

Caracterización y predicción del homicidio intencional en Colombia: comparación entre modelos estadísticos y de aprendizaje automático con datos del SIEDCO

Fabián Eduardo Camelo Sánchez

Eduardo José Fontalvo Romero

Bryan García Munévar

Universidad Ean

Facultad de Ingeniería

Maestría en Ciencia de Datos

Bogotá, Colombia

06/Agosto/2026

Caracterización y predicción del homicidio intencional en Colombia: comparación entre modelos estadísticos y de aprendizaje automático con datos del SIEDCO.

Fabián Eduardo Camelo Sánchez

Eduardo José Fontalvo Romero

Bryan García Munévar

Trabajo de grado presentado como requisito para optar al título de:

Magister en Ciencia de Datos

Director (a):

Carolina María Luque Zabala

Modalidad:

Monografía

Universidad Ean

Facultad de Ingeniería

Maestría en Ciencia de Datos

Bogotá, Colombia

06/Agosto/2026

Nota de aceptación:

Firma del jurado

Firma del jurado

Firma del director del trabajo de grado

Ciudad, día/mes/año

A mi familia, profesores y compañeros que fueron parte de este proceso de formación.

Eduardo F.

A Dios, mi amada Jenny, mi hija Isa y Manchitas por su amor inconmensurable. A mis amigos Fabio, Carlos, Bryan, Edu y Camilo, por tantas enseñanzas y buenos momentos.

Fabián C.

A Cronos, guardián del tiempo largo que exige lo que vale la pena hacer bien. A Hermes, que camina entre los datos hasta que empiezan a decir algo cierto.

Bryan G.

Agradecimientos

Expresamos nuestro agradecimiento a la profesora Carolina Luque, tutora del presente trabajo de grado, por su dedicación, orientación y acompañamiento durante todo el proceso de investigación, así como por sus valiosas correcciones y observaciones, que contribuyeron significativamente a la culminación de este trabajo.

Agradecemos también a la Universidad EAN, por brindar el espacio académico, los recursos y la formación necesarios para llevar a cabo este trabajo de grado.

Resumen

Este trabajo aborda el pronóstico del homicidio intencional en Colombia, fenómeno prioritario para la seguridad pública. Si bien la literatura ha explorado modelos espaciales y de aprendizaje automático sobre datos delictivos, persiste una brecha en la comparación sistemática entre enfoques estadísticos clásicos y de ciencia de datos evaluados frente a datos reales no observados durante el entrenamiento, con control riguroso de fuga de información. El objetivo fue comparar el desempeño de ambos enfoques en el pronóstico mensual del homicidio a escala nacional e identificar las variables más influyentes en dicho pronóstico. Siguiendo la metodología CRISP-DM (sin incluir la fase de despliegue, dado que el alcance del proyecto no contempla la puesta en producción del modelo), se construyó una serie mensual de víctimas de homicidio intencional para 2015–2025 a partir de 110.381 registros, extraídos de una base consolidada de aproximadamente 3,78 millones de registros de distintas temáticas delictivas del Sistema de Información Estadístico, Delincuencial, Contravencional y Operativo (SIEDCO). Estos registros corresponden a 141.774 víctimas. La serie se complementó con variables exógenas socioeconómicas y delictivas (hurto a personas, lesiones personales, violencia intrafamiliar, incautación de estupefacientes, actividad económica y desempleo). Se entrenaron y compararon nueve modelos, tres estadísticos (SARIMA, SARIMAX y VAR) y seis de aprendizaje automático (XGBoost, LightGBM, Random Forest, LSTM, GRU y Prophet), evaluados frente al homicidio efectivamente observado entre enero y mayo de 2026, periodo no utilizado en ningún momento del entrenamiento ni de la selección de modelos. LightGBM obtuvo el mejor desempeño en la ventana evaluada, con un margen amplio sobre el resto (RMSE 16,61; MASE 0,19; R^2 0,84, este último de valor únicamente indicativo por calcularse sobre cinco observaciones de prueba), seguido de SARIMAX,

Random Forest y VAR, los únicos otros modelos con ajuste positivo; XGBoost, LSTM y GRU tuvieron un desempeño deficiente. El error de pronóstico no mostró una relación sistemática con el horizonte, contrario a lo esperado inicialmente para los modelos estadísticos. El análisis de interpretabilidad mediante valores SHAP identificó a la violencia intrafamiliar del mes anterior como la variable más influyente en los modelos de árboles, incluido el ganador, por encima del rezago anual de la propia serie. Se concluye que el aprendizaje automático puede superar de forma sustancial a los modelos estadísticos clásicos cuando se evalúa frente a datos reales prospectivos, con la salvedad de que, con cinco observaciones de prueba, la ventaja del mejor modelo sobre el mejor modelo estadístico resulta indicativa y no concluyente en términos inferenciales; finalmente este resultado depende fuertemente del algoritmo específico y no se generaliza por igual a toda la familia de aprendizaje automático.

Palabras clave: homicidio intencional, pronóstico de series temporales, aprendizaje automático, SARIMA, LightGBM, valores SHAP, CRISP-DM, SIEDCO.

Abstract

This study addresses the forecasting of intentional homicide in Colombia, a phenomenon of major relevance to public security. Although previous research has explored spatial and machine-learning models applied to crime data, a gap remains in the systematic comparison of classical statistical approaches and data science methods when evaluated against actual observations not seen during training, under rigorous controls to prevent information leakage. The objective of this study was to compare the predictive performance of both approaches for monthly homicide forecasting at the national level and to identify the variables with the greatest influence on the forecasts.

Following the CRISP-DM methodology—excluding the deployment phase, as the project did not include putting the model into production—a monthly time series of intentional homicide victims covering 2015–2025 was constructed from 110,381 records representing 141,774 victims. These records were extracted from a consolidated database of approximately 3.78 million records covering multiple crime categories in the Statistical, Criminal, Contraventional, and Operational Information System (SIEDCO). The series was complemented with socioeconomic and crime-related exogenous variables, including theft from persons, personal injury, domestic violence, drug seizures, economic activity, and unemployment.

Nine models were trained and compared: three statistical models (SARIMA, SARIMAX, and VAR) and six machine-learning models (XGBoost, LightGBM, Random Forest, LSTM, GRU, and Prophet). Their forecasts were evaluated against the actual number of homicides observed between January and May 2026, a period that was not used at any stage of model training or model selection.

LightGBM achieved the best performance over the evaluation window, outperforming the remaining models by a wide margin (RMSE = 16.61; MASE = 0.19; R^2 = 0.84, with the latter interpreted only as indicative because it was calculated from five test observations). It was followed by SARIMAX, Random Forest, and VAR, which were the only other models to exhibit positive predictive fit, whereas XGBoost, LSTM, and GRU performed poorly. Forecast errors did not show a systematic relationship with the forecast horizon, contrary to the initial expectation for the statistical models.

Interpretability analysis using SHAP values identified domestic violence in the previous month as the most influential predictor in the tree-based models, including the best-performing model, exceeding even the importance of the homicide series' own annual lag.

The findings indicate that machine-learning methods can substantially outperform classical statistical models when evaluated against actual prospective observations. However, given that the test window consisted of only five observations, the advantage of the best-performing model over the best statistical model should be regarded as indicative rather than inferentially conclusive. Furthermore, this result depends strongly on the specific algorithm employed and should not be generalized uniformly across the entire family of machine-learning methods.

Keywords: intentional homicide, time series forecasting, machine learning, SARIMA, LightGBM, SHAP values, CRISP-DM, SIEDCO.

Contenido

Pág.

Lista de Figuras 13

Lista de Tablas 15

1. Introducción 19

2. Objetivos 23

Objetivo general 23

Objetivos específicos 23

3. Justificación 24

4. Marco Teórico 27

4.1 Fundamentos conceptuales del homicidio intencional como fenómeno social 27

4.2 La dimensión espacial del homicidio 28

4.3 Modelos estadísticos de series temporales para el análisis del delito 28

4.4 Ingeniería de variables temporales 30

4.5 Fundamentos teóricos de la ciencia de datos 31

4.6 Marcos metodológicos en ciencia de datos para proyectos predictivos 33

4.7 Modelos predictivos y métricas de desempeño 34

4.8 Interpretabilidad de modelos: valores SHAP 38

4.9 Validación temporal y fuga de información 38

4.10 Estado del arte en modelos de predicción del homicidio intencional 39

5. Hipótesis 41

6. Variables 44

7. Metodología 47

7.1 Enfoque y alcance de la investigación 47

7.2 Fuente de datos y evaluación de su calidad 48

7.3 Población y delimitación temporal 52

7.4 Preparación de datos e ingeniería de variables 53

7.5 Especificación e implementación de los modelos 56

7.6 Entorno de reproducibilidad 63

7.7 Esquema de validación 64

7.8 Métricas de evaluación 66

7.9 Interpretabilidad (SHAP) 68

8. Modelamiento y Evaluación de Resultados 70

8.1 Procesamiento de los datos 70

8.2 Análisis exploratorio 71

8.3 Modelos estadísticos 80

8.4 Modelos de aprendizaje automático 98

8.5 Verificación de hipótesis y propuestas sobre resultados 141

9. Discusión 149

9.1 Contraste de hipótesis 149

9.2 Recomendaciones metodológicas para el pronóstico del homicidio intencional en Colombia 154

9.3 Consideraciones éticas y de uso responsable del modelo 156

10. Conclusiones y Trabajo Futuro 160

Conclusiones 160

Trabajo futuro 164

Referencias 166

Lista de Figuras

Pág.

