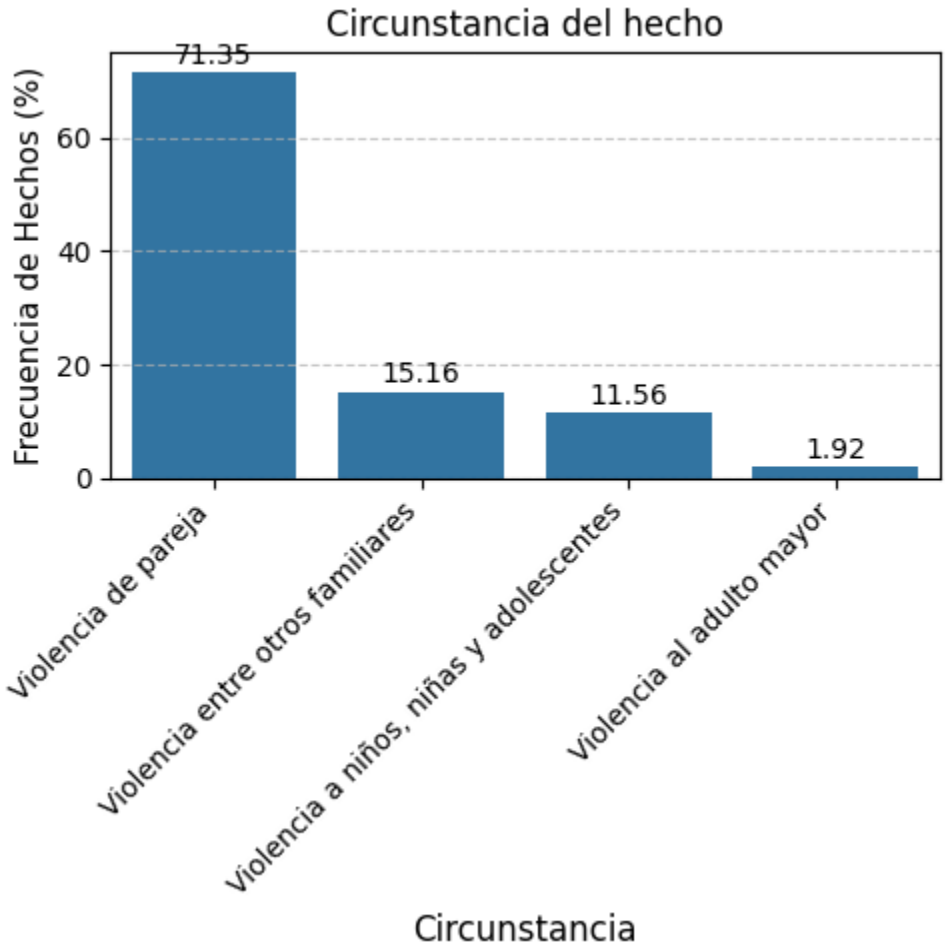


Fuente: elaboración propia.

El gráfico muestra que la mayor parte de los casos de violencia intrafamiliar contra la mujer ocurrieron mientras las víctimas realizaban actividades relacionadas con el cuidado personal (32,03 %) y actividades de desplazamiento de un lugar a otro (31,24 %). En tercer lugar, se encuentran los trabajos domésticos no remunerados (24,30 %). Estos resultados evidencian que gran parte de los hechos se desarrollan en contextos cotidianos y domésticos, especialmente durante actividades asociadas a la vida familiar y personal. Diversos organismos internacionales y estudios sobre violencia basada en

género han señalado que el entorno doméstico constituye uno de los principales espacios de exposición al riesgo para las mujeres, debido a dinámicas de convivencia, relaciones de poder y situaciones de dependencia o conflicto dentro del núcleo familiar.[62] Por otro lado, las demás categorías presentan porcentajes considerablemente menores, lo que indica una menor concentración de eventos en actividades específicas como el estudio, el trabajo remunerado o las actividades deportivas. En términos generales, esta variable permite comprender mejor el contexto situacional en el que ocurren los hechos analizados.

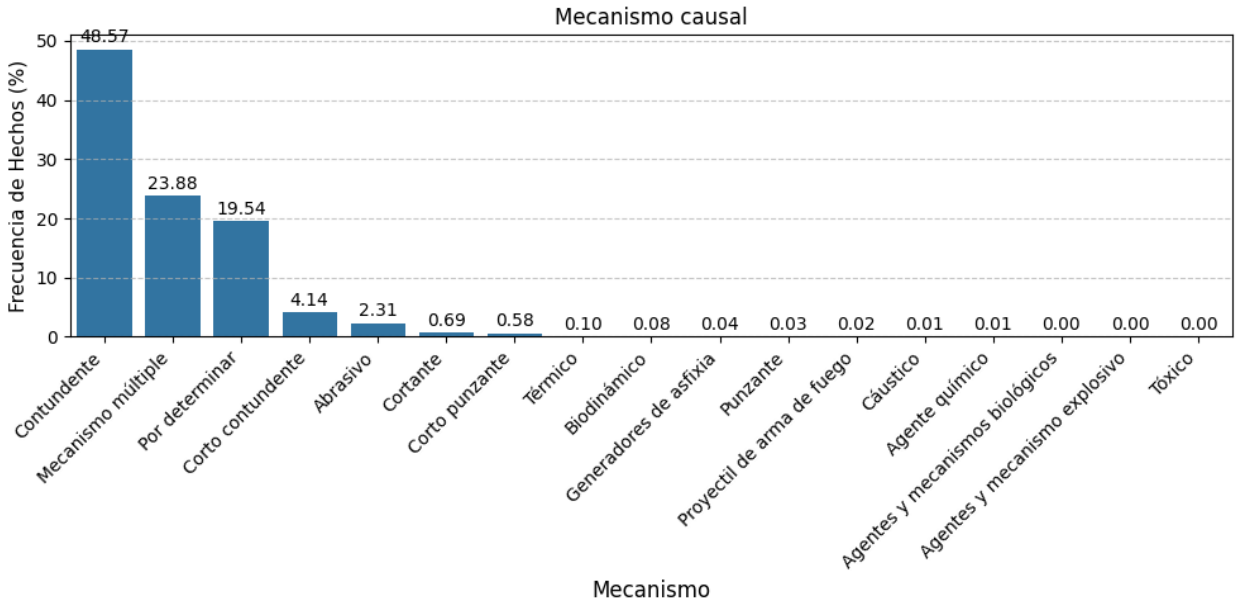
Gráfico 14.Circunstancia del hecho



Fuente: elaboración propia.

El gráfico evidencia que la principal circunstancia asociada a los casos de violencia intrafamiliar contra la mujer corresponde a la violencia de pareja, categoría que concentra el 71,35 % de los hechos registrados. En segundo lugar, se encuentra la violencia entre otros familiares (15,16 %), seguida de la violencia a niños, niñas y adolescentes (11,56 %). Estos resultados muestran una fuerte concentración de los hechos en contextos familiares y relaciones cercanas de convivencia. Diversos estudios y organismos internacionales han señalado que la violencia ejercida por parejas o familiares representa una de las formas más frecuentes de violencia basada en género, debido a dinámicas de control, dependencia y desigualdad presentes en el entorno doméstico.[63] Asimismo, informes de ONU Mujeres indican que una alta proporción de homicidios de mujeres ocurre precisamente en el ámbito familiar o por parte de parejas sentimentales.[64]] Por otro lado, las demás circunstancias presentan porcentajes considerablemente menores, lo que evidencia una menor concentración de casos en escenarios relacionados con delitos comunes, conflictos externos u otras motivaciones específicas.

Gráfico 15. Mecanismo causal

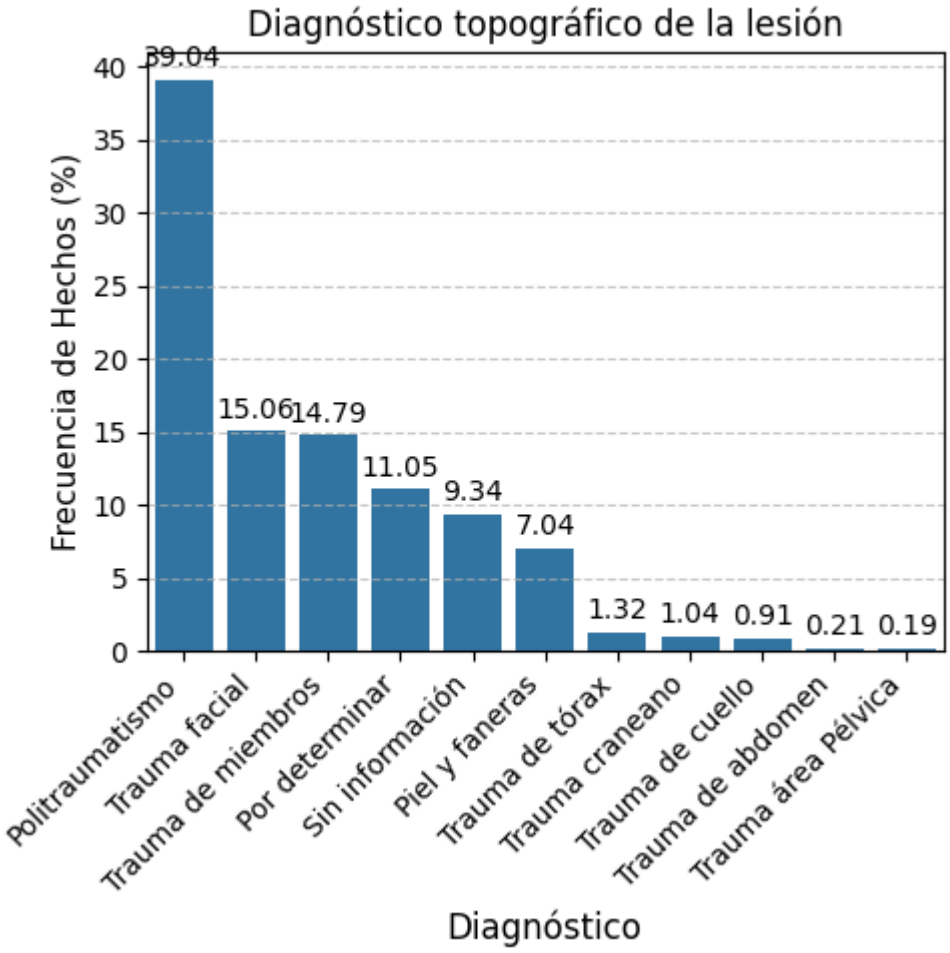


Fuente: elaboración propia.

La Gráfica 15 muestra que el principal mecanismo causal asociado a los casos de

violencia intrafamiliar contra la mujer corresponde al uso de mecanismos contundentes, categoría que concentra el 48,57 % de los hechos registrados. En segundo lugar, se encuentran los mecanismos múltiples (23,88 %) y los casos por determinar (19,54 %). Estos resultados sugieren una alta prevalencia de agresiones físicas directas dentro de los eventos analizados. Asimismo, mecanismos como armas cortopunzantes (0,58 %) y armas de fuego (0,020 %) presentan participaciones considerablemente menores dentro del conjunto de datos. La presencia de registros clasificados como “Por determinar” evidencia limitaciones en algunos procesos de registro o recolección de información; sin embargo, estos valores son tratados posteriormente durante la etapa de preparación de los datos mediante técnicas de imputación orientadas a mejorar la integridad y consistencia del conjunto de datos analizado.

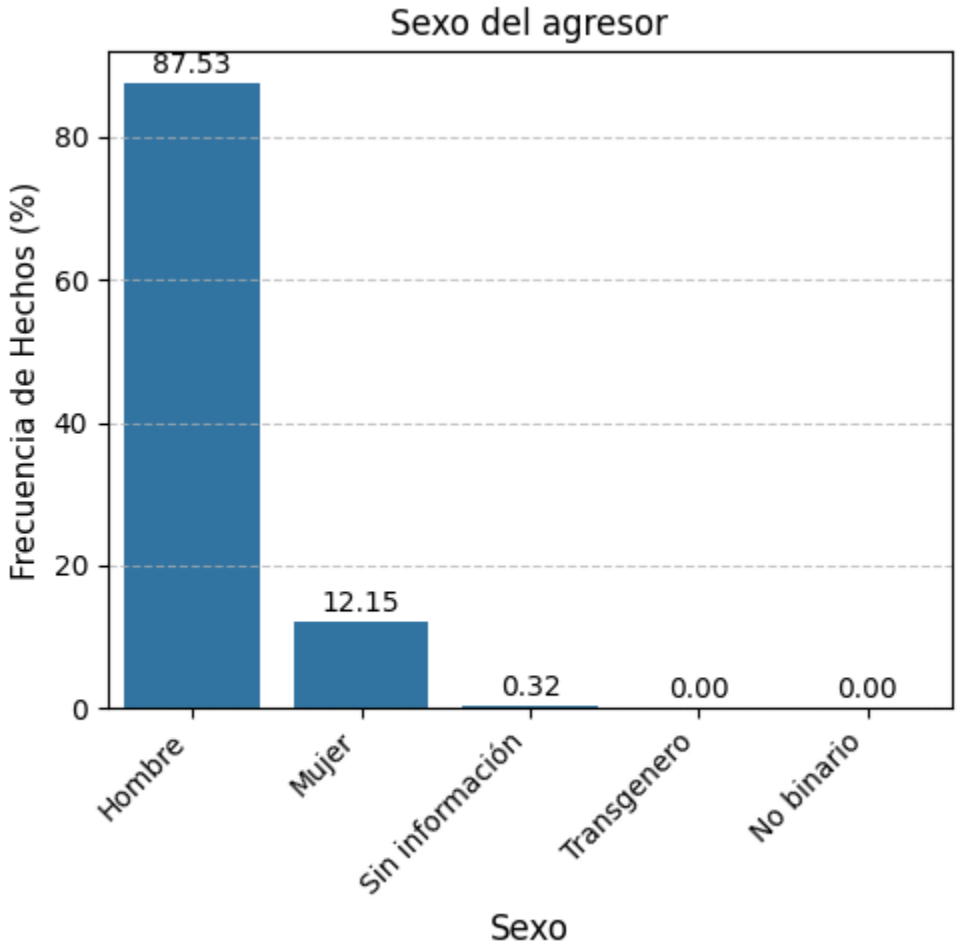
Gráfico 16. Diagnóstico topográfico de la lesión



Fuente: elaboración propia.