Figura 1 *Tendencia del comportamiento del homicidio intencional - 10 años* 24

Figura 2 *Relación entre objetivos específicos e hipótesis de investigación*..... 42

Figura 3 *Marco metodológico del estudio* 47

Figura 4 *Proceso de preparación de datos e ingeniería de variables* 57

Figura 5 *Serie mensual de homicidios en Colombia*..... 72

Figura 6 *Descomposición aditiva de la serie de homicidios (2015-2025)* 73

Figura 7 *Estacionalidad de la serie de tiempo* 74

Figura 8 *Distribución de los homicidios por género de la víctima (2015-2025)* ... 75

Figura 9 *Distribución de los homicidios por zona (2015-2025)* 76

Figura 10 *Distribución de los homicidios por arma o medio empleado (2015-2025)* 77

Figura 11 *Distribución espacial de la tasa de homicidios (2025)* 79

Figura 12 *ACF y PACF* 82

Figura 13 *Diagnóstico de los residuales del modelo SARIMA: función de autocorrelación (ACF) e histograma.* 85

Figura 14 *Serie histórica de homicidios (2015-2025) y pronóstico del modelo SARIMA para enero-mayo de 2026, con intervalo de confianza del 95%.* 86

Figura 15 *Serie histórica de homicidios (2015-2025) y pronóstico del modelo SARIMAX para enero-mayo de 2026, con intervalo de confianza del 95%.* 90

Figura 16 *Serie histórica de homicidios (2015-2025) y pronóstico del modelo VAR HOMICIDIOS~ISE para enero-mayo de 2026, con intervalo de confianza del 95%.* 94

Figura 17 *Comparación de los pronósticos de enero-mayo de 2026 de los tres modelos estadísticos frente a la serie histórica de homicidios.* 97

Figura 18 *Importancia de variables (valores SHAP) del modelo XGBoost.* 102

Figura 19 *Serie histórica de homicidios (2015-2025) y pronóstico de XGBoost para enero-mayo de 2026 con intervalo aproximado del 95 %.* 104

Figura 20 *Importancia de variables (valores SHAP) del modelo LightGBM.* 109

Figura 21 *Serie histórica de homicidios (2015-2025) y pronóstico de LightGBM para enero-mayo de 2026 con intervalo aproximado del 95 %.* 111

Figura 22 *Importancia de variables (valores SHAP) del modelo Random Forest.* 115

Figura 23 *Serie histórica de homicidios (2015-2025) y pronóstico de Random Forest para enero-mayo de 2026 con intervalo aproximado del 95 %.* 117

Figura 24 *Importancia de variables (valores SHAP) del modelo LSTM.* 122

Figura 25 *Serie histórica de homicidios (2015-2025) y pronóstico de LSTM para enero-mayo de 2026 con intervalo aproximado del 95 %.* 124

Figura 26 *Importancia de variables (valores SHAP) del modelo GRU.* 128

Figura 27 *Serie histórica de homicidios (2015-2025) y pronóstico de GRU para enero-mayo de 2026 con intervalo aproximado del 95 %.* 130

Figura 28 *Descomposición aditiva del modelo Prophet: tendencia, estacionalidad anual y efecto de las variables exógenas.* 135

Figura 29 *Coeficientes estimados de las variables exógenas del modelo Prophet.* 136

Figura 30 *Serie histórica de homicidios (2015-2025) y pronóstico de Prophet para enero-mayo de 2026, con intervalo de confianza del 95%.* 138

Figura 31 *Comparación del pronóstico de enero-mayo de 2026 de los nueve modelos frente al valor realmente observado.* 144

Lista de Tablas

Pág.

Tabla 1 *Descripción de las fases de la metodología CRISP-DM* 34

Tabla 2 *Resumen de las métricas de desempeño para regresión*..... 36

Tabla 3 *Resumen de las métricas de desempeño para clasificación* 37

Tabla 4 *Operacionalización de las variables*..... 44

Tabla 5 *Comparación de cantidad de homicidios según SIEDCO y Forensis* 51

Tabla 6 *Mallas de hiperparámetros por modelo* 58

Tabla 7 *Especificaciones de referencia para la identificación de órdenes SARIMA y SARIMAX* 60

Tabla 8 *Entorno de ejecución y librerías* 63

Tabla 9 *Métricas de evaluación*..... 66

Tabla 10 *Mejores modelos SARIMA considerados* 83

Tabla 11 *Pronóstico puntual e intervalo de confianza del 95% del modelo SARIMA, enero-mayo de 2026.*..... 85

Tabla 12 *Métricas de desempeño del modelo SARIMA frente al valor observado real de enero-mayo de 2026.* 86

Tabla 13 *Mejores modelos SARIMAX considerados.* 88

Tabla 14 *Pronóstico puntual e intervalo de confianza del 95% del modelo SARIMAX, enero-mayo de 2026.* 89

Tabla 15 *Métricas de desempeño del modelo SARIMAX frente al valor observado real de enero-mayo de 2026.* 90

Tabla 16 *Prueba de causalidad de Granger* 92

Tabla 17 *Modelos VAR considerados.* 92

Tabla 18 *Pronóstico puntual e intervalo de confianza del 95% del modelo VAR HOMICIDIOS~ISE, enero-mayo de 2026.* 93

Tabla 19 *Métricas de desempeño del modelo VAR HOMICIDIOS~ISE frente al valor observado real de enero-mayo de 2026.* 95

Tabla 20 *Pronósticos de enero a mayo de 2026 de los tres modelos estadísticos (SARIMA, SARIMAX y VAR).* 95

Tabla 21 *Métricas de desempeño de los tres modelos estadísticos.* 96

Tabla 22 *Mejores 10 combinaciones de variables exógenas e hiperparámetros identificados para XGBoost.* 101

Tabla 23 *Pronóstico puntual e intervalo aproximado del 95% de XGBoost, enero-mayo de 2026.* 103

Tabla 24 *Métricas de desempeño de XGBoost frente al valor observado real de enero-mayo de 2026.* 105

Tabla 25 *Mejores 10 combinaciones de variables exógenas e hiperparámetros identificados para LightGBM.* 107

Tabla 26 *Pronóstico puntual e intervalo aproximado del 95% de LightGBM, enero-mayo de 2026.* 110

Tabla 27 *Métricas de desempeño de LightGBM frente al valor observado real de enero-mayo de 2026.* 112

Tabla 28 *Mejores 10 combinaciones de variables exógenas e hiperparámetros identificados para Random Forest.* 114

Tabla 29 *Pronóstico puntual e intervalo aproximado del 95% de Random Forest, enero-mayo de 2026.* 116

Tabla 30 *Métricas de desempeño de Random Forest frente al valor observado real de enero-mayo de 2026.* 118

Tabla 31 *Mejores 10 combinaciones de variables exógenas e hiperparámetros identificados para LSTM.* 120

Tabla 32 *Pronóstico puntual e intervalo aproximado del 95% de LSTM, enero-mayo de 2026.* 123

Tabla 33 *Métricas de desempeño de LSTM frente al valor observado real de enero-mayo de 2026.* 125

Tabla 34 *Mejores 10 combinaciones de variables exógenas e hiperparámetros identificados para GRU.* 126

Tabla 35 *Pronóstico puntual e intervalo aproximado del 95% de GRU, enero-mayo de 2026.* 129

Tabla 36 *Métricas de desempeño de GRU frente al valor observado real de enero-mayo de 2026.* 131

Tabla 37 *Mejores 10 combinaciones de variables exógenas e hiperparámetros identificados para Prophet.* 133

Tabla 38 *Coeficientes de las seis variables exógenas del modelo Prophet.* 136

Tabla 39 *Pronóstico puntual e intervalo de confianza del 95% de Prophet, enero-mayo de 2026.* 137

Tabla 40 *Métricas de desempeño de Prophet frente al valor observado real de enero-mayo de 2026.* 139

Tabla 41 *Síntesis de las variables más influyentes por modelo, según SHAP (modelos de árboles y redes recurrentes) o coeficientes estimados (Prophet).* 140

Tabla 42 *Métricas de desempeño de los nueve modelos, ordenadas de mejor a peor por RMSE.* 141

Tabla 43 *Pronósticos de enero a mayo de 2026 de los nueve modelos evaluados.* 143

Tabla 44 *Error de pronóstico mensual (real menos predicho) de los nueve modelos frente al valor observado, enero-mayo de 2026.* 145

Tabla 45 *Prueba de Diebold-Mariano de LightGBM frente a los demás modelos, enero-mayo de 2026.*..... 147

1. Introducción

El homicidio intencional constituye uno de los problemas de seguridad más graves en América Latina y el Caribe. Con apenas el 8 % de la población mundial, la región concentra cerca de un tercio de los homicidios globales, lo que la convierte en la zona más violenta del planeta (United Nations Office on Drugs and Crime [UNODC], 2023). Colombia no es una excepción: en 2022 se registraron 13.536 homicidios, equivalentes a una tasa de 23,2 por cada 100.000 habitantes, con una tendencia que muestra incrementos sostenidos en los últimos años (Núñez Rivera et al., 2023). Este fenómeno no es aleatorio ni homogéneo en el territorio: investigaciones previas han documentado que las tasas de homicidio se asocian con factores estructurales como la desigualdad, la pobreza, el desempleo, la presencia de actores armados y las dinámicas del crimen organizado (Fajnzylber et al., 2002; Soares y Naritomi, 2010). A nivel regional, el Fondo Monetario Internacional estima que los municipios con tasas de homicidio un 10 % más altas presentan una actividad económica inferior en aproximadamente un 4 %, lo que evidencia el impacto macroeconómico del fenómeno más allá de sus consecuencias humanas (Fondo Monetario Internacional [FMI], 2024).

La comprensión de las dinámicas del homicidio en Colombia ha avanzado considerablemente gracias a enfoques descriptivos y econométricos. Santos-Marquez (2022) evaluó cinco métodos de pronóstico, incluyendo ARIMA, ETS y modelos de convergencia espacial beta, para proyectar tasas de homicidio en 1.120 municipios colombianos entre 2003 y 2018, encontrando que los modelos que incorporan efectos espaciales superan a las técnicas univariadas clásicas en horizontes de cuatro años. Por su parte, Santos-Marquez y Ribeiro-Silveira (2021) demostraron que la autocorrelación espacial de los homicidios es significativa a escala municipal pero no departamental, lo

que subraya la importancia de la granularidad territorial en el diseño de políticas de seguridad. Sin embargo, estos trabajos se centran en tasas anuales desagregadas por municipio y no exploran la capacidad predictiva de modelos de aprendizaje automático sobre series temporales, ni establecen protocolos de validación que prevengan la fuga de información temporal, un problema metodológico que puede inflar artificialmente las métricas de desempeño en contextos de pronóstico secuencial (Kapoor y Narayanan, 2023).

En paralelo, la ciencia de datos ha emergido como una alternativa metodológica con creciente aplicación en el pronóstico del crimen. A nivel nacional, Ordóñez-Eraso et al. (2020) aplicaron modelos de Random Forest sobre registros históricos de la Policía Nacional de Colombia para predecir tendencias de homicidio violento en un horizonte de cinco años, demostrando la viabilidad del aprendizaje automático sobre datos policiales colombianos. A escala de ciudad intermedia, Gelvez-Ferreira et al. (2022) desarrollaron un modelo de predicción del crimen en Bucaramanga combinando procesamiento de grafos y vectorización espacio-temporal, evidenciando que las relaciones espaciales entre barrios mejoran la capacidad predictiva a nivel local. En el ámbito internacional, Verrey et al. (2023) evaluaron la factibilidad de pronosticar homicidios domésticos en Londres mediante un ensamble de algoritmos y encontraron que supera significativamente a los instrumentos estadísticos convencionales. Estos avances confirman el potencial del aprendizaje automático, pero también revelan una limitación común: la mayoría de los estudios se focalizan en una sola ciudad, en un único tipo de modelo, o en crimen de naturaleza distinta al homicidio intencional, sin ofrecer comparaciones sistemáticas entre enfoques estadísticos y de *machine learning* sobre series temporales nacionales.

En el campo específico del pronóstico temporal del crimen mediante modelos estadísticos, SARIMA ha mostrado ser altamente competitivo. Kontopoulou et al. (2023),

en una revisión sistemática de enfoques ARIMA frente a modelos de aprendizaje automático en series temporales, concluyen que los modelos estadísticos clásicos mantienen ventajas comparativas en horizontes cortos y cuando la señal temporal es dominante, mientras que los modelos de *machine learning* ganan relevancia al aumentar la dimensionalidad de los datos de entrada. Este hallazgo es consistente con lo reportado por Alsubayhin et al. (2024), quienes compararon *Random Forest*, regresión logística y *LightGBM* en tareas de predicción del crimen, encontrando que el rendimiento relativo de cada modelo varía sustancialmente según el contexto de datos disponibles. A pesar de estos avances, no existe en Colombia un estudio que evalúe de forma sistemática y bajo criterios de validación temporal el desempeño comparativo de modelos estadísticos frente a múltiples algoritmos de *machine learning* sobre series mensuales nacionales de homicidio con más de una década de registro.

Esta brecha metodológica es relevante no solo desde la perspectiva académica, sino también desde la formulación de política pública. Los tomadores de decisiones en seguridad requieren herramientas que no solo predigan, sino que sean evaluadas con rigor sobre datos reales, construidas sin fuga de información temporal y acompañadas de análisis de interpretabilidad que permitan comprender qué factores impulsan el pronóstico. En este contexto, la presente investigación se plantea la siguiente pregunta: ¿Cuál es el desempeño comparativo de los modelos estadísticos y de ciencia de datos en el pronóstico del homicidio intencional en Colombia, y qué variables resultan más influyentes en dicho pronóstico?

Para responder a esta pregunta, se construyó una serie temporal mensual de víctimas de homicidio intencional en Colombia para el periodo 2015–2025. Se utilizaron 110.381 registros correspondientes a 141.774 víctimas, extraídos de una base consolidada de aproximadamente 3,78 millones de registros de distintas temáticas delictivas del

Sistema de Información Estadístico, Delincuencial, Contravencional y Operativo (SIEDCO) de la Policía Nacional. La serie se complementó con variables exógenas de naturaleza socioeconómica y delictiva. Sobre esta serie se entrenaron y compararon nueve modelos bajo un esquema común de validación temporal: tres modelos estadísticos (SARIMA, SARIMAX y VAR) y seis modelos de aprendizaje automático (XGBoost, LightGBM, Random Forest, LSTM, GRU y Prophet). A diferencia de una evaluación basada únicamente en un holdout histórico, el desempeño de los nueve modelos se contrastó frente al homicidio efectivamente observado entre enero y mayo de 2026, un periodo nunca utilizado durante el entrenamiento ni la selección de los modelos. La interpretabilidad se evaluó mediante valores SHAP.

Este trabajo ofrece una comparación sistemática y metodológicamente rigurosa entre modelos estadísticos y de aprendizaje automático para el pronóstico mensual del homicidio intencional en Colombia a escala nacional, evaluada frente a datos reales prospectivos y no solo frente a un holdout histórico. Además, establece un pipeline reproducible con control explícito de fuga de información temporal, un estándar metodológico que la literatura colombiana sobre predicción del crimen aún no ha consolidado. También aporta evidencia sobre las variables temporales y exógenas más influyentes en el comportamiento predictivo de los modelos, con hallazgos que no eran anticipables desde el planteamiento inicial de la investigación.

2. Objetivos

Objetivo general

Comparar el desempeño de modelos estadísticos clásicos y de aprendizaje automático en el pronóstico mensual del homicidio intencional a escala nacional en Colombia, bajo un esquema de validación temporal riguroso, con el fin de establecer una línea base metodológica que oriente futuros desarrollos predictivos en seguridad.

Objetivos específicos

- Construir una serie temporal mensual nacional de homicidios intencionales a partir de los registros del SIEDCO (2015-2025), complementada con variables exógenas de naturaleza socioeconómica y delictiva, aplicando un proceso reproducible de limpieza, agregación e ingeniería de variables con control explícito de fuga de información.
- Implementar y evaluar modelos estadísticos y de aprendizaje automático para el pronóstico del homicidio, comparando su capacidad predictiva mediante métricas de error frente a datos reales no utilizados durante el entrenamiento ni la selección de modelos.
- Identificar, mediante valores SHAP, las variables más influyentes en el pronóstico e interpretar su aporte al comportamiento de los modelos.
- Determinar en qué condiciones cada familia de modelos resulta más competitiva, derivando recomendaciones metodológicas para el pronóstico del homicidio intencional en Colombia.