El gráfico evidencia que el diagnóstico topográfico de lesión más frecuente corresponde al politraumatismo, categoría que concentra el 39,04% de los casos registrados. En segundo lugar, se encuentran los traumatismos faciales (15,06 %) y los traumas de miembros (14,78 %). Asimismo, se observa una proporción relevante de registros clasificados “por determinar” (11,05 %) y “sin información” (9,34 %), lo que evidencia la presencia de datos incompletos dentro de esta variable. Estos valores son tratados posteriormente durante la etapa de preparación de los datos mediante técnicas de imputación, con el propósito de mejorar la calidad e integridad del conjunto de datos analizado.

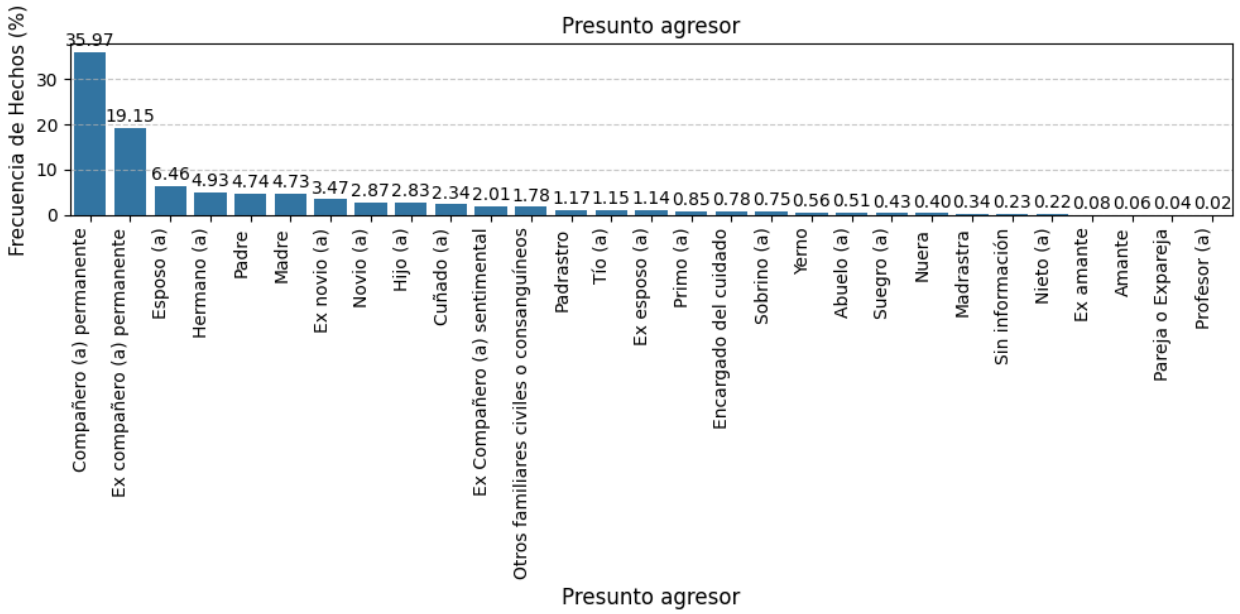
Gráfico 17. Sexo del agresor



Fuente: elaboración propia.

La gráfica evidencia que la gran mayoría de los agresores corresponden al sexo masculino, categoría que concentra el 87,53 % de los casos registrados, mientras que las mujeres representan el 12,15 %. Estos resultados mantienen coherencia con múltiples estudios e informes internacionales sobre violencia basada en género, los cuales señalan que las mujeres son las principales víctimas de agresiones ejercidas por hombres dentro de contextos de violencia intrafamiliar y de pareja. Asimismo, la baja proporción de registros “Sin información” (0,32 %) evidencia una alta completitud de esta variable dentro del conjunto de datos analizado.[63]

Gráfico 18. Presunto agresor



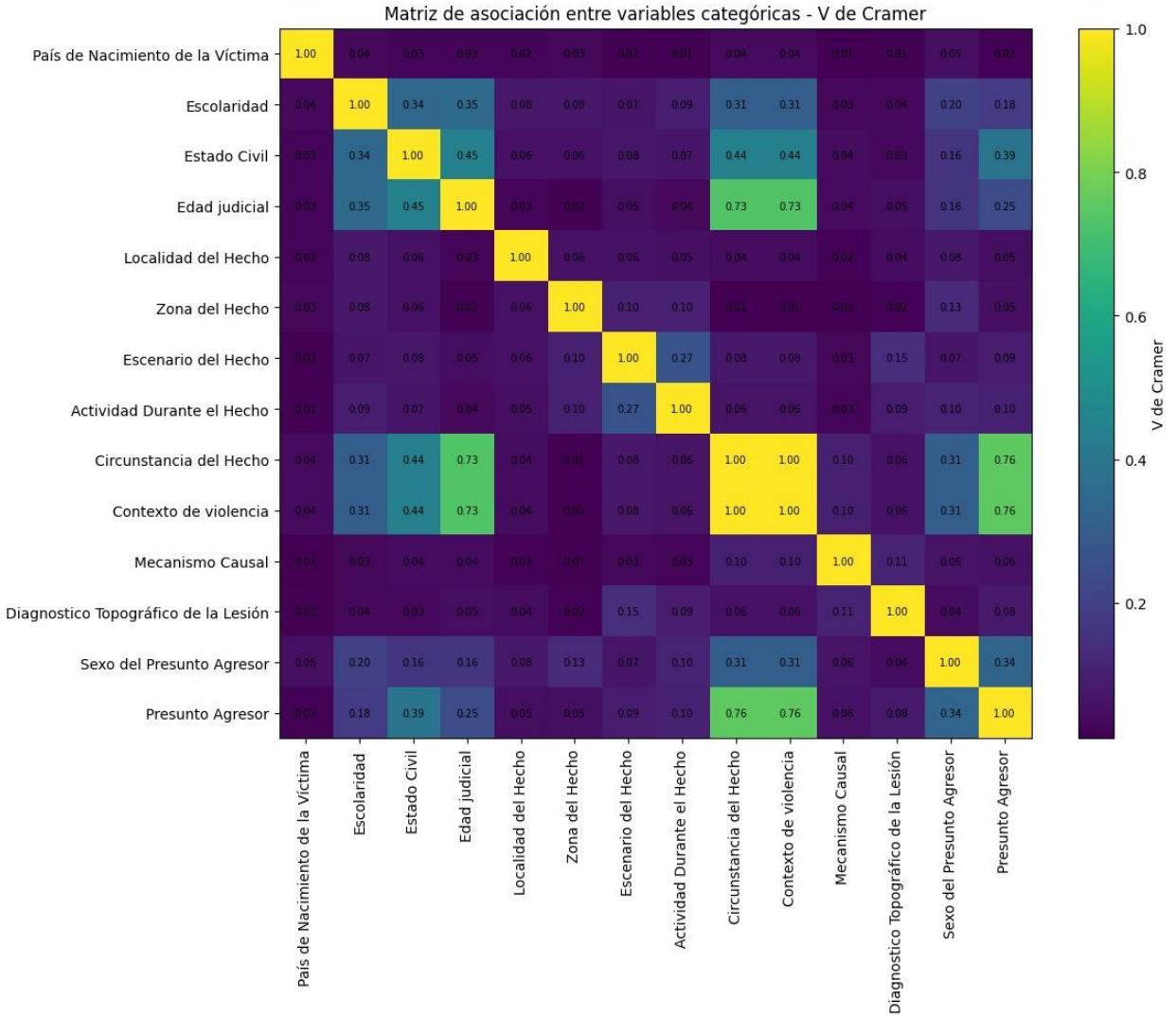
Fuente: elaboración propia.

La gráfica evidencia que los principales presuntos agresores corresponden a la pareja permanente (35,97 %) y la expareja permanente (19,15 %), seguidos por el esposo, hermano y padre, aunque con porcentajes considerablemente menores. Estos resultados refuerzan los hallazgos observados previamente en las variables de contexto y circunstancia del hecho, donde predominaban los entornos familiares y de convivencia cercana. En términos generales, la información sugiere que gran parte de los casos de violencia contra la mujer ocurre dentro de relaciones afectivas o familiares, comportamiento ampliamente documentado en estudios sobre violencia de género y violencia intrafamiliar [65].

Una vez agotado el análisis univariado, se procede a realizar uno multivariado. El primer paso en este sentido ha sido construir una matriz de asociación de Cramer para identificar posibles relaciones entre las variables categóricas consideradas para el proceso de segmentación. Esta medida estadística, derivada del estadístico Chi-cuadrado, permite cuantificar la intensidad de la relación existente entre pares de variables categóricas mediante valores comprendidos entre 0 y 1, donde valores cercanos a cero indican ausencia de asociación y valores próximos a uno reflejan

relaciones más fuertes.[66], [67]

Gráfico 19. Matriz de asociación de Cramer



Fuente: elaboración propia.

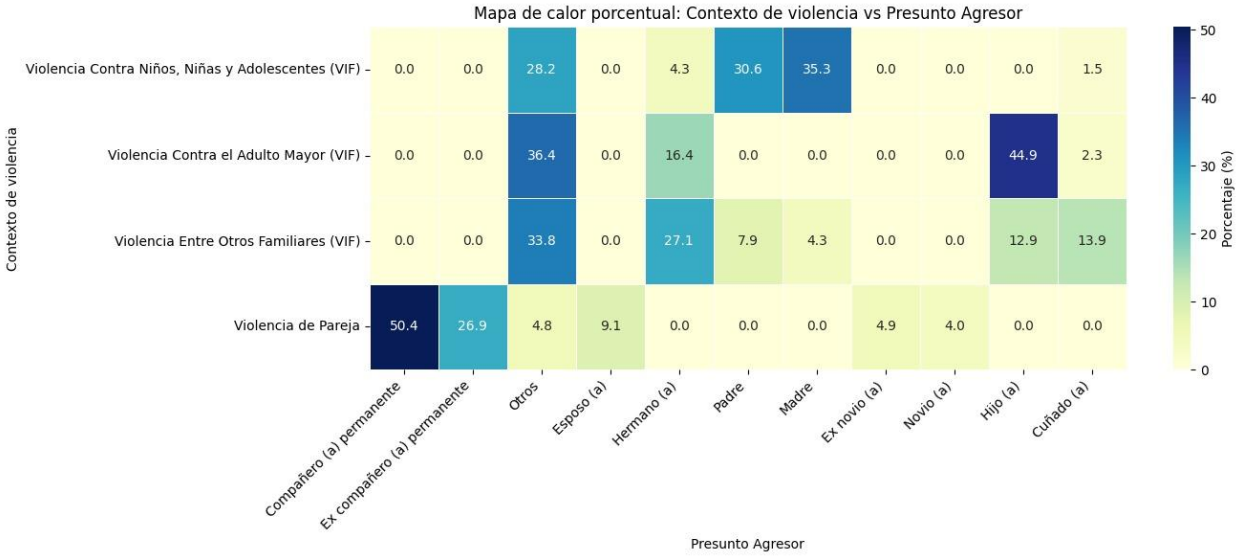
La matriz obtenida permitió evaluar simultáneamente el grado de dependencia entre las variables sociodemográficas, territoriales y de caracterización de los hechos, proporcionando una visión integral de las interacciones presentes en el conjunto de datos antes de aplicar las técnicas de reducción de dimensionalidad y agrupamiento.[66], [68]

Los resultados evidencian que la mayoría de las variables presentan asociaciones

débiles o moderadas, lo que sugiere que cada una aporta información complementaria al análisis. Sin embargo, se identificaron relaciones de mayor intensidad entre el contexto de violencia y el presunto agresor (V = 0,76) y circunstancia del hecho y presunto agresor (V = 9,75), así como entre el contexto de violencia y la edad judicial de la víctima (V = 0,73), indicando que estos factores se encuentran estrechamente relacionados con la dinámica de los eventos analizados. También se observan asociaciones moderadas entre el contexto de violencia y el estado civil (V = 0,44), y entre el estado civil y el presunto agresor (V = 0,39).

En contraste, variables como la localidad del hecho, la zona del hecho y el país de nacimiento de la víctima presentan niveles de asociación bajos respecto al resto de variables, lo que evidencia una mayor independencia relativa dentro de la estructura de datos. En conjunto, estos resultados sugieren la existencia de patrones subyacentes entre varias de las variables categóricas analizadas, justificando la aplicación de técnicas de escalamiento óptimo mediante PRINCUAL para sintetizar la información y facilitar la posterior identificación de segmentos homogéneos a través del algoritmo K-Means.

Gráfico 20. Tipo de casos



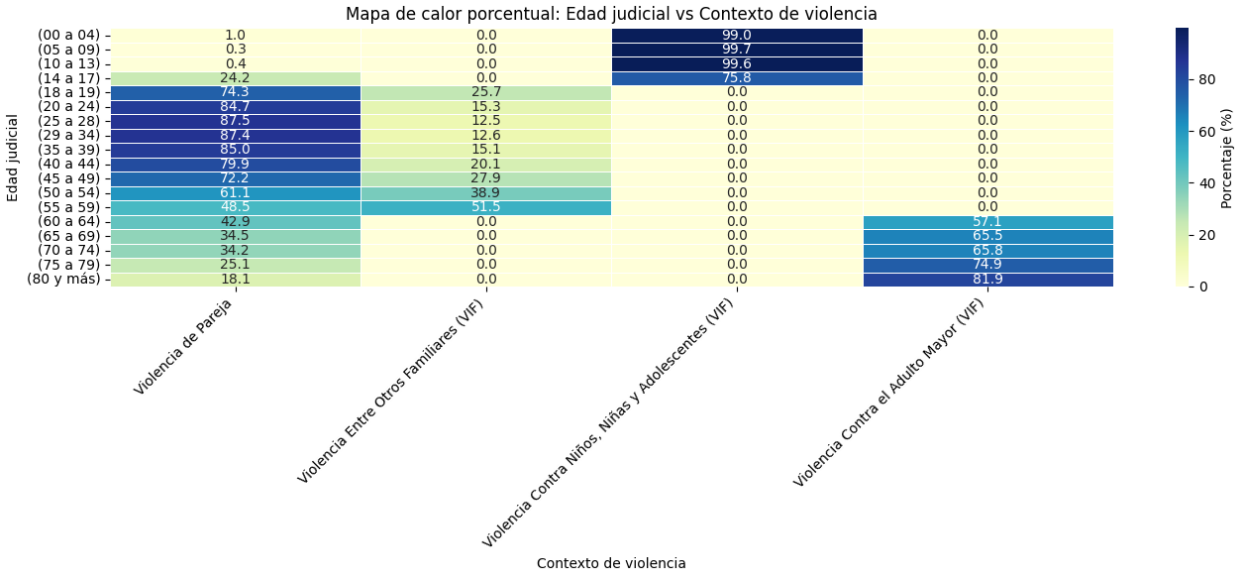
Fuente: elaboración propia.