3. Justificación

La magnitud y persistencia del homicidio intencional en Colombia justifican por sí solas la necesidad de herramientas cuantitativas que permitan anticipar su comportamiento. Sin embargo, la pertinencia de este trabajo no se agota en la gravedad del fenómeno, sino que responde a una decisión metodológica específica sobre cómo abordarlo: establecer, en primer lugar, una línea base de pronóstico a escala nacional y con rigor temporal, antes de avanzar hacia la heterogeneidad territorial que la literatura ya ha documentado, como se detalla en la Figura 1

Tendencia del comportamiento del homicidio intencional - 10 años

Figura 1
Tendencia del comportamiento del homicidio intencional - 10 años



Nota. Tomado de Factores sociodemográficos en la dinámica del comportamiento delictivo: análisis descriptivo de criminalidad en Colombia, año 2022, por L. C. Núñez Rivera, F. Tolentino Pulido y H. Rodríguez Barrios, 2023, Revista Criminalidad, pp. 161-280.

La literatura ha documentado ampliamente que la violencia homicida no es homogénea en el territorio colombiano: existen agrupamientos de municipios con niveles de violencia sistemáticamente más altos, como el Bajo Cauca antioqueño, Arauca, Meta y el Eje Cafetero, frente a otros con niveles bajos y sostenidos, como Atlántico y Boyacá

(Sánchez Jabba et al., 2012). Esta evidencia espacial es valiosa para el diseño de política pública territorial, pero también plantea una pregunta metodológica previa: antes de modelar esa heterogeneidad territorial, es necesario contar con una línea base confiable a escala nacional que permita comparar, bajo un mismo estándar de rigor, el desempeño de distintas familias de modelos de pronóstico. Este trabajo se ubica precisamente en esa etapa previa.

Estudios como el de Soares Filho et al. (2024) en Brasil muestran que, una vez se cuenta con un pronóstico temporal confiable, incorporar la dimensión espacial permite identificar clústeres de alta incidencia y comprender cómo la desigualdad, la urbanización y la debilidad institucional inciden en la distribución territorial del crimen. Este trabajo no aborda esa dimensión espacial: se concentra deliberadamente en la serie nacional agregada, con el fin de aislar y evaluar con rigor metodológico el desempeño comparativo de los modelos estadísticos y de aprendizaje automático en el dominio temporal, antes de introducir la complejidad adicional que supondría modelar simultáneamente el tiempo y el territorio. Esta decisión de alcance no resta relevancia al fenómeno espacial, documentado también en la Sección 4.2, sino que reconoce que una comparación metodológica rigurosa de modelos de pronóstico requiere, en esta primera etapa, controlar la mayor cantidad de fuentes de variación posible; la heterogeneidad territorial queda planteada como la extensión natural de este trabajo en la Sección 10.

De igual manera, investigaciones recientes en Colombia han mostrado la estrecha relación entre el conflicto armado y la falta de acceso a la educación de calidad, sugiriendo que la violencia homicida tiene impactos multidimensionales que trascienden lo meramente criminal (Arbona et al., 2024). Esta evidencia refuerza la pertinencia de contar con herramientas predictivas rigurosas en Colombia, no solo para dimensionar la magnitud del

problema, sino también para orientar decisiones de política pública con evidencia cuantitativa.

En consecuencia, este trabajo estableció una línea base metodológica que compara nueve modelos estadísticos y de aprendizaje automático para el pronóstico del homicidio intencional a escala nacional, incorporando tanto la memoria propia de la serie como variables exógenas de naturaleza socioeconómica y delictiva, y evaluándolos frente a datos reales no utilizados durante el entrenamiento. Al hacerlo con rigor metodológico y sin fuga de información temporal, este proyecto ofrece insumos prácticos para orientar el diseño de herramientas predictivas de apoyo a la toma de decisiones en seguridad, y sienta las bases empíricas y metodológicas necesarias para que estudios futuros incorporen la dimensión territorial del fenómeno.

4. Marco Teórico

4.1 Fundamentos conceptuales del homicidio intencional como fenómeno social

El homicidio intencional constituye un indicador “punta de iceberg” de la violencia letal y un reflector de desórdenes sociales, económicos y territoriales. A nivel comparado, la UNODC documenta patrones persistentes de concentración geográfica y temporal del homicidio, así como su asociación con economías ilegales, disputas entre actores armados y contextos urbanos desiguales (UNODC, 2023). En Colombia, estas regularidades se combinan con dinámicas propias de conflicto, urbanización acelerada y desigualdad territorial, lo que exige modelos capaces de capturar la persistencia temporal y los choques que caracterizan su evolución.

Desde la teoría criminológica, al menos tres teorías son clave para la modelación predictiva. Primero, la desorganización social postula que la debilidad de redes y capacidades comunitarias eleva la violencia; su medición mediante indicadores de heterogeneidad, movilidad residencial o precariedad institucional ha mostrado fuerte poder explicativo en contextos urbanos (Sampson y Groves, 1989). Segundo, la actividad rutinaria enfatiza la convergencia espacio-temporal entre ofensores motivados, objetivos adecuados y ausencia de guardianes, lo que justifica incluir en los modelos variables de exposición y oportunidad (Cohen y Felson, 1979). Tercero, los enfoques económicos del crimen destacan incentivos y costos esperados de la conducta delictiva, conectando desempleo, informalidad y desigualdad con ciclos de violencia (Becker, 1968).

La evidencia colombiana converge con estos marcos. Sánchez y Núñez (2001) mostraron que desempleo, desigualdad y presencia de actores armados se asocian a mayor homicidio; Sandoval (2014) vinculó transición demográfica, urbanización y tasas de homicidio; y estudios regionales han subrayado la importancia de condiciones

socioeconómicas y capacidades institucionales diferenciadas entre territorios. Estas regularidades respaldan la inclusión de covariables sociodemográficas, económicas y de conflictividad en el análisis del fenómeno.

En suma, el abordaje teórico-empírico articula por qué ocurren los homicidios (teorías sociales y económicas), cómo evolucionan en el tiempo (dependencia temporal) y cómo anticiparlos (modelos estadísticos y de ciencia de datos). Este andamiaje sustenta el objetivo de la tesis: pronosticar el homicidio intencional en Colombia para aportar evidencia cuantitativa a la formulación y evaluación de políticas de seguridad y prevención.

4.2 La dimensión espacial del homicidio

La literatura criminológica documenta también una marcada dimensión espacial del homicidio: su concentración en clústeres o *hotspots* persistentes y los mecanismos de difusión entre territorios vecinos, estudiados mediante econometría espacial e indicadores locales de asociación espacial (Anselin, 1995; Messner et al., 1999). Si bien esta dimensión es relevante para comprender el fenómeno, el presente trabajo se circunscribe a la serie temporal mensual de homicidios a escala nacional; por ello, los enfoques espaciales se reseñan únicamente como contexto y se proponen como línea de extensión futura, sin constituir el núcleo metodológico de esta investigación.

4.3 Modelos estadísticos de series temporales para el análisis del delito

El análisis y pronóstico del delito exige modelar dependencias dinámicas en el tiempo, estacionalidad, como ciclos semanales o mensuales, y choques exógenos asociados a políticas públicas o eventos macro. La metodología Box–Jenkins ofrece un marco clásico para series de tiempo de conteos delictivos, con un proceso ordenado de identificación, estimación, diagnóstico y validación fuera de muestra (Box et al., 2015; Hyndman y Athanasopoulos, 2021).

En términos prácticos, Box–Jenkins se implementa mediante modelos ARIMA y sus extensiones estacionales (SARIMA). Cuando existen determinantes externos, por ejemplo, desempleo, presencia de actores armados o intervenciones policiales, se emplean especificaciones con regresores exógenos, ARIMAX o, si además se incorpora la componente estacional, SARIMAX, y, cuando el impacto de estos factores sigue una dinámica propia, funciones de transferencia para análisis de intervención y cuantificación de efectos diferidos (Box y Tiao, 1975; Box et al., 2015). Este enfoque permite evaluar cambios estructurales y medir impactos en el corto y mediano plazo con criterios comparables de desempeño predictivo (Hyndman y Athanasopoulos, 2021).

Cuando el objetivo es capturar relaciones dinámicas bidireccionales entre el homicidio y otras variables, en lugar de tratarlas como determinantes exclusivamente unidireccionales, resulta pertinente el modelo de vectores autorregresivos (VAR), que trata todas las series involucradas como endógenas y permite que cada una dependa de sus propios rezagos y de los rezagos de las demás (Sims, 1980). Este enfoque es especialmente útil cuando existe razón teórica para sospechar una relación de retroalimentación, como ocurre entre el homicidio y la actividad económica, y se complementa con pruebas de causalidad en el sentido de Granger para contrastar la dirección de dicha relación.

Dado que los conteos delictivos suelen presentar sobredispersión y asimetría, se recurre a transformaciones estabilizadoras o a formulaciones alternativas en espacio de estados y marcos de modelos lineales generalizados. No obstante, la disciplina Box–Jenkins permanece: (i) identificación mediante patrones de autocorrelación y pruebas de raíz unitaria; (ii) estimación por máxima verosimilitud; (iii) diagnóstico de residuos para verificar ruido blanco; y (iv) validación con particiones bloqueadas en el tiempo (p. ej., “*rolling origin*”) (Box et al., 2015; Hyndman y Athanasopoulos, 2021).

En síntesis, el andamiaje Box–Jenkins constituye una base sólida para el pronóstico temporal en conteos delictivos: identificación, estimación, diagnóstico y validación con particiones bloqueadas en el tiempo ("*rolling origin*"). Combinado con métricas de error comparables, este enfoque fortalece la capacidad de pronosticar el homicidio intencional a escala nacional y de evaluar el efecto de políticas de prevención (Box et al., 2015; Hyndman y Athanasopoulos, 2021).

4.4 Ingeniería de variables temporales

Bontempi et al. (2013) postulan que es posible reencuadrar una serie temporal como un problema de aprendizaje supervisado mediante el uso de rezagos, es decir, una ventana finita de valores pasados que se emplea como vector de entrada, de manera que el pronóstico se convierte en una tarea de regresión convencional donde un modelo aprende a mapear la dependencia funcional entre estos patrones históricos y los valores futuros. Además, como lo explican Polyzos y Siriopoulos (2023), en muchos modelos que manejan la variable temporal es necesario incluir los rezagos, que permiten aportar información sobre los efectos de eventos pasados y ayudan a predecir valores futuros de la variable dependiente del estudio.

Asimismo, otro de los conceptos aplicados es la media móvil, un indicador que busca atenuar las fluctuaciones de la serie en un periodo determinado y, teniendo en cuenta la tendencia, reducir el ruido (Herrán et al., 2008). Esto se logra mediante una ventana móvil, un subconjunto de tamaño fijo que recorre los datos de forma secuencial.

Otro punto para tener en cuenta es la estacionariedad, propiedad según la cual las características estadísticas de la serie se mantienen constantes a lo largo del tiempo y no dependen del momento en que se observan. Dado que muchos modelos asumen series estacionarias, suele aplicarse la diferenciación u otras transformaciones que estabilizan la media y la varianza antes del modelado (Hyndman y Athanasopoulos, 2021).

Por último, la estacionalidad corresponde a los patrones que se repiten a intervalos regulares y conocidos y constituye un componente frecuente en las series de homicidios registradas con periodicidad mensual. Para que los modelos aprovechen esta información, la estacionalidad puede incorporarse mediante variables de calendario (como el mes del año) o mediante términos de Fourier que representan los ciclos de la serie; en los modelos estadísticos como SARIMA, este componente se modela de forma explícita a través de su parte estacional (Hyndman y Athanasopoulos, 2021).

4.5 Fundamentos teóricos de la ciencia de datos

La ciencia de datos surgió como término en la década de 1990, aunque sus fundamentos se remontan a dos hilos históricos más antiguos: la recolección y el análisis de datos (Kelleher y Tierney, 2018). Históricamente, sus aportes se encontraban dispersos entre diversas disciplinas, incluyendo la estadística y las matemáticas, la computación y la gestión de datos. Sin embargo, la proliferación masiva de datos en prácticamente todos los ámbitos ha diluido las fronteras entre estas áreas, dando lugar a un perfil profesional integral en el que se espera que el científico de datos combine competencias en las tres (OpenStax, 2025).

En términos prácticos, consiste en extraer conclusiones útiles a partir de conjuntos de datos grandes y diversos mediante la exploración, la predicción y la inferencia (Adhikari et al., 2022). Estas funciones constituyen la base del trabajo del científico de datos y permiten transformar datos crudos en conocimiento útil.

Exploración

Es imposible hablar de ciencia de datos sin antes explorar los datos en sí mismos. Los datos pueden provenir de múltiples fuentes y presentarse en diversas formas; en esencia, son colecciones de texto, números o símbolos que por sí solos carecen de

significado. Para que aporten información, requieren ser procesados o interpretados en un contexto específico (Cambridge International Examinations, 2015).

Para llevar a cabo un proceso exploratorio riguroso, es fundamental contar con datos estructurados, es decir, organizados en filas y columnas, como en una tabla (Bruce y Bruce, 2017). De forma convencional, las variables o características se representan en columnas, mientras que las unidades de análisis se ubican en filas (Kroese et al., 2024).

Una vez estructurados, se aplican técnicas estadísticas que permiten explorar, describir y resumir la naturaleza de los datos. Estas técnicas facilitan identificar patrones y comprender las relaciones entre variables, con el objetivo de maximizar la interpretación y generar conocimiento accionable (Ministerio de Asuntos Económicos y Transformación Digital, 2021).

Predicción

Para abordar la predicción es necesario introducir el concepto *de machine learning*, entendido como un conjunto de herramientas de propósito general diseñadas para generalizar a partir de datos. Estos algoritmos han demostrado ser capaces de resolver problemas en múltiples disciplinas sin requerir un conocimiento profundo del dominio específico (Blum et al., 2018).

Los algoritmos de *machine learning* se clasifican, en su mayoría, en aprendizaje supervisado y aprendizaje no supervisado. En el primero, existe un atributo objetivo, y cada instancia del conjunto de datos incluye tanto los valores de entrada como el valor de salida, lo que permite que el modelo aprenda la función que los relaciona. En el segundo, no hay atributo objetivo definido; el algoritmo debe identificar patrones o regularidades en los datos para agrupar o estructurar la información de manera significativa (Kelleher y Tierney, 2018).

Inferencia

Busca que las conclusiones obtenidas a partir de un conjunto limitado de datos puedan aplicarse a un contexto más amplio. Para ello, se emplean técnicas estadísticas como la significancia estadística, las pruebas de hipótesis y los intervalos de confianza, que permiten extrapolar los resultados de la muestra hacia la población de interés (Bruce y Bruce, 2017).

En otras palabras, la inferencia tiene como objetivo explicar y cuantificar relaciones: identificar qué variables son relevantes, estimar la magnitud de sus efectos y determinar si los cambios observados son estadísticamente significativos.

4.6 Marcos metodológicos en ciencia de datos para proyectos predictivos

Los marcos metodológicos en ciencia de datos proporcionan una estructura ordenada para planificar y ejecutar proyectos predictivos. Facilitan la definición de objetivos, la selección de datos y la evaluación de modelos, asegurando que el proceso sea riguroso, reproducible y alineado con la meta del negocio o investigación. De los marcos metodológicos disponibles, este trabajo adoptó la metodología CRISP-DM, que se describe a continuación, aunque sin abordar su fase final de despliegue, dado que el alcance de este trabajo no contempla la puesta en producción ni el monitoreo continuo del modelo.

CRISP-DM (Cross Industry Standard Process for Data Mining) es un modelo de proceso abierto desarrollado por SPSS y Teradata en 1996 que presenta un proceso estructurado, bien definido y ampliamente documentado. Divide el ciclo de vida de un proyecto de minería de datos en seis fases: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue (Martinez et al., 2021). Wirth y Hipp (2000) dan una breve explicación de cada una de estas fases, que se presentan en la Tabla 1.

Tabla 1
Descripción de las fases de la metodología CRISP-DM

Fase	Objetivo
Comprensión del negocio	Definir los objetivos del proyecto y convertirlos en un problema de minería de datos.
Comprensión de los datos	Recolectar, explorar y evaluar la calidad de los datos para generar hipótesis.
Preparación de los datos	Seleccionar, limpiar, transformar y crear variables para construir el conjunto de datos final.
Modelado	Elegir técnicas, ajustar parámetros y construir modelos óptimos.
Evaluación	Verificar que los modelos cumplen los objetivos del negocio y decidir su uso.

Nota. Esta tabla resume los objetivos de cada una de las fases de la metodología CRISP-DM a partir de lo descrito por Wirth y Hipp (2000).

4.7 Modelos predictivos y métricas de desempeño

En ciencia de datos, la selección de un modelo predictivo, su validación y la medición de su desempeño son pasos fundamentales para garantizar resultados confiables y útiles para la toma de decisiones. A continuación, se describen los modelos de aprendizaje automático efectivamente empleados en este estudio, agrupados según su lógica de funcionamiento.

Los árboles de decisión son otro tipo de modelo relevante, aplicable tanto a problemas de clasificación como de regresión. Representan el proceso de decisión mediante una estructura jerárquica que divide el espacio de las variables predictoras en regiones sucesivamente más simples, ya sea para asignar una categoría o, como en el

caso de este estudio, para estimar un valor numérico continuo (OpenStax, 2025). Entre los métodos más conocidos se encuentra Random Forest, que combina múltiples árboles de decisión entrenados con datos y variables seleccionados aleatoriamente y promedia sus predicciones. Este enfoque de ensamblado por agregación (bagging) genera modelos diversos y reduce la varianza y el riesgo de sobreajuste, lo que mejora la precisión y la estabilidad de las predicciones (Bentéjac et al., 2021).