La gráfica anterior muestra cómo se distribuye el tipo de presunto agresor dentro de cada

contexto de violencia. En la violencia contra niñas y adolescentes, los principales agresores registrados son la madre (35,3 %), el padre (30,6 %) y la categoría “Otros” (28,2 %); los hermanos representan 4,3 %. En la violencia contra el adulto mayor, predominan los hijos (44,9 %), seguidos de “Otros” (36,4 %) y los hermanos (16,4 %), lo que evidencia una concentración de este contexto en relaciones familiares descendentes o cercanas.

En la violencia entre otros familiares, la distribución es más diversa: “Otros” representa 33,8 %, los hermanos 27,1 %, los cuñados 13,9 % y los hijos 12,9 %, mientras que padres y madres tienen participaciones menores. Por su parte, la violencia de pareja presenta una estructura claramente diferenciada: el principal presunto agresor es el compañero permanente (50,4 %), seguido del excompañero permanente (26,9 %), el esposo (9,1 %), otros agresores (4,8 %), la expareja o exnovio (4,9 %) y el novio (4,0 %). En conjunto, la gráfica muestra una asociación coherente entre el contexto de violencia y el vínculo con el agresor, lo cual coincide con el valor elevado de V de Cramer observado anteriormente.

Gráfico 21. Edad judicial y contexto de violencia

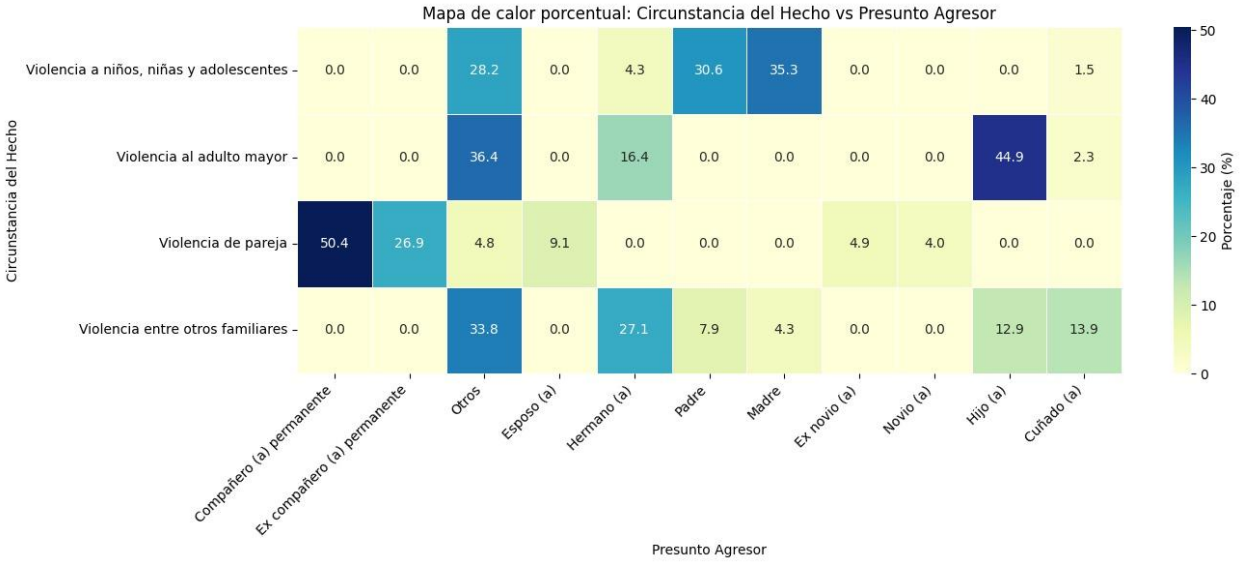


Fuente: elaboración propia.

La gráfica anterior muestra la distribución porcentual del contexto de violencia dentro de cada grupo de edad. En las víctimas menores de 14 años predomina casi totalmente la violencia contra niñas y adolescentes: representa 99,0 % entre 0 y 4 años, 99,7 % entre 5 y 9 y 99,6 % entre 10 y 13. En el grupo de 14 a 17 años, esta categoría continúa siendo mayoritaria con 75,8 %, pero aparece una proporción importante de violencia de pareja, equivalente al 24,2 %. A partir de los 18 años, la violencia de pareja se convierte en el contexto predominante, alcanzando sus valores más altos entre los 25 y 34 años: 87,5 % en el grupo de 25 a 28 y 87,4 % entre 29 y 34 años.

Desde los 35 años, la proporción de violencia de pareja disminuye progresivamente y aumenta la violencia entre otros familiares: pasa de 15,1 % entre los 35 y 39 años a 38,9 % entre los 50 y 54, hasta superar a la violencia de pareja en el grupo de 55 a 59 años, con 51,5 %. Desde los 60 años aparece como categoría principal la violencia contra el adulto mayor, cuya participación aumenta con la edad: 57,1 % entre 60 y 64 años, 74,9 % entre 75 y 79 y 81,9 % en personas de 80 años o más. El patrón evidencia una transición clara del contexto registrado a lo largo del ciclo vital; sin embargo, parte de esta asociación es estructural, porque las categorías “violencia contra niñas y adolescentes” y “violencia contra el adulto mayor” se definen precisamente según la edad de la víctima. Por ello, los valores de cero no necesariamente significan ausencia absoluta de violencia, sino que pueden responder a las reglas empleadas para clasificar cada caso.

Gráfico 22. Circunstancia del hecho *vs.* Presunto agresor



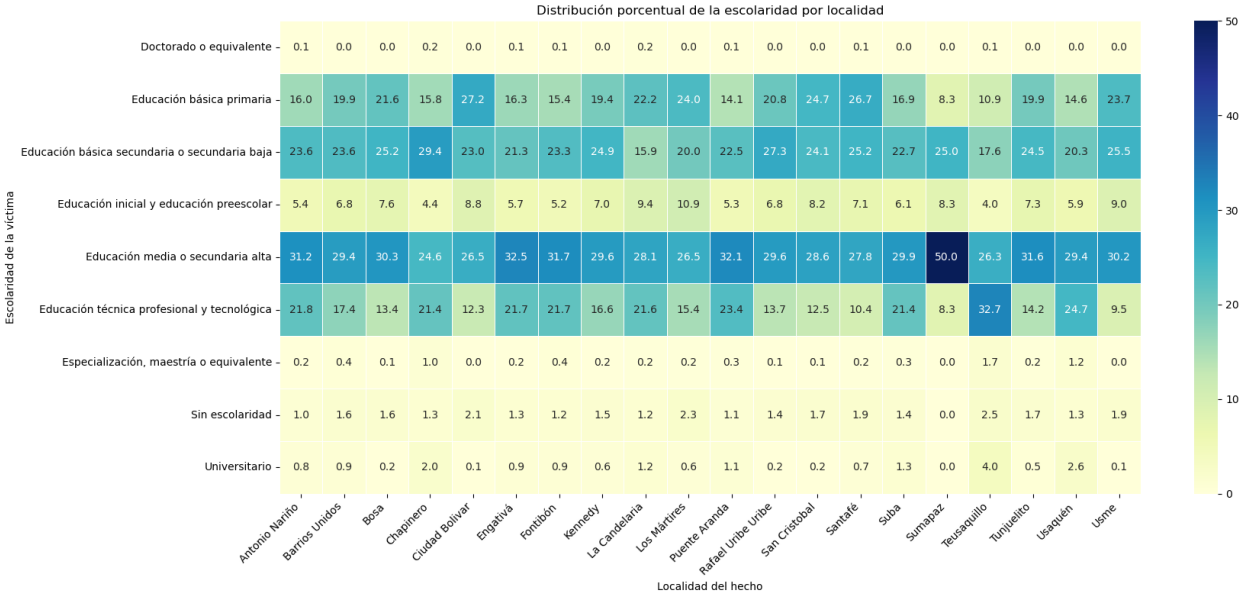
Fuente: elaboración propia.

La gráfica anterior muestra la distribución del presunto agresor dentro de cada circunstancia del hecho. En las circunstancias asociadas a violencia familiar aparecen patrones claros. En la violencia de pareja, predominan el compañero permanente (50,4 %) y el excompañero permanente (26,8 %), seguidos del esposo (9,1 %), “Otros” (4,8 %), exnovio (4,9 %) y novio (4,0 %). En la violencia contra niñas y adolescentes, los principales agresores son la madre (35,3 %) y el padre (30,6 %), mientras que “Otros” representa 28,2 %. En la violencia contra el adulto mayor, sobresalen los hijos (44,9 %), “Otros” (36,4 %) y los hermanos (16,4 %). Para la violencia entre otros familiares, la distribución es más heterogénea: “Otros” alcanza 33,8 %, los hermanos 27,1 %, los cuñados 13,9 % y los hijos 13,0 %.

El feminicidio presenta una composición más diversa: “Otros” concentra 47,1 %, seguido del compañero permanente con 21,6 %, el esposo con 11,8 %, el excompañero con 9,8 % y el novio con 5,9 %. En numerosas circunstancias (como hurto, riña, sicariato, ajuste de cuentas, negligencia, violencia sexual y “sin información”) el 100 % aparece en la categoría “Otros”. Esto no significa necesariamente que todos los agresores tengan el mismo vínculo, sino que las categorías disponibles en presunto agresor posiblemente no identifican relaciones específicas para esos casos, o que los valores desconocidos

fueron agrupados como “Otros”.

Gráfico 23. Localidad y escolaridad de las víctimas



Fuente: elaboración propia.

Esta gráfica muestra la correlación entre las localidades y el nivel de estudios obtenidos por las mujeres. La gráfica muestra que, en la mayoría de las localidades, la categoría predominante es educación media o secundaria alta, con porcentajes generalmente cercanos al 25 % y 32 %. Los valores más altos se observan en Engativá (32,5 %), Puente Aranda (32,1 %), Fontibón (31,7 %), Tunjuelito (31,6 %) y Antonio Nariño (31,2 %). La educación básica secundaria o secundaria baja constituye, igualmente, una proporción importante, especialmente en Chapinero (29,4 %), Rafael Uribe Uribe (27,3 %), Usme (25,5 %), Bosa (25,2 %) y Santa Fe (25,2 %). Por su parte, la educación primaria presenta mayor participación en Ciudad Bolívar (27,2 %), Santa Fe (26,7 %), San Cristóbal (24,7 %), Los Mártires (24,0 %) y Usme (23,7 %), lo que evidencia diferencias en la composición educativa de las víctimas registradas entre localidades.

La educación técnica y tecnológica tiene una presencia destacada en Teusaquillo (32,7 %), Usaquén (24,7 %), Puente Aranda (23,4 %) y Antonio Nariño (21,8 %). En contraste, las categorías de universitario, especialización, maestría y doctorado representan

proporciones muy bajas en casi todas las localidades, generalmente inferiores al 2 %, aunque Teusaquillo alcanza 4 % en educación universitaria y 1,7 % en especialización o maestría. Sumapaz tiene un 50 % de los datos en educación media. Sin embargo, hay que entender esta localidad no tiene una gran cantidad de registros, por lo que pequeñas variaciones en el número de casos producen cambios porcentuales elevados.

En términos generales, el análisis exploratorio de los datos permitió identificar patrones temporales, espaciales y contextuales asociados a los casos de violencia intrafamiliar contra la mujer. Los resultados evidenciaron una alta concentración de los hechos en entornos domésticos y relaciones de pareja, así como una mayor incidencia durante horarios nocturnos y fines de semana. Asimismo, se identificó la presencia de variables con valores faltantes o categorías no informativas, las cuales fueron consideradas dentro del proceso de preparación de los datos para garantizar la calidad e integridad de la información. Esta etapa permitió comprender la estructura y comportamiento del conjunto de datos, proporcionando una base sólida para las fases posteriores de preprocesamiento y aplicación de técnicas de *machine learning*.

2.2. Preparación de los datos

Una vez finalizadas las etapas de comprensión y exploración de los datos, se desarrolló un proceso de preparación orientado a transformar los registros originales en estructuras analíticas aptas para la implementación de modelos de *machine learning*. Esta fase incluyó actividades de limpieza, normalización, imputación, reducción de cardinalidad, ingeniería de variables y transformación dimensional, buscando garantizar calidad, consistencia y representatividad en la información utilizada durante el modelado.[69]

Como punto de partida, se usó el conjunto de datos correspondiente a violencia intrafamiliar . Posteriormente, se aplicaron filtros territoriales y poblacionales, conservando únicamente registros correspondientes a mujeres víctimas en la ciudad de Bogotá. Como resultado de este procedimiento se obtuvo una base inicial compuesta por 153.597 registros y 38 variables.

El flujo general del proceso de preparación de datos es el siguiente:

Imagen 13. Flujo de preparación de los datos
Fuente: elaboración propia

Después de lo anterior, se realizó un proceso de normalización textual orientado a facilitar el procesamiento computacional de las variables categóricas. Este procedimiento incluyó la conversión de nombres de columnas y categorías a minúsculas, eliminación de espacios adicionales y normalización de caracteres especiales y tildes. Asimismo, se eliminaron variables no informativas como “Sin información”, “Por determinar” y “No aplica”, debido a que estas podían introducir ruido y afectar el desempeño de los algoritmos implementados [69] donde el *set* datos final queda con 153.597 registros y 20 variables.

Con el propósito de garantizar la calidad de la información utilizada en el modelado, se realizó un análisis de completitud sobre las variables disponibles. Para ello, se calcularon porcentajes de disponibilidad de información y se definió un umbral mínimo de

completitud del 60 % como criterio de inclusión para variables categóricas dentro de los modelos predictivos. Bajo este criterio se seleccionaron variables relacionadas con el escenario del hecho, actividad durante el hecho, contexto de violencia, zona, presunto agresor, escolaridad, mecanismo causal, sexo del presunto agresor y diagnóstico topográfico de la lesión.

Resumen general de la preparación de datos:

Tabla 4. Resumen de preparación de los datos

Etapa	Objetivo	Técnicas utilizadas
Filtrado	Delimitar población objetivo	Bogotá y mujeres
Limpieza	Normalizar registros	Minúsculas, eliminación de caracteres y categorías no informativas
Análisis de completitud	Selección de variables útiles	Umbral mínimo del 60 %
Imputación	Reconstrucción de categorías faltantes	Imputación por la moda
Reducción de cardinalidad	Disminución del ruido analítico	Agrupación de categorías poco frecuentes
Reducción dimensional	Representación factorial de variables categóricas	MCA
Ingeniería temporal	Captura de tendencias y comportamiento histórico	Lags, medias móviles y variables calendario

Fuente: elaboración propia.

En paralelo, se implementó un proceso de imputación para reconstruir categorías faltantes en variables críticas como localidad del hecho, mecanismo causal, contexto de violencia, sexo del presunto agresor y diagnóstico topográfico de la lesión. Para ello, se aplicó la técnica de imputación por moda, reemplazando los valores faltantes por la categoría con mayor frecuencia observada en cada variable.[54] Esta estrategia permitió mejorar la completitud de la base de datos y garantizar una mayor consistencia en el

análisis posterior, minimizando la pérdida de información y facilitando la aplicación de los algoritmos de segmentación y predicción desarrollados en el estudio.[55]

A posteriori, se desarrolló un proceso de reducción de cardinalidad sobre variables categóricas con alta dispersión. Para ello, las categorías menos frecuentes fueron agrupadas bajo la etiqueta “Otros”, priorizando aquellas que acumulaban mayor representatividad estadística dentro de cada variable. Este procedimiento permitió disminuir el ruido analítico y evitar la generación de agrupamientos excesivamente pequeños o poco representativos durante los procesos de segmentación.

Para los modelos de segmentación se empleó la técnica PRINCUAL (Principal Components Analysis by Alternating Least Squares), perteneciente a la familia de métodos de escalamiento óptimo, debido a la naturaleza predominantemente categórica de las variables analizadas. Esta técnica permitió reducir la dimensionalidad del conjunto de datos mediante la obtención de dimensiones latentes que preservan las principales estructuras de asociación entre las variables, proporcionando una representación más adecuada para la posterior aplicación de algoritmos de agrupamiento.[70]

El proceso de preparación para modelos de segmentación es el siguiente:

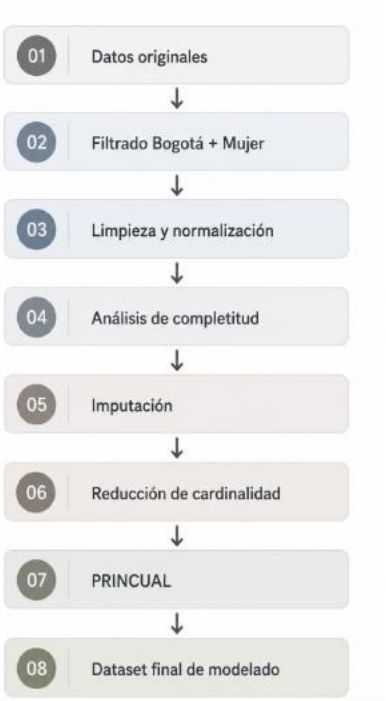


Imagen 14. Proceso preparación datos para segmentación
Fuente: Elaboración propia

En el caso de los modelos predictivos, la preparación de los datos estuvo orientada a la construcción de una estructura temporal basada en la combinación localidad-mes. Para cada combinación territorial y temporal se calcularon variables agregadas relacionadas con el número total de casos de violencia intrafamiliar, constituyendo la serie histórica utilizada para el entrenamiento del modelo.

Adicionalmente, se incorporaron variables demográficas derivadas de la distribución de edades de las víctimas, incluyendo edad promedio, mediana y proporciones por grupos etarios. Posteriormente, se construyeron variables calendario y características derivadas de series temporales, tales como año, mes, trimestre y transformaciones seno y coseno orientadas a representar comportamientos estacionales [5].

Con el propósito de capturar patrones históricos y comportamiento reciente del fenómeno, se generaron variables temporales derivadas como rezagos (lags), promedios móviles, sumas móviles, diferencias mensuales y variaciones porcentuales para los

principales indicadores de violencia. Estas características permitieron incorporar dependencia temporal y tendencias acumuladas dentro de los modelos predictivos [5].

Preparación temporal para modelos predictivos:

Imagen 15. Preparación para modelos predictivos
Fuente: elaboración propia.

Finalmente, se construyeron las variables objetivo correspondientes a horizontes de predicción de uno, dos y tres meses hacia adelante. Debido a que las últimas observaciones no contaban con información futura suficiente para todos los horizontes, dichos registros fueron excluidos del conjunto final de modelado. Como resultado integral de esta fase, se consolidó una base analítica robusta y estructurada, apta para la implementación de modelos supervisados de predicción y algoritmos no supervisados de segmentación aplicados al estudio de la violencia contra la mujer en Bogotá.

Capítulo III.

Machine learning aplicado a la violencia contra la mujer

Una vez finalizada la etapa de preparación de los datos, se procedió a la implementación de los modelos de *machine learning* definidos para el presente proyecto aplicado. Considerando la naturaleza del fenómeno estudiado y los objetivos planteados en la investigación, se desarrollaron dos enfoques analíticos complementarios: modelos supervisados orientados a la predicción de la cantidad de casos de violencia contra la mujer y modelos no supervisados enfocados en la identificación de patrones y segmentaciones asociadas al comportamiento del fenómeno.

Para el componente predictivo se implementaron los algoritmos Random Forest y XGBoost, debido a su capacidad para modelar relaciones no lineales, manejar estructuras complejas de datos y capturar comportamientos temporales asociados a fenómenos multicausales [14], [26]. Ambos modelos fueron desarrollados utilizando una estructura temporal basada en la combinación localidad-mes, incorporando variables históricas, territoriales, sociodemográficas y contextuales derivadas del proceso de preparación de datos.

Por otra parte, para el componente de segmentación se implementaron los algoritmos K-Means y BIRCH, los cuales permitieron identificar agrupamientos y perfiles asociados a características sociodemográficas, territoriales y de *modus operandi* presentes en los registros analizados. Debido al predominio de variables categóricas dentro del conjunto de datos, previo a la segmentación se aplicó un Análisis de Correspondencias Múltiples (MCA), con el propósito de representar las asociaciones entre categorías dentro de un espacio factorial de menor dimensión [58].

A continuación, se presentan los flujos metodológicos implementados para los componentes predictivo y de segmentación desarrollados durante la etapa de modelado.

Imagen 16. Flujos modelos supervisado y no supervisado
Fuente: elaboración propia.

3.1. Modelos predictivos

Para el desarrollo de los modelos supervisados se construyó una estructura longitudinal basada en la combinación localidad-mes, permitiendo representar temporalmente el comportamiento de los casos registrados de violencia contra la mujer en Bogotá. Esta estructura facilitó la incorporación de variables históricas y temporales destinadas a capturar patrones de persistencia, tendencia y estacionalidad dentro del fenómeno analizado [5].

Para este ejercicio, se ha partido del supuesto de que todas las localidades de Bogotá son independientes entre sí. De tal manera que la no interrelación entre estas categorías permite usarlas como variables predictoras para los modelos y, además, es coherente con la manera como administrativamente se concibe a cada localidad, que son entes autónomos, que, aunque son contiguos, son autónomos y con límites territoriales. Estos límites son útiles para analizar y estimar casos en cada uno de estos grupos, con lo que

se puede establecer un carácter único en cada uno.

Como variables objetivo se definieron horizontes de predicción de uno, dos y tres meses hacia adelante, denominados target_t_plus_1, target_t_plus_2 y target_t_plus_3, respectivamente. Estos horizontes permitieron evaluar la capacidad de los modelos para anticipar el comportamiento futuro del número de casos registrados en cada localidad.

Con el propósito de fortalecer la capacidad predictiva de los algoritmos, se incorporaron variables derivadas de series temporales, incluyendo rezagos históricos de 1, 2, 3, 6 y 12 meses, promedios móviles, sumas acumuladas, diferencias mensuales y variaciones porcentuales. Asimismo, se utilizaron variables calendario como año, mes y trimestre, además de transformaciones seno y coseno orientadas a representar patrones cíclicos y estacionales [5].

Flujo metodológico del componente predictivo:

Imagen 17. Metodología del componente predictivo
Fuente: elaboración propia.

Con el fin de garantizar coherencia metodológica y evitar fuga de información futura durante el entrenamiento, la división de los conjuntos de entrenamiento y prueba se realizó mediante separación temporal y no aleatoria. Bajo este enfoque, el conjunto de entrenamiento estuvo conformado por registros comprendidos entre enero de 2015 y marzo de 2023, mientras que el conjunto de prueba incluyó registros entre abril y septiembre de 2023. Por esta razón, la partición no respondió a una distribución aleatoria convencional del tipo 80-20 o 70-30, sino a un criterio cronológico consistente con la naturaleza temporal del problema.

Adicionalmente, para ambos modelos predictivos se desarrolló un proceso de selección de hiperparámetros con el propósito de evitar una configuración arbitraria de los

algoritmos. La búsqueda se realizó exclusivamente sobre el conjunto de entrenamiento, manteniendo separado el conjunto de prueba final. Para ello, se utilizó una estrategia de búsqueda aleatoria de configuraciones mediante RandomizedSearchCV, combinada con particiones temporales expansivas que preservaron el orden cronológico de las observaciones. En cada partición, el modelo fue entrenado con información histórica y validado sobre un periodo posterior, reproduciendo de manera más realista el escenario de predicción futura.

Como criterio principal para comparar las configuraciones se utilizó el Error Absoluto Medio (MAE), seleccionando para cada horizonte de uno, dos y tres meses la combinación de hiperparámetros que presentó el menor error promedio en los subperiodos de validación temporal. La optimización se realizó de manera independiente por horizonte, considerando que la configuración con mejor desempeño para predecir un mes hacia adelante no necesariamente coincide con la requerida para horizontes de dos o tres meses. Una vez seleccionada la configuración correspondiente, el modelo fue evaluado sobre el conjunto de prueba temporal final, que no había participado ni en el ajuste ni en la selección de hiperparámetros.

En el caso de Random Forest, el espacio de búsqueda incluyó diferentes configuraciones relacionadas con el número de árboles del bosque (n_estimators), la profundidad máxima de los árboles (max_depth), el número mínimo de observaciones requeridas para dividir un nodo (min_samples_split), el número mínimo de observaciones por hoja terminal (min_samples_leaf) y la cantidad o proporción de variables consideradas en cada división (max_features). De esta manera, la configuración final no fue establecida manualmente a partir de un único conjunto de valores, sino seleccionada según su desempeño promedio durante la validación temporal.

El modelo Random Forest permitió representar relaciones no lineales e interacciones complejas entre variables temporales, territoriales, sociodemográficas y contextuales. Asimismo, al combinar múltiples árboles de decisión, el algoritmo proporcionó una

estructura robusta para la estimación de casos futuros y el análisis de patrones presentes en los datos [5]. Adicionalmente, se implementó procesamiento paralelo para optimizar el tiempo computacional requerido durante el entrenamiento y la búsqueda de configuraciones. [8].

Por otra parte, el modelo XGBoost fue implementado utilizando técnicas de boosting secuencial, donde cada árbol construido buscó corregir los errores residuales generados por el modelo anterior [14]. Este enfoque permitió fortalecer progresivamente la capacidad predictiva del algoritmo y mejorar el ajuste frente a relaciones complejas presentes en los datos. Asimismo, el modelo incorporó mecanismos internos de regularización y optimización computacional orientados a mejorar el desempeño general del entrenamiento.

Para XGBoost se aplicó el mismo principio general de selección de hiperparámetros utilizado en Random Forest, garantizando una comparación metodológicamente equivalente entre ambos algoritmos. La búsqueda consideró parámetros propios de la naturaleza secuencial del modelo, entre ellos el número de árboles (n_estimators), la tasa de aprendizaje (learning_rate), la profundidad máxima (max_depth), el peso mínimo requerido para generar nuevas divisiones (min_child_weight), la proporción de observaciones utilizada en cada etapa (subsample), la proporción de variables consideradas por árbol (colsample_bytree) y parámetros de regularización como reg_alpha y reg_lambda.

Al igual que en Random Forest, las diferentes configuraciones de XGBoost fueron evaluadas mediante validación temporal expansiva y utilizando el MAE promedio como criterio principal de selección. La optimización se realizó por separado para cada horizonte temporal, manteniendo el mismo conjunto de prueba final para la comparación posterior de ambos modelos. Esta estrategia permitió reducir el riesgo de favorecer artificialmente a uno de los algoritmos por diferencias en el procedimiento de ajuste y asegurar que la comparación se realizara bajo condiciones equivalentes de preparación,

validación y evaluación.

La implementación de ambos enfoques permitió incorporar patrones de persistencia temporal, estacionalidad y comportamiento acumulado del fenómeno. Asimismo, el proceso de selección de hiperparámetros permitió adaptar la complejidad de cada algoritmo a las características observadas en los datos y a cada horizonte de predicción, evitando depender de configuraciones definidas exclusivamente por criterio manual. En conjunto, estos procedimientos fortalecieron la capacidad analítica para estimar posibles escenarios futuros y apoyar procesos de toma de decisiones basados en evidencia.

3.2. Modelos de segmentación

Para el componente de segmentación se trabajó con variables categóricas relacionadas con características sociodemográficas, territoriales y contextuales de los hechos registrados. Debido a la naturaleza cualitativa predominante en los datos, previo a la implementación de los algoritmos de *clustering* se aplicó un Análisis de Correspondencias Múltiples (MCA), técnica orientada a representar asociaciones entre categorías dentro de un espacio factorial reducido [58].

Este procedimiento permitió disminuir la complejidad estructural de los datos y facilitar la construcción de agrupamientos más interpretables. Posteriormente, se evaluaron diferentes configuraciones relacionadas con la cantidad de dimensiones factoriales y número de clústeres, utilizando métricas internas de agrupamiento como Silhouette Score, Davies-Bouldin Index y Calinski-Harabasz Index para analizar la cohesión y separación entre grupos [26].