Otro método relevante es el gradient boosting, una técnica de *ensemble learning* que construye un modelo robusto combinando varios modelos débiles de forma iterativa. En cada iteración, aprende de los errores del modelo previo minimizando una función de pérdida de gradientes, lo que mejora progresivamente el desempeño predictivo (Kurama, 2025). Una de sus implementaciones más populares es XGBoost, que crea árboles de decisión de forma secuencial para corregir errores anteriores y optimizar la función de pérdida utilizando los residuos entre los valores reales y predichos (Kavlakoglu y Russi, s.f.). Otra implementación destacada es LightGBM, una versión eficiente de gradient boosting que acelera el entrenamiento mediante la selección inteligente de observaciones (GOSS) y la reducción del espacio de características (EFB), manteniendo un desempeño predictivo competitivo (Ke et al., 2017).

Los métodos de deep learning, y en particular las redes neuronales recurrentes, están diseñados para modelar secuencias en las que el orden de las observaciones es informativo, como ocurre en una serie temporal. A diferencia de una red neuronal convencional, una red recurrente mantiene un estado interno que se actualiza en cada paso de tiempo, lo que le permite, en principio, conservar información de observaciones pasadas para informar la predicción del paso siguiente (International Business Machines [IBM], s.f.). La arquitectura Long Short-Term Memory (LSTM) incorpora un mecanismo de compuertas que regula qué información se conserva, se olvida o se actualiza en el estado

interno, lo que mitiga el problema de desvanecimiento del gradiente que afecta a las redes recurrentes simples en secuencias largas (Hochreiter y Schmidhuber, 1997). La arquitectura Gated Recurrent Unit (GRU) simplifica este mecanismo de compuertas, combinando las funciones de olvido y actualización en una sola, lo que reduce el número de parámetros a estimar y puede resultar ventajoso cuando la serie de entrenamiento es relativamente corta (Cho et al., 2014).

Prophet es un modelo aditivo estructural desarrollado para el pronóstico de series temporales de negocio, que descompone la serie en una componente de tendencia, una componente estacional modelada mediante series de Fourier y, cuando están disponibles, regresores externos adicionales (Taylor y Letham, 2018). A diferencia de los modelos de caja negra, cada componente de Prophet puede inspeccionarse por separado, lo que facilita su interpretación por audiencias no especializadas, a costa de una estructura más rígida que la de los modelos de árboles o de las redes neuronales.

En cuanto a la evaluación del desempeño de los modelos, las métricas se dividen en métricas de regresión y métricas de clasificación, que se presentan en la Tabla 2 y la Tabla 3, respectivamente. Dado que este estudio aborda exclusivamente un problema de pronóstico (regresión) y no contempla ninguna tarea de clasificación, las métricas de la Tabla 2 son las que efectivamente se aplican, por igual, a los nueve modelos evaluados en este estudio; las de la Tabla 3 se incluyen únicamente como referencia conceptual del marco teórico.

Tabla 2

Resumen de las métricas de desempeño para regresión

Métrica	**Definición**
Raíz del Error Cuadrático Medio (RMSE)	Raíz cuadrada del promedio de los errores al cuadrado en una regresión. Es la métrica más utilizada para comparar modelos de regresión.

Métrica	Definición
Error Estándar Residual (RSE)	Igual que el RMSE, pero ajustado por los grados de libertad.
R-cuadrado (R^2)	Proporción de la varianza de la variable dependiente explicada por el modelo.
Error Absoluto Medio (MAE)	Promedio de los valores absolutos de los errores. Es menos sensible a los valores atípicos que el RMSE.
Error Porcentual Absoluto Medio (MAPE)	Promedio de los errores expresados como porcentaje del valor real; facilita la comparación entre series de distinta escala.

Nota. Elaboración propia a partir de Practical Statistics for Data Scientists (Bruce y Bruce, 2017).

Tabla 3

Resumen de las métricas de desempeño para clasificación

Métrica	Definición
Exactitud	Proporción de casos clasificados correctamente.
Matriz de confusión	Tabla que muestra el conteo de registros según su clasificación real y predicha.
Sensibilidad (Recall)	Proporción de casos positivos correctamente clasificados.
Especificidad	Proporción de casos negativos correctamente clasificados.
Precisión	Proporción de predicciones positivas que son realmente positivas.

Métrica	Definición
F1-Score	Medida armónica de la precisión y la sensibilidad.

Nota. Elaboración propia a partir de Practical Statistics for Data Scientists (Bruce y Bruce, 2017) y Machine Learning Aplicado al Análisis del Rendimiento de Desarrollos de Software (Gil-Vera y Seguro-Gallego, 2022). Estas métricas no se aplican en el desarrollo del presente estudio, al no contemplarse ninguna tarea de clasificación.

4.8 Interpretabilidad de modelos: valores SHAP

Muchos de los modelos de machine learning se consideran modelos de caja negra, ya sea porque son difíciles de comprender para un ser humano o porque la función que emplean es propietaria, esto es, está sujeta a derechos comerciales o legales (Rudin, 2019). Por otro lado, los modelos interpretables son aquellos que permiten a un usuario predecir de forma eficiente y correcta los resultados de dicho modelo (Kim et al., 2016).

Para apoyar la interpretabilidad de los modelos de tipo caja negra se proponen los valores SHAP (SHapley Additive exPlanations), una medida unificada de la importancia de las características que atribuye a cada una su contribución al resultado de la predicción (Lundberg y Lee, 2017). Estos valores adaptan el concepto de valor de Shapley, propio de la teoría de juegos cooperativos (Shapley, 1953), en el que cada variable se trata como un «jugador» cuyo aporte a la predicción se calcula promediando su contribución marginal sobre las posibles combinaciones de variables.

4.9 Validación temporal y fuga de información

En clasificación y regresión, la evaluación de los modelos suele apoyarse en funciones de pérdida cuadrática, pues se asume una distribución normal del error, y en la validación cruzada como procedimiento estándar para la selección de modelos. Sin embargo, este enfoque no puede aplicarse a las series de tiempo, ya que los datos presentan dependencia y pueden estar sujetos a efectos que evolucionan con el tiempo (Bergmeir y Benítez, 2012).

Para ello se propone el uso de evaluaciones con origen móvil, que permiten dividir la serie respetando su orden cronológico, obtener distribuciones de error diferenciadas según el horizonte de predicción y reducir la sensibilidad de las medidas de error (Tashman, 2000). De este modo se reproducen las condiciones reales del pronóstico y se evita la fuga de información (*data leakage*), entendida como el uso, durante el entrenamiento, de información que no estaría disponible al momento de predecir. Esta fuga es indeseable, pues constituye una fuente de mala generalización y conduce a sobreestimar el desempeño de los modelos.

4.10 Estado del arte en modelos de predicción del homicidio intencional

La predicción de homicidios intencionales ha sido un campo de creciente interés en la criminología cuantitativa y en la ciencia de datos aplicada a la seguridad ciudadana. A nivel global, los primeros esfuerzos estuvieron centrados en modelos estadísticos clásicos, principalmente de series temporales, orientados a comprender la evolución de la criminalidad en el tiempo. Box y Tiao (1975) y, posteriormente, Box et al. (2015) sentaron las bases para el análisis de tendencias, estacionalidades y choques externos en fenómenos sociales. Con el tiempo, los avances en ciencia de datos y en modelos estadísticos han abierto nuevas posibilidades para comprender los patrones subyacentes de la violencia letal y, en consecuencia, aportar a la formulación de políticas de prevención y control (UNODC, 2023; Becker, 1968; Cohen y Felson, 1979).

En América Latina, investigaciones como las de Sánchez Jabba et al. (2012) y Sandoval (2014) han explorado la evolución geográfica del homicidio en Colombia, señalando que la violencia letal tiene una marcada concentración territorial y temporal. Estudios más recientes aplican modelos espaciales y multinomiales para analizar fenómenos asociados, como la mortalidad por armas de fuego en Ecuador (Sosa et al., 2023) o la relación entre violencia y factores sociodemográficos en Brasil (Soares Filho et

al., 2024). Estos trabajos confirman la pertinencia de captar la multicausalidad y la heterogeneidad regional de la violencia homicida.

La incorporación de machine learning en este campo ha representado un salto cualitativo. Ordóñez-Eraso et al. (2020) aplicaron algoritmos de clasificación y detección de patrones a los homicidios en Colombia, logrando identificar tendencias que pasaban desapercibidas con métodos tradicionales. De manera similar, Zhang et al. (2020) realizaron una comparación entre algoritmos como Random Forest, Gradient Boosting y Support Vector Machines para predecir hotspots criminales, encontrando que los métodos de ensamblado ofrecen ventajas en precisión frente a enfoques lineales.