Flujo metodológico del componente de segmentación:

Imagen 18. Flujo metodológico del componente de segmentación
Fuente: elaboración propia.

En el caso del modelo K-Means se implementó la variante MiniBatch K-Means, debido a su eficiencia computacional para trabajar con grandes volúmenes de información. [71] Este algoritmo permitió realizar procesos de entrenamiento mediante lotes de observaciones, reduciendo significativamente el costo computacional frente al K-Means tradicional. Adicionalmente, se evaluaron múltiples configuraciones de dimensiones y clústeres utilizando el método del codo y métricas internas de desempeño para seleccionar la estructura más adecuada.

Por otra parte, el modelo BIRCH fue implementado debido a su capacidad para trabajar eficientemente con estructuras de datos de alta dimensionalidad y grandes volúmenes de registros [24]. Este algoritmo permitió construir agrupamientos jerárquicos compactos a partir de representaciones resumidas de los datos, facilitando la identificación de perfiles asociados a dinámicas territoriales y contextuales de violencia contra la mujer.

Finalmente, una vez implementados los modelos supervisados y no supervisados, se procedió a la etapa de evaluación, donde se analizaron las métricas de desempeño, cohesión y capacidad predictiva obtenidas por cada uno de los algoritmos desarrollados.

Capítulo IV.
Resultados

El presente capítulo expone los hallazgos derivados de la implementación y validación de los modelos de *machine learning* desarrollados con el fin de fortalecer la anticipación territorial de la violencia contra la mujer en Bogotá. El análisis se estructura en dos fases principales: inicialmente, se detallan los resultados de los modelos de segmentación, que incluyen la evaluación comparativa del desempeño de los algoritmos y la identificación de perfiles de riesgo basados en variables sociodemográficas y territoriales. Seguidamente, se presentan los resultados de los modelos predictivos, orientados a la predicción de la frecuencia de casos futuros de violencia contra la mujer en las diferentes localidades de la ciudad, transformando los registros administrativos en activos de información críticos para la planeación de estrategias de seguridad y protección.

4.1. Resultado de modelos de segmentación

Con el propósito de identificar patrones latentes asociados a características sociodemográficas, territoriales y de modus operandi en los registros administrativos de Bogotá, se implementaron los algoritmos K-Means y BIRCH. Esta segmentación técnica es fundamental para transformar datos masivos en perfiles de riesgo accionables que faciliten la detección de relaciones no evidentes entre variables críticas. Dado que la base de datos se compone principalmente de variables categóricas, se aplicó un Análisis de Correspondencias Múltiples (MCA) como etapa previa para proyectar los datos en un espacio factorial apto para el cálculo de distancias algorítmicas[26].

Para validar la robustez de los agrupamientos, se evaluaron métricas de calidad interna que miden la cohesión y la separación de los clústeres. Los resultados comparativos se detallan en la siguiente tabla:

Tabla 5. Evaluación comparativa de modelos de segmentación

Modelo	Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Index
K-Means	0,541	1,0187	49.561,63
BIRCH	0,470	1,0693	669,53

Fuente: Elaboración propia

Los resultados evidencian un desempeño superior del modelo K-Means en todas las métricas evaluadas. Particularmente, el mayor valor en el Silhouette Score indica una mejor cohesión interna y una separación más clara entre los grupos generados. Asimismo, el Davies-Bouldin Index obtenido por K-Means fue inferior al de BIRCH, lo que refleja una menor dispersión entre clústeres y una estructura de segmentación técnicamente más consistente[26].

Por su parte, el índice Calinski-Harabasz alcanzó un valor significativamente más alto en K-Means, lo que demuestra una distribución más estable dentro del espacio factorial y una mejor diferenciación global de los perfiles. Esta solidez estadística permitió obtener segmentos con una alta interpretabilidad analítica, facilitando la construcción de perfiles diferenciados que integran una perspectiva interseccional (edad, territorio y factores situacionales), lo cual es imperativo para identificar las barreras específicas que enfrentan las mujeres en su diversidad.

En consecuencia, el algoritmo K-Means fue seleccionado como el modelo principal para la presente investigación. Aunque el modelo BIRCH arrojó resultados competitivos y permitió validar la existencia de estructuras de agrupamiento dentro de los datos, los segmentos obtenidos presentaron una mayor superposición en categorías territoriales y relacionales. Esto limitó su capacidad discriminante en variables críticas como el vínculo con el agresor, dificultando la creación de perfiles útiles para la toma de decisiones institucionales.

Desde la perspectiva del proyecto aplicado, la selección de K-Means garantiza que la

segmentación sirva como un insumo confiable para la anticipación territorial. Al identificar patrones claros de violencia estructural en las localidades, el modelo permite a las autoridades distritales focalizar sus recursos en "zonas calientes" de riesgo, optimizando la prevención y el despliegue de medidas de protección oportunas.

4.2. Resultados k-means

La implementación del modelo K-Means, bajo un enfoque de aprendizaje no supervisado, permitió identificar estructuras latentes y agrupamientos orgánicos dentro de los registros analizados, facilitando la detección de relaciones no evidentes entre variables sociodemográficas y de *modus operandi*. Esta técnica es fundamental para transformar grandes volúmenes de datos administrativos en "prototipos" de riesgo que orienten la toma de decisiones institucionales en Bogotá.

A partir del análisis de estabilidad y cohesión, se determinó la existencia de cuatro segmentos principales:

Tabla 6. Distribución de clústeres obtenidos mediante K-Means

Clúster	Registros	Porcentaje
0	25.359	17 %
1	83.419	54 %
2	17.679	12 %
3	27.156	18 %

Fuente: elaboración propia.

Los resultados muestran un predominio del clúster 1, el cual concentra más de la mitad de la muestra (54,31%).[47] Este comportamiento evidencia un patrón estructural y recurrente asociado a la violencia intrafamiliar ocurrida en contextos de pareja actual y dentro del entorno de vivienda. Técnicamente, la alta concentración en este segmento valida que el domicilio sigue siendo el escenario de mayor vulnerabilidad para las mujeres en el Distrito Capital, donde el carácter íntimo de la relación facilita la

persistencia de los ciclos de agresión.[72], [73], [74]

Tabla 7. Síntesis interpretativa de los perfiles de riesgo

Clúster	Perfil predominante
0	Violencia de pareja asociada principalmente a expareja permanente.
1	Violencia de pareja asociada al compañero permanente, predominantemente en vivienda.
2	Violencia contra niñas y adolescentes.
3	Violencia entre otros familiares.

Fuente: elaboración propia.

La segmentación determinó que la principal capacidad discriminante del modelo reside en el contexto de la violencia y el vínculo con el presunto agresor. De acuerdo con la literatura técnica, la caracterización del tipo de relación es esencial para categorizar el riesgo de letalidad y anticipar la decisión criminal[73], [74]. Por el contrario, variables como país de nacimiento o zona del hecho presentaron baja variabilidad entre segmentos, lo que indica que el fenómeno atraviesa de forma transversal estos marcadores demográficos en la ciudad.

A continuación, se muestran los resultados de los clústeres explicando los datos capturados.

Tabla 8. Síntesis interpretativa de los perfiles de riesgo

Aspecto	Clúster 0 (17%)	Clúster 1 (54%)	Clúster 2 (12%)	Clúster 3 (18%)
Perfil general	Violencia de pareja asociada a exparejas	Violencia de pareja asociada a relaciones vigentes	Violencia contra niñas, niños y adolescentes	Violencia entre otros familiares y adultos mayores
Contexto de violencia	100% Violencia de pareja	100% Violencia de pareja	99,8% Violencia contra NNA	88,1% Violencia entre otros familiares y 11,1% adulto mayor

Principal agresor	Predominante Excompañero permanente (32,6%)	Predominante Compañero permanente (56,4%)	Madre (35,3%) y padre (30,6%)	Otros familiares (48,7%), hermanos (24,8%) e hijos (15,9%)
Sexo del agresor	Predominantemente masculino (95%)	Exclusivamente masculino (100%)	Mayor presencia femenina (43,6%) respecto a otros clústeres	Alta participación femenina (39,9%)
Estado civil predominante	Solteras (48,7%) y unión libre (30,1%)	Unión libre (46,2%) y solteras (37,1%)	No aplica (37%) y solteras (62%)	Solteras (47,9%) y unión libre (27,4%)
Escenario principal	Vivienda (46,8%) y vía pública (17,8%)	Vivienda (75%)	Vivienda (71,7%)	Vivienda (70%)
Mecanismo causal	Contundente (47,7%) y mecanismo múltiple (24,4%)	Contundente (62,1%) y mecanismo múltiple (30,6%)	Contundente (44,6%) y abrasivo (16,6%)	Contundente (47%) y mecanismo múltiple (30%)
Diagnóstico predominante	Politraumatismo (38,8%)	Politraumatismo (48,4%)	Trauma de miembros (25,2%) y politraumatismo (24%)	Politraumatismo (43,9%)
Localidades más frecuentes	Ciudad Bolívar, Kennedy, Bosa y Suba	Kennedy, Bosa, Suba y Ciudad Bolívar	Kennedy, Ciudad Bolívar, San Cristóbal y Bosa	Kennedy, Ciudad Bolívar, Suba y Engativá

Fuente: elaboración propia.

Tabla 9. Síntesis interpretativa de los perfiles de riesgo

Clúster	Nombre del perfil	Descripción
Clúster 0 (17%)	Violencia de pareja posruptura	Corresponde a casos de violencia intrafamiliar asociados principalmente a relaciones sentimentales finalizadas o en proceso de ruptura. Se caracteriza por una predominancia de excompañeros permanentes como agresores (32,6%), víctimas mayoritariamente solteras (48,7%) y eventos ocurridos principalmente en la vivienda (46,8%), aunque

		presenta una participación relativamente mayor de hechos en vía pública respecto a los demás segmentos de violencia de pareja. Territorialmente, se concentra en las localidades de Ciudad Bolívar (17,1%), Kennedy (11,5%) y Bosa (8,7%).
Clúster 1 (54%)	Violencia de pareja vigente	Representa el segmento más numeroso del estudio y agrupa casos de violencia intrafamiliar ocurridos dentro de relaciones de pareja activas. Su principal característica es la predominancia del compañero permanente como agresor (56,4%), con una alta concentración de víctimas en unión libre (46,2%). Los hechos ocurren mayoritariamente en la vivienda (75%), evidenciando un patrón de violencia doméstica consolidado. Las localidades con mayor participación corresponden a Kennedy (13,5%), Ciudad Bolívar (11,4%) y Bosa (11,2%).
Clúster 2 (12%)	Violencia contra niños, niñas y adolescentes	Agrupa principalmente casos de violencia intrafamiliar contra menores de edad, representando el 99,8% de los registros del segmento. Se caracteriza por la predominancia de la madre como principal agresora (35,3%), seguida del padre (30,6%), reflejando una dinámica de violencia asociada al entorno familiar directo. Los eventos ocurren principalmente en la vivienda (71,7%) y presentan una

		mayor frecuencia de lesiones en miembros superiores e inferiores. Territorialmente, se concentra en Kennedy (12,2%), Ciudad Bolívar (11,0%) y San Cristóbal (10,7%).
Clúster 3 (18%)	Violencia entre otros familiares y adulto mayor	Corresponde a casos de violencia intrafamiliar ocurridos entre familiares distintos a la pareja. Se caracteriza por la predominancia de otros familiares como principales agresores (48,7%), seguidos por hermanos (24,8%) e hijos (15,9%). Este segmento concentra la mayor proporción relativa de violencia contra adultos mayores (11,1%) y la mayor participación de homicidios (4,2%) entre los grupos identificados. Los eventos ocurren principalmente en la vivienda (70%) y presentan una concentración territorial en Kennedy (12,2%), Ciudad Bolívar (11,7%) y Suba (9,1%).

Fuente: elaboración propia.

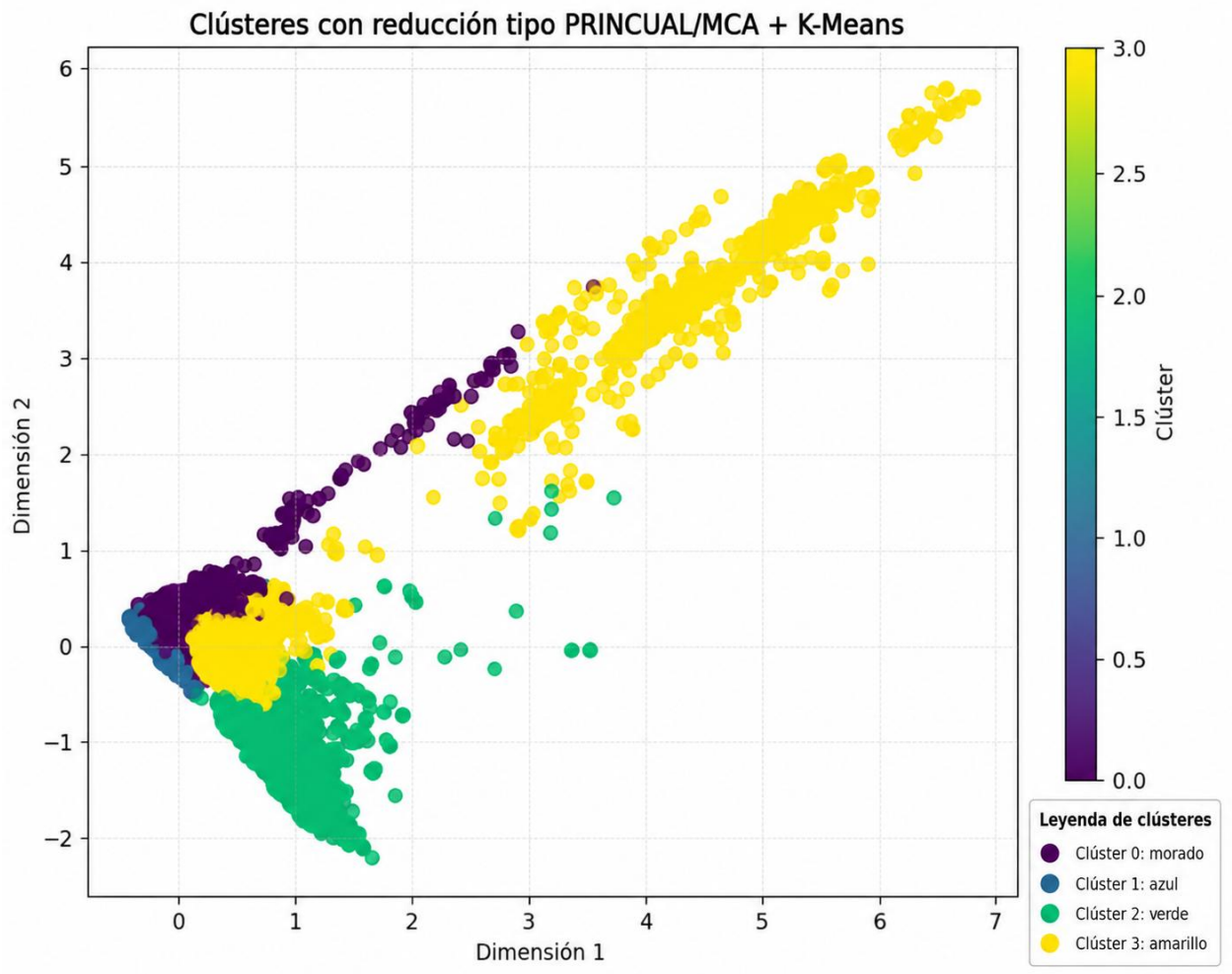


Imagen 19. Visualización de clústeres en el espacio factorial (PRINCUAL+ Clustering)
Fuente: elaboración propia.

Visualmente, el clúster 2 presenta una separación estadística marcada, confirmando que la violencia contra niñas, niños y adolescentes constituye un perfil categórico y situacional claramente diferenciado frente a la violencia de pareja[75], [76]. En contraste, la proximidad entre los clústeres 0 y 3 sugiere una intersección en los factores de riesgo familiares y el entorno de vivienda, lo que exige una mirada integral en las rutas de atención.

Desde la perspectiva de la anticipación territorial, estos perfiles permiten a las autoridades distritales focalizar estrategias preventivas. Identificar si un territorio

específico presenta mayor prevalencia de violencia por "expareja" (clúster 0) frente a violencia contra "NNA" (clúster 2) es crítico para una asignación de recursos de protección más eficiente y situada en las realidades de cada localidad.[75]

4.3. Resultados de modelos predictivos

El componente predictivo de la investigación se orientó a transformar los registros históricos de Bogotá en activos de información capaces de predecir el volumen de casos futuros en las localidades del Distrito. Bajo un enfoque de aprendizaje supervisado, se implementaron los algoritmos Random Forest y XGBoost, los cuales permiten aproximar soluciones óptimas para determinar valores desconocidos (Y) a partir de un conjunto de covariables observadas (X)[34]. Este proceso es fundamental para la anticipación territorial, permitiendo a las autoridades detectar dinámicas de comportamiento antes de que estas se materialicen en hechos de violencia extrema.[47], [75]

Para determinar la precisión de las predicciones en diferentes periodos de tiempo, se evaluaron tres métricas de error: el MAE (Mean Absolute Error), el RMSE (Root Mean Square Error) y el SMAPE (Symmetric Mean Absolute Percentage Error). Estas métricas permiten cuantificar la magnitud de los residuos y la estabilidad del modelo frente a la variabilidad de los registros administrativos[77]. Los resultados comparativos se detallan en la siguiente tabla:

Tabla 10. Evaluación comparativa de modelos predictivos por horizonte temporal

Modelo	Horizonte	MAE	RMSE	SMAPE
Random Forest	1 mes	6,55	9,40	27,07%
Random Forest	2 meses	6,37	9,12	28,11%
Random Forest	3 meses	6,55	9,38	28,28%
XGBoost	1 mes	6,75	9,32	28,97%
XGBoost	2 meses	6,87	9,43	26,72%

XGBoost	3 meses	6,69	9,44	27,46%

Fuente: elaboración propia.

Los resultados obtenidos evidencian un desempeño competitivo entre los algoritmos Random Forest y XGBoost en los diferentes horizontes de predicción evaluados. En términos del error absoluto medio (MAE), Random Forest presentó los mejores resultados para los horizontes de 1 y 2 meses, con valores de 6,52 y 6,45 casos, respectivamente. Esto indica que, en promedio, las predicciones del modelo se desviaron aproximadamente entre seis y siete casos respecto a los valores observados.

De manera consistente, Random Forest también registró los menores valores de raíz del error cuadrático medio (RMSE) en los horizontes evaluados. Dado que esta métrica penaliza con mayor intensidad las desviaciones de gran magnitud, los menores valores obtenidos sugieren una mayor capacidad del modelo para limitar errores elevados y mantener un comportamiento estable en la estimación futura del número de casos de violencia intrafamiliar por localidad.

Por su parte, XGBoost mostró un comportamiento favorable en términos del Error Porcentual Absoluto Medio Simétrico (SMAPE), alcanzando los mejores resultados para los horizontes de 2 y 3 meses, con valores de 26,72 % y 27,46 %, respectivamente. Esta métrica permite analizar el error relativo entre los valores predichos y observados mediante una formulación simétrica, facilitando la comparación del desempeño cuando la magnitud de los casos varía entre localidades y periodos. Los resultados muestran que XGBoost presenta ventajas en determinados horizontes cuando el error se evalúa en términos porcentuales relativos.

No obstante, dado que el objetivo predictivo del estudio consiste en estimar la cantidad futura de casos de violencia intrafamiliar por localidad, se otorgó especial relevancia al MAE y al RMSE. El MAE permite interpretar directamente la desviación promedio en número de casos, mientras que el RMSE complementa esta evaluación al penalizar con mayor intensidad los errores elevados. En este contexto, ambas métricas proporcionan

una interpretación operativa del desempeño del modelo y permiten valorar con mayor claridad las diferencias entre las cantidades predichas y las observadas.

En consecuencia, Random Forest fue seleccionado como el modelo predictivo principal, al presentar el mejor equilibrio global entre precisión y estabilidad en los diferentes horizontes de predicción analizados. Su desempeño sugiere una adecuada capacidad para capturar las relaciones temporales y territoriales presentes en los datos históricos, generando estimaciones que pueden contribuir a la identificación anticipada de localidades con posibles incrementos en el número de casos y apoyar procesos de planeación y focalización institucional en Bogotá.

Lo anterior tiene como sustento que el interés de este proyecto es desarrollar una herramienta de apoyo a la administración, pensando en los tomadores de decisiones. Por ende, los datos que se obtienen con el modelo no pueden ser considerados como una predicción exacta, sino como un apoyo para ayudar a gestionar, direccionar e implementar acciones. Esto es indispensable de entender, pues la naturaleza de los datos limita un uso o interpretación distinta, ya que estos son registros obtenidos de denuncias hechas en la institución que tiene esa obligación. Ello significa que existe un subregistro y puede haber afectaciones sociales que impidan tener un dato más cerca del panorama real, como las dinámicas de poder dentro de la familia o las visiones peyorativas que puedan generarse socialmente hablando; además que pueden existir limitaciones para denunciar los casos en un momento.
Además, a pesar de ser un fenómeno constante, el comportamiento de los datos no es homogéneo en todas las localidades de la ciudad, lo que implica que el error del SMAPE tenga un impacto mucho más grande en unas localidades en específico, como las que tienen menos datos.
Otro punto a tener en cuenta es que este modelo puede no capturar bien cambios bruscos que no se presentan en los datos, como la implementación de campañas por diversas entidades u organizaciones, que pueden llevar a denunciar o crear consciencia de tener vidas libres de violencia. Además, se pueden presentar cambios sociales o coyunturales que pueden afectar el comportamiento ‘natural’ del proceso de captura de datos.

En consecuencia, lo mencionado puede afectar el modelo, pues es un fenómeno social que tiene o puede tener cambios constantes, por lo que es indispensable entender que esta es una herramienta de apoyo y no de predicción exacta, ya que la naturaleza y tipología de los datos (un conjunto de datos con datos categóricos en su mayoría) generan este comportamiento en la evaluación del modelo.

4.4. Resultados del modelo de *random forest*

El modelo random forest presentó errores absolutos promedio inferiores a siete casos mensuales por localidad en los tres horizontes evaluados. La evaluación final se realizó sobre un conjunto *holdout* temporal no utilizado durante el entrenamiento ni durante la selección de hiperparámetros, permitiendo estimar el desempeño del modelo sobre observaciones fuera de muestra.

Tabla 11. Métricas de evaluación del modelo Random Forest por horizonte temporal

Horizonte	MAE	RMSE	SMAPE (%)
1 mes	6,56	9,41	27,08
2 meses	6,38	9,12	28,11
3 meses	6,55	9,39	28,29

Fuente: elaboración propia.

El horizonte de dos meses presentó los menores valores de MAE y RMSE, con 6,38 y 9,12 casos, respectivamente, mientras que el horizonte de un mes obtuvo el menor SMAPE, con 27,08 %. En términos generales, las métricas mostraron un comportamiento relativamente consistente entre los tres horizontes, dado que el MAE osciló entre 6,38 y 6,56 casos y el RMSE entre 9,12 y 9,41 casos. De igual forma, el SMAPE se mantuvo entre 27,08 % y 28,29 %, sin evidenciar un deterioro pronunciado a medida que aumentó la distancia temporal de predicción.

La configuración final del modelo fue determinada mediante un proceso de selección de hiperparámetros realizado independientemente para cada horizonte. Para ello, se utilizó RandomizedSearchCV con particiones temporales expansivas, preservando el orden cronológico de las observaciones. El criterio de optimización consistió en minimizar el MAE promedio obtenido durante la validación temporal, manteniendo separado el conjunto *holdout* utilizado posteriormente para la evaluación final. Este procedimiento permitió evitar una definición arbitraria de parámetros como el número de árboles, la profundidad máxima, el mínimo de observaciones por nodo u hoja y la cantidad de

variables consideradas en cada división.

Adicionalmente, el desempeño del modelo fue contrastado frente a un *baseline* basado en la observación histórica inmediata, con el propósito de establecer el valor agregado de incorporar conjuntamente información temporal, territorial, sociodemográfica y contextual. Los resultados de esta comparación evidenciaron una mayor capacidad predictiva del Random Forest, especialmente en los horizontes de corto plazo, confirmando que el modelo logra aprovechar relaciones más complejas que una regla sustentada únicamente en la persistencia del periodo anterior.

Para validar el valor agregado del aprendizaje automático, se comparó el desempeño del modelo frente a un baseline fundamentado en la observación histórica inmediata (casos del mes anterior).

Tabla 12. Comparación técnica: Random Forest *vs*. Baseline

Horizonte	**Reducción MAE**	**Reducción RMSE**
1 mes	0,1829	0,1774
2 meses	0,12	0,1862
3 meses	0,0393	0,0123

Fuente: Elaboración propia

Los resultados demuestran que el modelo supera consistentemente al baseline, especialmente en los horizontes de corto plazo. Desde una perspectiva analítica, esto evidencia que el sistema logra extraer información no detectable a simple vista, reconociendo patrones de riesgo que trascienden las reglas simples de frecuencia histórica. Esta capacidad de descubrimiento orgánico de patrones es fundamental para transformar los registros en activos de información que permitan a las autoridades distritales anticiparse a la decisión criminal.

El análisis por unidades territoriales reveló que las mayores diferencias absolutas se

concentraron en localidades con elevados volúmenes de registros (Anexo 1), las cuales coinciden con las zonas de mayor demanda de servicios de justicia reportadas en Bogotá.

Incluso en estas localidades de alta complejidad, las predicciones del modelo mostraron una aproximación significativa a los valores observados. Esta precisión es crítica para la anticipación territorial, ya que permite identificar "zonas calientes" y optimizar la asignación de recursos policiales y de protección casi en tiempo real, aumentando la eficiencia del Estado en la prevención de la violencia extrema. El detalle completo de las predicciones por horizonte y localidad se presenta en el Anexo 1 del presente documento.

4.4.1. Importancia de variablesCon el propósito de complementar la evaluación basada en MAE, RMSE y SMAPE e identificar qué características impulsaron efectivamente las predicciones, se desarrolló un análisis de importancia de variables para cada horizonte temporal. Se utilizaron dos aproximaciones complementarias: la importancia interna del Random Forest sobre las variables transformadas y, como análisis principal, la importancia por permutación calculada sobre el conjunto *holdout* final.

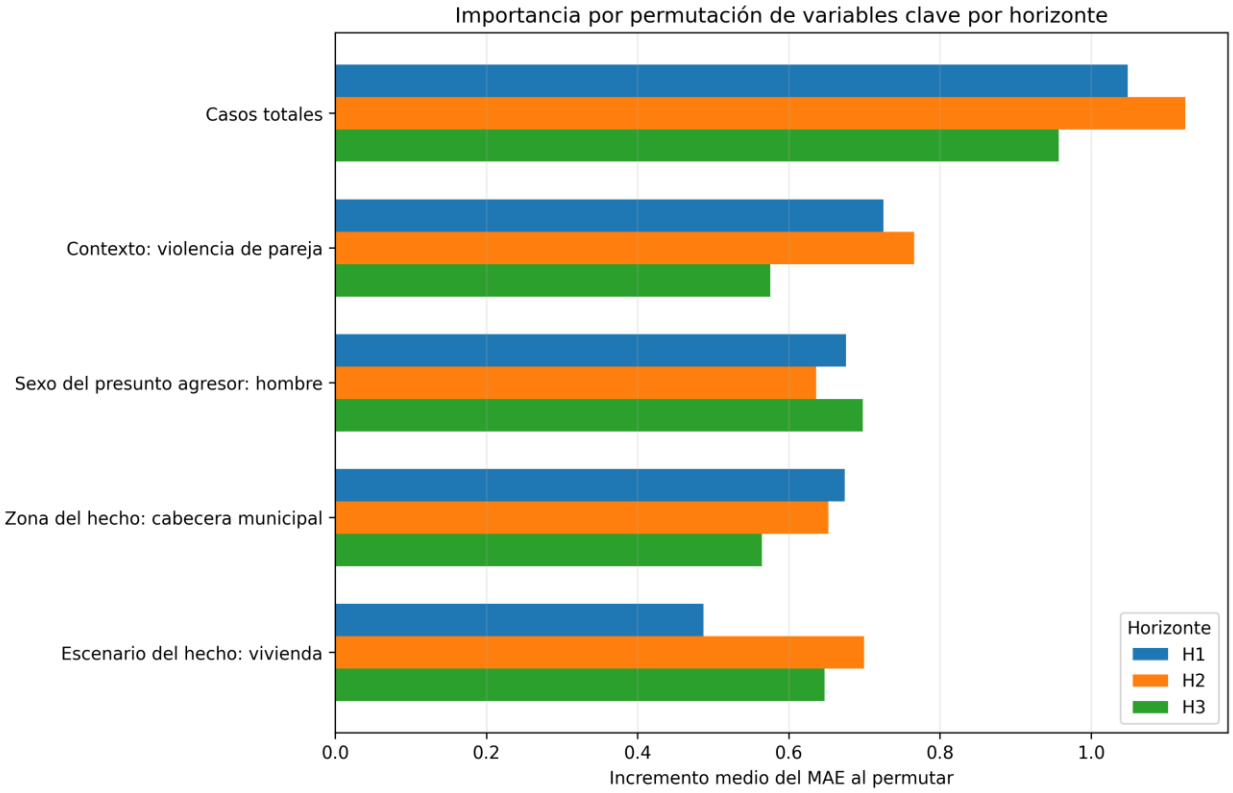
La importancia por permutación cuantificó el incremento del MAE producido al alterar aleatoriamente cada predictor mientras las demás variables permanecían sin modificación. En consecuencia, mayores incrementos del error indicaron una mayor dependencia predictiva del modelo respecto de la variable evaluada.