En Colombia, los aportes recientes de Pérez et al. (2024) y Arbona et al. (2024) refuerzan el vínculo entre homicidios, factores socioeconómicos y educativos, utilizando analítica de datos para comprender las interacciones entre violencia y desarrollo. Igualmente, investigaciones del Observatorio de la Unidad para las Víctimas (2023) y del Instituto Nacional de Medicina Legal y Ciencias Forenses (2022) evidencian la relevancia de integrar bases de datos oficiales con técnicas de predicción para mejorar la planeación en seguridad pública.

En síntesis, el estado del arte evidencia una convergencia entre tres tradiciones: (i) los modelos estadísticos de series temporales que priorizan la interpretabilidad, (ii) los algoritmos de machine learning que maximizan la capacidad predictiva, y (iii) los enfoques híbridos que buscan equilibrar precisión y explicabilidad. En el caso colombiano, donde la violencia homicida responde a múltiples factores sociales, económicos y geográficos, esta convergencia resulta estratégica para diseñar modelos predictivos más robustos y útiles para la política pública.

5. Hipótesis

Las hipótesis que se formulan a continuación operacionalizan los objetivos específicos de naturaleza comparativa y explicativa de este trabajo. El objetivo específico de construir la serie temporal a partir del SIEDCO no se traduce en una hipótesis, dado que corresponde a un componente metodológico y constructivo del trabajo, cuyo cumplimiento se verifica directamente en la metodología. En este sentido, las hipótesis se plantean como proposiciones susceptibles de verificación empírica, las cuales orientan el análisis comparativo y la interpretación de los resultados obtenidos.

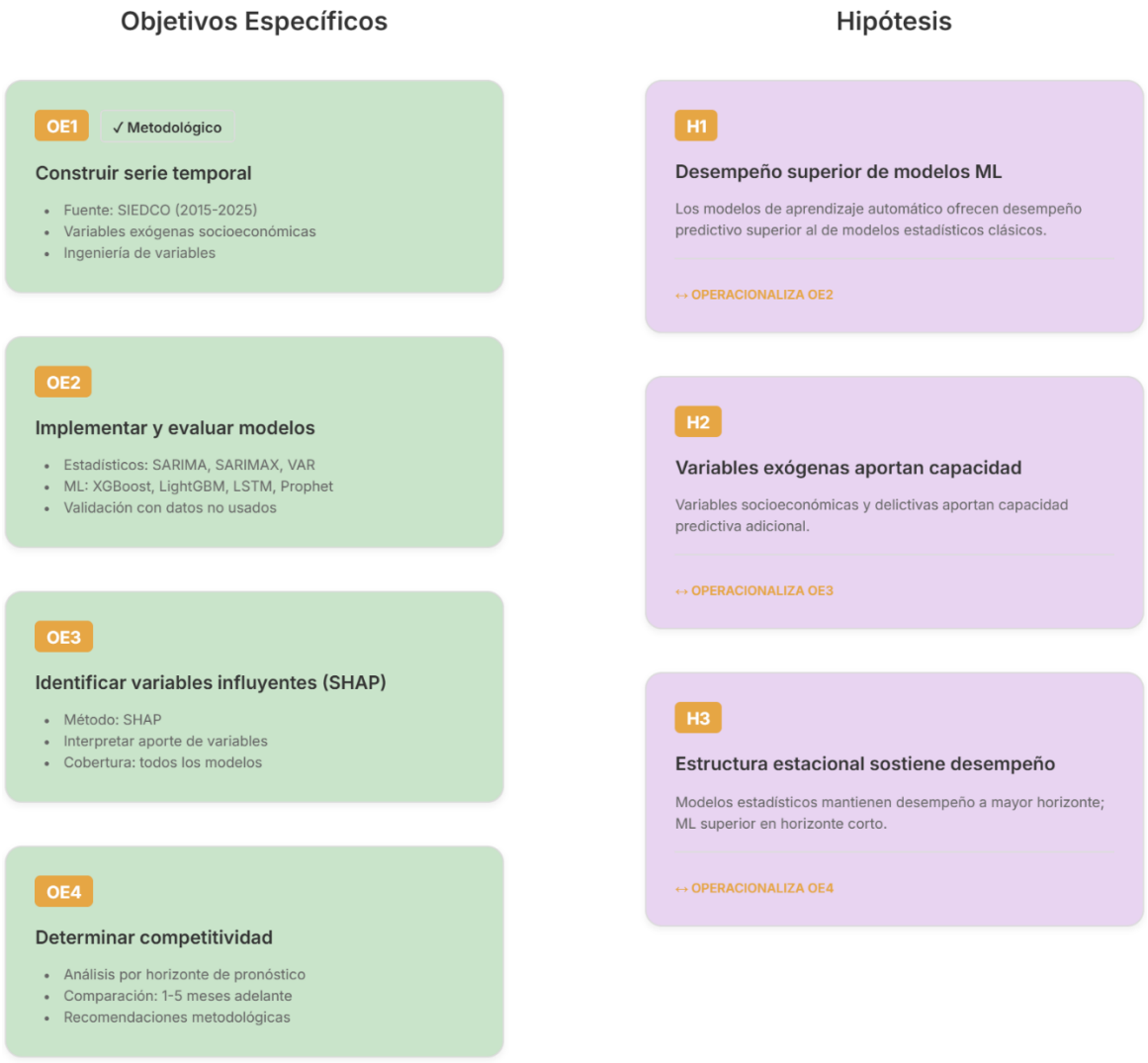
Hipótesis 1. Dada su mayor flexibilidad para capturar relaciones no lineales, los modelos de aprendizaje automático ofrecen un desempeño predictivo superior al de los modelos estadísticos clásicos en el pronóstico mensual del homicidio intencional a escala nacional.

Hipótesis 2. Las variables exógenas de naturaleza socioeconómica y delictiva aportan capacidad predictiva adicional sobre el pronóstico del homicidio intencional, más allá de la información contenida en la propia historia de la serie.

Hipótesis 3. Dentro de la ventana de prueba, el error de pronóstico de los modelos estadísticos con estructura estacional explícita crece con el horizonte (de uno a cinco meses) de manera más moderada que el de los modelos de aprendizaje automático, mientras que estos últimos resultan relativamente más competitivos en el horizonte más corto (un mes), donde su mayor sensibilidad a la información reciente de la serie resulta ventajosa.

La Figura 2 sintetiza la relación entre los objetivos específicos planteados en la Sección 2 y las hipótesis de investigación formuladas en esta sección, indicando qué hipótesis opera cada objetivo.

Figura 2
Relación entre objetivos específicos e hipótesis de investigación



Nota. Elaboración propia.

En la propuesta inicial de este trabajo, la Hipótesis 3 se formuló para un diseño con dos horizontes de evaluación separados: un holdout histórico de doce meses y un horizonte prospectivo de un mes. Durante el desarrollo, ese diseño se reemplazó por una única ventana de prueba prospectiva de cinco meses continuos (enero a mayo de 2026), por dos razones: se privilegió la evaluación frente a datos reales no observados sobre la evaluación frente a un holdout histórico, y reservar doce meses adicionales de 2025 como holdout habría reducido de forma apreciable el periodo de entrenamiento de una serie de 132 observaciones. En consecuencia, la Hipótesis 3 se reformuló en los términos anteriores, verificables con el diseño de ventana única mediante el error de pronóstico por horizonte (Tabla 44). La formulación original, que contrasta explícitamente un horizonte corto y uno largo, se conserva como línea de trabajo futuro bajo un esquema de validación de origen móvil.

6. Variables

Las variables de este estudio se derivan de las hipótesis planteadas y se organizan en torno al problema de pronóstico. Al tratarse de un estudio con propósito predictivo, la relación entre las variables adopta una estructura de dependencia: una variable de respuesta, cuyo comportamiento se busca pronosticar, y un conjunto de variables predictoras a partir de las cuales se genera el pronóstico. La Tabla 4 presenta la definición conceptual y operacional de cada una.

Tabla 4
Operacionalización de las variables

Variable	Tipo	Definición conceptual	Definición operacional
Homicidio intencional mensual	Dependiente	Magnitud mensual del fenómeno del homicidio intencional a escala nacional que se busca pronosticar.	Número total de víctimas de homicidio intencional por mes, como suma del campo Cantidad de SIEDCO (temática homicidios), enero 2015 a diciembre 2025.
Rezagos de la serie	Independiente (temporal)	Memoria histórica del fenómeno: información que el pasado reciente de la serie aporta sobre su comportamiento futuro.	Valores de la variable dependiente en los doce meses anteriores (rezagos 1 a 12).
Mes del año	Independiente (temporal)	Patrón estacional del homicidio a lo largo del año.	Número del mes (1 a 12) asociado a cada observación.