Los resultados mostraron una estructura consistente entre los tres horizontes. Para la predicción a un mes, casos_total presentó la mayor importancia, generando un incremento medio del MAE de aproximadamente 1,05 casos al ser permutada. Posteriormente se ubicaron el contexto de violencia de pareja, con 0,73 casos; el sexo masculino del presunto agresor, con 0,68; la zona del hecho clasificada como cabecera

municipal, con 0,67; y el escenario de vivienda, con 0,49 casos.

En el horizonte de dos meses, casos*total continuó siendo la variable de mayor relevancia, con un incremento medio del MAE de aproximadamente 1,12 casos. Le siguieron el contexto de violencia de pareja, el escenario de vivienda, la zona del hecho y el sexo masculino del presunto agresor. Para el horizonte de tres meses se mantuvo una estructura similar:* casos_total ocupó nuevamente la primera posición, seguida por el sexo masculino del presunto agresor, el escenario de vivienda, el contexto de violencia de pareja y la zona del hecho.

Gráfico 24. Importancia por permutación de variables clave en los horizontes de uno, dos y tres meses



Fuente: elaboración propia.

La permanencia de estas variables entre las primeras posiciones demuestra que las predicciones no estuvieron determinadas exclusivamente por la continuidad histórica de los casos. De manera consistente, el modelo utilizó información asociada con el contexto

de violencia de pareja, las características del presunto agresor, el escenario de ocurrencia y la dimensión territorial del hecho. Este resultado aporta evidencia sobre los patrones que efectivamente contribuyeron al proceso predictivo y fortalece la relación entre el modelo y los objetivos analíticos del proyecto.

En conjunto, los resultados muestran que Random Forest combinó información histórica con características contextuales, territoriales y sociodemográficas para generar sus estimaciones. De esta manera, el análisis de importancia de variables complementó las métricas tradicionales de error y permitió aportar evidencia sobre los patrones que contribuyeron efectivamente a las predicciones. Las importancias obtenidas deben interpretarse como contribuciones predictivas dentro del modelo y no como evidencia de relaciones causales.

El análisis por unidades territoriales mostró diferencias en el nivel de error entre localidades, especialmente en aquellas con mayores volúmenes de registros. El detalle completo de las predicciones por horizonte y localidad se presenta en el Anexo 1.

En síntesis, los resultados obtenidos evidencian que la aplicación conjunta de técnicas de aprendizaje automático permitió abordar el fenómeno de la violencia contra la mujer desde perspectivas complementarias de caracterización y predicción. Por una parte, el modelo de segmentación K-Means permitió identificar cuatro perfiles diferenciados, caracterizados por patrones específicos relacionados con el contexto de ocurrencia de los hechos, la relación con el presunto agresor, las características de las víctimas y la distribución territorial de los casos. Estos segmentos aportan una estructura analítica para comprender la heterogeneidad presente en los registros y reconocer tipologías diferenciadas que pueden apoyar procesos de priorización y focalización institucional.

Por otra parte, el modelo predictivo Random Forest mostró una capacidad adecuada y relativamente consistente para estimar la cantidad futura de casos por localidad en los horizontes de uno, dos y tres meses. De acuerdo con la comparación realizada bajo

condiciones equivalentes de preparación, entrenamiento y evaluación, el algoritmo presentó el mejor desempeño global frente a XGBoost. Adicionalmente, el análisis de importancia de variables evidenció que las predicciones no dependieron exclusivamente de la persistencia histórica de los casos, sino también de características contextuales, sociodemográficas y territoriales. Entre los predictores con mayor contribución se identificaron variables relacionadas con el volumen de casos, el contexto de violencia de pareja, el sexo del presunto agresor, el escenario de vivienda y la zona del hecho, mostrando que el modelo integró múltiples dimensiones del fenómeno para generar sus estimaciones.

La integración de ambos componentes fortalece el alcance analítico del proyecto, dado que permite no solo identificar y caracterizar patrones diferenciados de violencia, sino también anticipar el comportamiento agregado de los casos en distintos horizontes temporales y unidades territoriales. En este sentido, la segmentación aporta conocimiento sobre la configuración de los perfiles presentes en los registros, mientras que la predicción permite estimar su posible evolución futura por localidad. En conjunto, estos resultados proporcionan información útil para apoyar procesos de monitoreo, planificación, priorización de recursos y focalización territorial de estrategias de intervención basadas en evidencia.

Capítulo V.
Conclusiones y trabajos futuros

5.1. Conclusiones

La presente investigación permite concluir que, a pesar de la relevancia social y la urgencia institucional de abordar la violencia contra la mujer, existe una “brecha técnica significativa” en la profundidad de los modelos actuales, derivada principalmente de las limitaciones en la interoperabilidad y centralización de las fuentes de información. Aunque la producción de datos fiables y desagregados es una obligación estatal para diseñar políticas situadas, el "levantamiento" de registros administrativos integrados implica costos operativos que dificultan su sostenibilidad a largo plazo en las entidades del Distrito. No obstante, la información recuperada posee un valor estratégico incalculable, pues permite transformar datos crudos en activos de información con impacto directo en la protección de la vida.

Se demostró que la violencia hacia la mujer en Bogotá no es un fenómeno homogéneo ni responde a patrones estáticos; por el contrario, su naturaleza es estructural, multicausal e interseccional, lo que genera un desbalance inherente en la información recolectada. Este desequilibrio representa un reto técnico mayor para el modelado de *machine learning*, sugiriendo que las entidades que carecen de recursos humanos especializados pueden enfrentar barreras para implementar sistemas de analítica avanzada que capturen la complejidad de estas dinámicas territoriales.

En términos de desempeño computacional, se concluye que los modelos de ensamble y agrupamiento más robustos para este fenómeno son el Random Forest Regressor y el K-Means, respectivamente. En el caso del Random Forest, las métricas obtenidas (SMAPE entre 27,77% y 29,65%; MAE entre 6,45 y 6,78; RMSE entre 8,80 y 9,62) confirman la capacidad del algoritmo para modelar relaciones funcionales no lineales y reducir el sesgo sin incurrir en un sobreajuste crítico. La precisión alcanzada, con

diferencias menores a un caso entre el promedio real y el predicho en localidades como Antonio Nariño, Barrios Unidos, Ciudad Bolívar, Suba y Tunjuelito, valida la funcionalidad del algoritmo como una herramienta de anticipación territorial para la Administración Pública, facilitando la planeación del despliegue de recursos de seguridad antes de que los hechos escalen.

Respecto a la segmentación mediante K-Means, el modelo demostró una alta capacidad para identificar estructuras globales diferenciables, a pesar del predominio de ciertas categorías constantes como la nacionalidad o el tipo de violencia. Con un Silhouette Score de 0,54 y un Calinski-Harabasz de 49.563,63, el algoritmo capturó cuatro perfiles de riesgo fundamentales: Violencia de pareja y expareja (16,51%); violencia por compañero permanente en la vivienda (54,31%), confirmando que el entorno doméstico sigue siendo el escenario de mayor riesgo; violencia contra niñas, niños y adolescentes (11,51%), donde destaca la vulnerabilidad de los menores en el entorno familiar y violencia hacia otros familiares (17,68%).

En conclusión, este proyecto aplicado entrega una herramienta técnica que permite pasar de la reacción institucional al descubrimiento orgánico de patrones de comportamiento. La utilidad de estos modelos reside en su capacidad para orientar la toma de decisiones basada en evidencia, asegurando que el uso de la Inteligencia Artificial en Bogotá sea un instrumento ético, transparente y auditable que contribuya efectivamente a salvaguardar los derechos humanos de las mujeres en el territorio.

5.2. Trabajos futuros

La presente investigación consolida una base técnica para la anticipación territorial de la violencia contra la mujer en Bogotá. No obstante, debido a la naturaleza dinámica y multicausal del fenómeno, se plantean las siguientes líneas de trabajo para fortalecer la capacidad analítica y la respuesta institucional:

- Ecosistemas de datos e interoperabilidad: Se recomienda avanzar hacia la

construcción de un sistema integrado de información que trascienda los actuales registros administrativos. Es prioritario implementar mecanismos de interoperabilidad que permitan cruzar datos en tiempo real entre el sector salud, el sistema judicial y las líneas de emergencia (C4-123). Esta integración debe centrarse en reconstruir el "historial o trayecto de la sobreviviente", permitiendo una gestión de casos informada que evite la duplicidad de registros y la victimización secundaria. Asimismo, se plantea la incorporación de técnicas de Big Data para analizar fuentes no estructuradas, como conversaciones en redes sociales, empleando modelos de procesamiento de lenguaje natural (NLP) como latent dirichlet allocation (LDA) para identificar tendencias y preocupaciones ciudadanas de forma orgánica.

- Evolución hacia modelos de aprendizaje profundo y refuerzo: En el ámbito metodológico, futuras investigaciones podrían explorar arquitecturas de aprendizaje profundo (*Deep Learning*), tales como redes recurrentes (LSTM) o Transformers, para modelar series temporales con mayor precisión ante comportamientos no lineales. De igual forma, se sugiere la implementación de aprendizaje por refuerzo (*Reinforcement Learning*) para optimizar la toma de decisiones dinámicas en la asignación de recursos de protección, permitiendo que el sistema aprenda de forma continua a partir de la efectividad de intervenciones previas.
- Analítica prescriptiva y apoyo a la decisión judicial: Es pertinente evolucionar desde modelos descriptivos y predictivos hacia un enfoque de analítica prescriptiva. Esto implica desarrollar herramientas de apoyo a la toma de decisiones que no solo estimen el riesgo, sino que sugieran rutas de atención personalizadas y proporcionen argumentos legales basados en la perspectiva de género para los equipos judiciales. Este avance facilitaría la clasificación de delitos complejos, como el paso de homicidio a feminicidio, acercando la respuesta estatal a la realidad del territorio.
- Gobernanza ética y mitigación de sesgos: Se propone una línea de investigación

dedicada exclusivamente a la transparencia algorítmica y la mitigación de sesgos. Es imperativo implementar auditorías algorítmicas constantes y el uso de model cards para documentar lecciones aprendidas, asegurando que el sistema sea auditable, justo y libre de la opacidad de la "caja negra" (black-box). El co-diseño con expertos interdisciplinarios y comunidades afectadas garantizará que la IA sea un instrumento ético al servicio de los derechos humanos y la dignidad de las mujeres en Bogotá.

En definitiva, el horizonte de los trabajos futuros debe trascender la optimización técnica para consolidarse en un marco de ética algorítmica y transparencia, donde la auditoría constante y la mitigación de sesgos garanticen que la innovación permanezca al servicio de los derechos humanos. La evolución de estos modelos, sustentada en la interoperabilidad y un enfoque multidisciplinario, será el motor para transformar la ciencia de datos en una herramienta efectiva de política pública que promueva la igualdad y asegure el derecho fundamental de las mujeres a una vida libre de violencias.

Referencias bibliográficas

[1] Observatorio de Asuntos de las Mujeres del Cauca, “Boletín sobre violencias contra las mujeres en el departamento del Cauca,” Popayán, 2020.

[2] Ministerio de Justicia, “Violencia intrafamiliar. Sobre la familia, la violencia al interior del núcleo, sus tipos y las rutas de atención disponibles para víctiams.”

[3] M. y S. Centro de Investigaciones y Estudios de Género, “Violencias basadas en género: percepciones con base en un ejercicio de cartografía social,” *Nómadas*, no. 51, pp. 155–171, Dec. 2019, doi: 10.30578/nomadas.n51a9.

[4] M. D. Sánchez Prada, “Lesiones infligidas por otros,” *Medicina Legal*, 2000.

[5] L. Breiman, “Random Forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[6] L. Breiman, “Manual on Setting Up, Using, and Understanding Random Forests v4.0,” http://oz.berkeley.edu/users/breiman/Using_random_forests_v4.0.pdf.

[7] C.-C. Pinto-Muñoz, J.-A. Zuñiga-Samboni, and H.-A. Ordoñez-Erazo, “Machine Learning Applied to Gender Violence: A Systematic Mapping Study,” *Revista Facultad de Ingeniería*, vol. 32, no. 64, p. e15944, Jun. 2023, doi: 10.19053/01211129.v32.n64.2023.15944.

[8] S. Borrero, L. Martínez, and C. Giraldo, “Aplicación de modelos predictivos al análisis de violencia de género en Colombia,” *Revista Colombiana de Coputación*, vol. 22, no. 2, pp. 45–59, 2021.

[9] F. Hernández Barajas, *Introducción a machine learning*. Universidad Nacional de Colombia, 2026.

[10] I. Rodríguez-Rodríguez, J.-V. Rodríguez, D.-J. Pardo-Quiles, P. Heras-González, and I. Chatzigiannakis, “Modeling and Forecasting Gender-Based Violence through Machine Learning Techniques,” *Applied Sciences*, vol. 10, no. 22, p. 8244, Nov. 2020, doi: 10.3390/app10228244.

[11] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in *22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, San Francisco, 2016, pp. 785–794.

[12] G. Shmueli, “To Explain or to Predict?,” *Statistical Science*, vol. 25, no. 3, pp. 289–310, 2010.

[13] J. H. Friedman, “Greedy Function Approximation: A Gradient Boosting Machine,” *Ann. Stat.*, vol. 29, no. 5, pp. 1189–1232, 2001.

[14] F. Doshi-Velez and B. Kim, *Towards A Rigorous Science of Interpretable Machine Learning*. ArXiv e-prints, 2017.

[15] Z. C. Lipton, “The Mythos of Model Interpretability,” *Queue*, vol. 16, no. 3, pp. 31–57, 2018.

[16] C. Molnar, *Interpretable Machine Learning*, 2nd ed. Munich: Leanpub, 2022.

[17] European Commission, “Ethics Guidelines for Trustworthy AI,” 2019, *Brussels*.

[18] C. M. Bishop, *Pattern Recognition and Machine Learning*. Nueva York: Springer, 2006.

[19] J. MacQueen, “Some methods for classification and analysis of multivariate observations,” in *5th Berkeley Symp. Mathematical Statistics and Probability*, Berkeley, 1967, pp. 281–297.

[20] J. C. Alonso Cifuentes, M. F. Largo Liévano, and C. C. Hoyos Bermeo, *Una introducción a los modelos de Clústering empleando R*. Universidad Icesi, 2025. doi: 10.18046/EUI/bda.h.6.

[21] T. Zhang, R. Ramakrishnan, and M. Livny, “BIRCH: An Efficient Data Clustering Method for Very Large Databases,” in *Management of Data, Montreal*, 1996.

[22] J. Otavalo, “Birch clustering qué es y cómo funciona?,” Medium.com.

[23] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Nueva York: Springer, 2009.

[24] C. C. Aggarwal, *Data Mining: The Textbook*. Cham: Springer, 2015.

[25] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques,* 3rd ed. Burlington: Morgan Kaufmann, 2011.

[26] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 3rd ed. Sebastopol: O'Reilly Media, 2022.

[27] P. J. Rousseeuw, "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, 1987.

[28] A. K. A. K. Jain, M. N. Murly, and P. J. Flynn, "Data Clustering: A Review," *ACM Comput. Surv.*, vol. 31, no. 3, pp. 264–323, 1999.

[29] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*, 3rd ed. Melbourne: OTexts, 2021.

[30] T. Calinski and J. Harabasz, "A Dendrite Method for Cluster Analysis," *Communications in Statistics*, vol. 3, no. 1, pp. 1–27, 1974.

[31] D. L. Davies and D. W. Bouldin, "A Cluster Separation Measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 1, no. 2, pp. 224–227, 1979.

[32] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning: Limitations and oportunities*. Londres: MIT Press, 2023.

[33] C. Wirth, "Data science in the public interest: Using analytics to inform social policy," *IBM J. Res. Dev.*, vol. 62, no. 1, pp. 1–10, 2018.

[34] J. Sáenz, M. Gutiérrez, and L. Moreno, "Predicting domestic violence recidivism with decision trees," *J. Crim. Justice*, vol. 68, pp. 101–110, 2020.

[35] N. Grgić-Hlača, M. B. Zafar, K. P. Gummadi, and A. Weller, "The Case for Process Fairness in Learning: Feature Selection for Fair Decision Making," in *Sympo-sium on Machine Learning and the Law at the 29th Conference on Neural Information Processing Systems*, 2016.

[36] I. J. Goodfellow *et al.*, *Generative Adversarial Networks*. ArXiv e-prints, 2014.

[37] Y. Wang, H. Li, and J. Wang, "Machine learning techniques for predicting gender-based violence: A systematic review," *IEEE Access*, vol. 9, pp. 115234–115246, 2021.

[38] UN Global Pulse, "Using big data to combat violence against women."

[39] D. E. Goin, K. E. Rudolph, and J. Ahern, “Predictors of firearm violence in urban communities: A machine-learning approach,” *Health Place*, vol. 51, pp. 61–67, May 2018, doi: 10.1016/j.healthplace.2018.02.013.

[40] L. M. Thompson Lee, A. Moore, and C. Crossland, “Harnessing the Power of Machine Learning to Prevent Gender-Based Violence: Using Big Data Techniques to Enhance Research on Violence Against Women,” *Violence Against Women*, Feb. 2025, doi: 10.1177/10778012251319701.

[41] IBM, *Manual CRISP-DM de IBM SPSS Modeler*. IBM, 2012.

[42] Data Science PM, “What is CRISP-DM,” https://www.datascience-pm.com/crisp-dm-2/.

[43] P. Chapman *et al.*, “CRISP-DM 1.0 Step-by-step data mining guide,” DaimlerChrysler, 1999.

[44] C. Roberto, “CRIPS-DM: Las 6 etapas de la metodología del futuro,” *MBAUSPESALQ*, Jun. 31, 2022.

[45] M. del P. Roa Avella, J. E. Sanabria-Moyano, and K. Dinas-Hurtado, “Herramientas de predicción de violencia basada en género y feminicidio mediante la Inteligencia Artificial,” *Revista Jurídica Mario Alario D’Filipo*, vol. 15, no. 30, pp. 360–390, 2023.

[46] Organización Mundial de la Salud, “Violencia contra la mujer,” https://www.who.int/es/news-room/fact-sheets/detail/violence-against-women.

[47] UN Women, “Global database on violence against woman and girls,” https://www.google.com/url?q=https://evaw-global-database.unwomen.org/en/countries/americas/colombia&sa=D&source=docs&ust=1779893041308426&usg=AOvVaw398HKos8iH54z5qqCLpA5i.

[48] Alcaldía Mayor de Bogotá, “Informe de gestión 2024,” Bogotá, 2024.

[49] P. Bruce, A. Bruce, and P. Gedeck, *Estadística práctica para ciencia de datos con R y Python*. Bogotá: Marcombo y Alpha Editorial, 2022.

[50] S. García, J. Luengo, and F. Herrera, *Data Preprocessing in Data Mining*. Cham: Springer, 2015.

[51] F. Provost and T. Fawcett, *Data Science for Business*. Sebastopol: O'Reilly Media, 2013.

[52] R. J. A. Litte and D. B. Rubin, *Statistical analysis with missing data*, 3rd ed. Hoboken: Wiley, 2019.

[53] Z. Zhang, "Missing data imputation: focusing on single imputation," *Ann. Transl. Med.*, vol. 4, no. 1, pp. 1–12, 2016.

[54] M. Kuhn and K. Johnson, *Feature engineering and selection: a practical approach for predictive models*. Boca Raton: CRC Press, 2019.

[55] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data mining: practical machine learning tools and techniques*, 4th ed. Burlington: Morgan Kaufmann, 2016.

[56] J. Pagès, *Multiple Factor Analysis by Example Using R.* Boca Raton: CRC Press, 2014.

[57] F. Husson, S. Lê, and J. Pagès, *Exploratory multivariate analysis by example using R*, 2nd ed. Boca Raton: CRC Press, 2017.

[58] M. Kuhn and K. Johnson, *Applied predictive modeling*. Nueva York: Springer, 2013.

[59] C. Hernández, "Por qué la celebración del Día de la Madre es una de las más violentas," El Tiempo.

[60] ONU Mujeres, "Datos y cifras: violencia contra las mujeres," https://www.unwomen.org/es/articulos/datos-y-cifras/datos-y-cifras-violencia-contra-las-mujeres.

[61] UN Woman, "Facts and figures: Ending violence against women," https://www.unwomen.org/en/articles/facts-and-figures/facts-and-figures-ending-violence-against-women.

[62] UN Woman, *Femicides in 2024: Global estimates of intimate partner/family member femicides*. UNODC, 2025.

[63] ONU Mujeres, "Acabar con la violencia contra las mujeres," https://www.unwomen.org/es/what-we-do/ending-violence-against-women.

[64] A. Agresti, *An introduction to categorical data analysis*, 3rd. Hoboken: John Wiley & Sons, 2018.

[65] H. Crámer, *Matematical methods of statistics*. Princeton: Princeton University Press, 1946.

[66] J. Cohen, *Statistical power analysis for the behavioral sciences*, 2nd ed. Hilsdale: Lawrence Eribaum Associates, 1988.

[67] A. Zheng and A. Casari, *Feature Engineering for Machine Learning*. Sebastopol: O'Reilly Media, 2018.

[68] A. J. van der Burg and J. de Leeuw, "Nonlinear principal component analysis and optimal scaling," in *Handbook of multivariate experimental psychology*, Boston: Springer, 1988, pp. 523–551.

[69] D. Sculley, "Web-scale K-Means clustering," in *Proceedings of the 19th International Conference on World Wide Web (WWW '10)*, Raleigh, 2010, pp. 1177–1178.

[70] UN Woman, "Global Database on Violence against Women and Girls," https://data.unwomen.org/global-database-on-violence-against-women.

[71] Comisión Económica para América Latina y el Caribe, "Panorama Social de América Latina y el Caribe," Santiago, 2025.

[72] Comisión Económica para América Latina y el Caribe, "ción para la igualdad, el desarrollo y la paz en América Latina y el Caribe: informe regional sobre el examen de la Declaración y Plataforma de Acción de Beijing a 30 años de su aprobación en sinergia con la implementación de la Agenda Regional de Género," Santiago, 2025.

[73] Programa de Naciones Unidas para el Desarrollo, "Programa Regional de la Iniciativa Spotlight para América Latina, Inteligencia artificial: una herramienta de apoyo para el análisis de la perspectiva de género en los homicidios de mujeres. Propuesta de guía para su elaboración," 2022.

[74] Ministerio de Justicia y del Derecho, *Guía para la valoración de la percepción del riesgo de feminicidio en el contexto familiar*. Bogotá: Ministerio de Justicia y del Derecho, 2023.

[75] C. Russell, M. J. Kusner, J. Loftus, and R. Silva, “When worlds collide: Integrating different counterfactual assumptions in fairness,” *Advances in Neural Information Processing Systems 30, Curran Associates*, pp. 6414–6423, 2017.

[76] P. Ball, “Fairness in Machine Learning with Causal Reasoning,” Sidney Sussex College, Sidney, 2018.

Anexos

Anexo 1. Resultados del random forest a 3 horizontes

localidad	fecha_base_modelo	horizonte_meses	fecha_prediccion	casos_predichos_final
Antonio Nariño	2023-09-01	1	2023-10-01	12
Barrios Unidos	2023-09-01	1	2023-10-01	18
Bosa	2023-09-01	1	2023-10-01	116
Chapinero	2023-09-01	1	2023-10-01	12
Ciudad Bolivar	2023-09-01	1	2023-10-01	114
Engativá	2023-09-01	1	2023-10-01	91
Fontibón	2023-09-01	1	2023-10-01	28
Kennedy	2023-09-01	1	2023-10-01	129
La Candelaria	2023-09-01	1	2023-10-01	9
Los Mártires	2023-09-01	1	2023-10-01	17
Puente Aranda	2023-09-01	1	2023-10-01	21
Rafael Uribe Uribe	2023-09-01	1	2023-10-01	40
San Cristobal	2023-09-01	1	2023-10-01	71
Santafé	2023-09-01	1	2023-10-01	17
Suba	2023-09-01	1	2023-10-01	104
Sumapaz	2023-09-01	1	2023-10-01	2
Teusaquillo	2023-09-01	1	2023-10-01	7
Tunjuelito	2023-09-01	1	2023-10-01	24
Usaquén	2023-09-01	1	2023-10-01	38
Usme	2023-09-01	1	2023-10-01	48
Antonio Nariño	2023-09-01	2	2023-11-01	13
Barrios Unidos	2023-09-01	2	2023-11-01	18
Bosa	2023-09-01	2	2023-11-01	114

Chapinero	2023-09-01	2	2023-11-01	13
Ciudad Bolivar	2023-09-01	2	2023-11-01	109
Engativá	2023-09-01	2	2023-11-01	91
Fontibón	2023-09-01	2	2023-11-01	27
Kennedy	2023-09-01	2	2023-11-01	127
La Candelaria	2023-09-01	2	2023-11-01	8
Los Mártires	2023-09-01	2	2023-11-01	17
Puente Aranda	2023-09-01	2	2023-11-01	21
Rafael Uribe Uribe	2023-09-01	2	2023-11-01	38
San Cristobal	2023-09-01	2	2023-11-01	71
Santafé	2023-09-01	2	2023-11-01	17
Suba	2023-09-01	2	2023-11-01	103
Sumapaz	2023-09-01	2	2023-11-01	2
Teusaquillo	2023-09-01	2	2023-11-01	7
Tunjuelito	2023-09-01	2	2023-11-01	23
Usaquén	2023-09-01	2	2023-11-01	37
Usme	2023-09-01	2	2023-11-01	46
Antonio Nariño	2023-09-01	3	2023-12-01	12
Barrios Unidos	2023-09-01	3	2023-12-01	18
Bosa	2023-09-01	3	2023-12-01	116
Chapinero	2023-09-01	3	2023-12-01	13
Ciudad Bolivar	2023-09-01	3	2023-12-01	113
Engativá	2023-09-01	3	2023-12-01	96
Fontibón	2023-09-01	3	2023-12-01	27
Kennedy	2023-09-01	3	2023-12-01	128
La Candelaria	2023-09-01	3	2023-12-01	8

Los Mártires	2023-09-01	3	2023-12-01	18
Puente Aranda	2023-09-01	3	2023-12-01	21
Rafael Uribe Uribe	2023-09-01	3	2023-12-01	39
San Cristobal	2023-09-01	3	2023-12-01	73
Santafé	2023-09-01	3	2023-12-01	17
Suba	2023-09-01	3	2023-12-01	107
Sumapaz	2023-09-01	3	2023-12-01	2
Teusaquillo	2023-09-01	3	2023-12-01	7
Tunjuelito	2023-09-01	3	2023-12-01	23
Usaquén	2023-09-01	3	2023-12-01	36
Usme	2023-09-01	3	2023-12-01	46

Fuente: elaboración propia.