Variable	Tipo	Definición conceptual	Definición operacional
Hurto a personas	Independiente (exógena)	Actividad delictiva contra el patrimonio como factor de contexto asociado a la violencia.	Serie mensual nacional de hurto a personas (SIEDCO), rezagada un mes.
Lesiones personales	Independiente (exógena)	Violencia interpersonal no letal como factor de contexto.	Serie mensual nacional de lesiones personales (SIEDCO), rezagada un mes.
Violencia intrafamiliar	Independiente (exógena)	Violencia en el ámbito doméstico como factor de contexto social.	Serie mensual nacional de violencia intrafamiliar (SIEDCO), rezagada un mes.
Incautación de estupefacientes	Independiente (exógena)	Actividad de control sobre economías ilegales asociadas a la violencia.	Serie mensual nacional de incautación de estupefacientes (SIEDCO), rezagada un mes.
Índice de Seguimiento a la Economía (ISE)	Independiente (exógena)	Condiciones de la actividad económica como factor estructural del contexto.	Serie mensual del ISE (DANE), rezagada un mes.

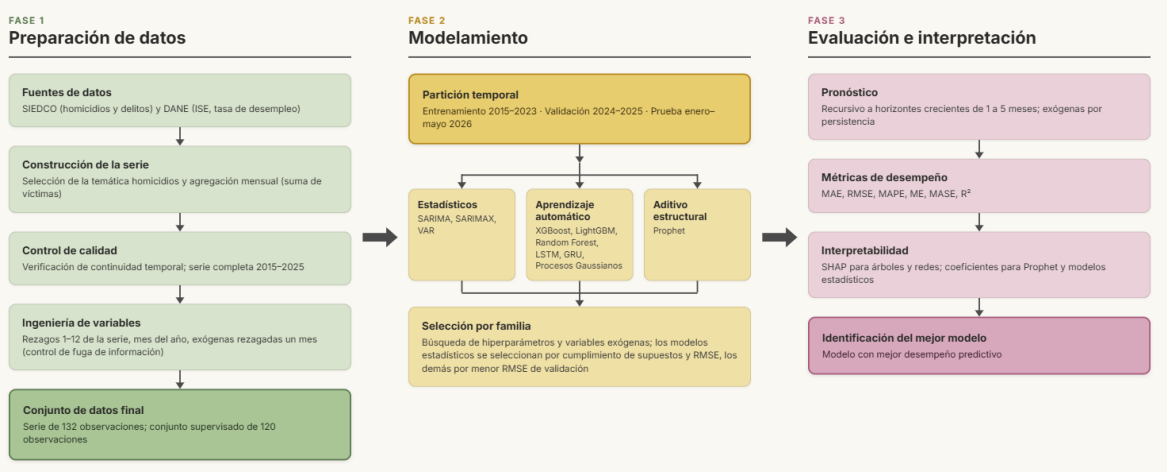
Variable	Tipo	Definición conceptual	Definición operacional
Tasa de desempleo	Independiente (exógena)	Condiciones del mercado laboral como factor socioeconómico asociado a la violencia.	Serie mensual de la tasa de desempleo (DANE), rezagada un mes.

Nota. Las variables independientes se agrupan en dos conjuntos: las temporales, derivadas de la propia serie, y las exógenas, provenientes de fuentes externas. Elaboración propia.

7. Metodología

El presente capítulo describe el proceso metodológico seguido en la investigación para construir, modelar y evaluar una serie temporal del homicidio intencional en Colombia. El proceso se organiza en tres fases sucesivas: la preparación de los datos, que abarca la construcción de la serie y de las variables predictoras; el modelamiento, que comprende la especificación, el ajuste y la selección de las distintas familias de modelos; y la evaluación e interpretación, que contrasta el desempeño de los modelos y analiza la influencia de las variables. La Figura 3 presenta una visión general de este proceso, cuyas etapas se detallan en los apartados siguientes.

Figura 3
Marco metodológico del estudio



Nota. Elaboración propia.

7.1 Enfoque y alcance de la investigación

La presente investigación adopta un enfoque cuantitativo, en tanto el fenómeno de estudio (el homicidio intencional en Colombia) se aborda a través de su medición numérica a lo largo del tiempo y del contraste de hipótesis mediante métricas de error. El objeto de análisis es una serie temporal de conteos mensuales, y las conclusiones se derivan del

desempeño predictivo de los modelos, expresado en indicadores cuantitativos, y no de la interpretación de significados o experiencias.

El diseño de la investigación es no experimental, dado que no existe manipulación de las variables. El estudio se realiza sobre registros administrativos ya generados por el Sistema de Información de la Policía Nacional, sin intervención del investigador sobre las condiciones en que ocurre el fenómeno. Es, además, de carácter longitudinal, puesto que la variable objetivo se observa de forma repetida a lo largo de una sucesión de periodos mensuales, lo que permite estudiar su evolución temporal, su estacionalidad y su tendencia.

En cuanto al alcance, el estudio es de tipo descriptivo-correlacional con propósito predictivo. Es descriptivo porque caracteriza el comportamiento del homicidio intencional a nivel nacional en el periodo analizado; es correlacional porque examina la relación entre la variable objetivo y un conjunto de variables temporales y socioeconómicas; y tiene propósito predictivo porque, a partir de esa estructura, construye y compara modelos orientados al pronóstico de la serie.

El estudio se circunscribe a la serie nacional agregada y no aborda la dimensión espacial ni la desagregación territorial del fenómeno. Esta delimitación obedece al propósito de establecer una línea base metodológica que permita comparar el desempeño de las distintas familias de modelos sobre una serie única y consistente, de modo que futuros estudios puedan incorporar la variabilidad territorial sobre una referencia previamente establecida.

7.2 Fuente de datos y evaluación de su calidad

La información utilizada en esta investigación proviene del Sistema de Información Estadístico, Delincuencial, Contravencional y Operativo de la Policía Nacional (SIEDCO), fuente oficial de registro administrativo de la actividad delictiva en Colombia. La base

consolidada reúne 3.785.395 registros correspondientes a distintas temáticas delictivas, cada uno de los cuales representa un evento y se describe mediante dieciocho campos, entre los que se encuentran la temática del delito, la fecha del hecho (año, mes y día), su localización (departamento, código DANE del municipio, municipio, zona y clase de sitio), las características de la víctima (género y grupo de edad), el arma o medio empleado y la cantidad de víctimas.

Para efectos de esta investigación, la temática de interés es HOMICIDIOS, que en SIEDCO corresponde exclusivamente a homicidio intencional; el sistema no incluye bajo esta categoría las muertes por accidente de tránsito ni el homicidio culposo. Al restringir la base a esta temática para el periodo 2015 a 2025 se obtienen 110.381 registros, equivalentes a 141.774 víctimas de homicidio. La diferencia entre ambas cifras se explica porque un mismo evento puede involucrar más de una víctima: el campo Cantidad registra el número de personas fallecidas en cada hecho, y alcanza valores superiores a uno en eventos como las masacres. En consecuencia, la variable objetivo de este estudio es el número mensual de víctimas de homicidio, obtenido como la suma del campo Cantidad para cada mes, y no el número de eventos registrados.

De los campos disponibles en la fuente, la mayoría se emplearon en la caracterización descriptiva del fenómeno y no como predictores del modelo. Esta decisión responde al alcance del estudio: dado que el pronóstico se realiza sobre la serie nacional agregada a nivel mensual, las variables de desagregación geográfica, demográfica o situacional no intervienen en la predicción, aunque resultan pertinentes para la descripción del fenómeno. Cabe precisar que cinco de los dieciocho campos corresponden a versiones auxiliares (cruda, estandarizada y normalizada) de los atributos Armas y Género, generadas durante la depuración de dichos campos de texto; no constituyen información adicional, sino registro del proceso de limpieza.

Las variables predictoras de naturaleza socioeconómica y delictiva se incorporan como series mensuales nacionales independientes, provenientes de fuentes oficiales: hurto a personas, lesiones personales, violencia intrafamiliar e incautación de estupefacientes, procedentes del propio SIEDCO, así como el Índice de Seguimiento a la Economía (ISE) y la tasa de desempleo, publicados por el DANE. Todas comparten la periodicidad mensual y el ámbito nacional de la variable objetivo, condición que permite su integración en un mismo modelo multivariado. La elección de la periodicidad mensual responde precisamente a esta necesidad: los indicadores socioeconómicos empleados como predictores se publican con frecuencia mensual y no admiten una desagregación temporal mayor, por lo que el mes constituye la mínima unidad temporal común que permite integrar la variable objetivo con el conjunto de predictores.

Al tratarse de un registro administrativo, SIEDCO presenta las limitaciones propias de este tipo de fuentes. El dato se origina en la actividad de denuncia y registro policial, de modo que su cobertura depende de que cada hecho sea efectivamente reportado y consolidado en el sistema. Esta característica lo distingue de las fuentes de base forense, como el sistema Forensis del Instituto Nacional de Medicina Legal y Ciencias Forenses, cuyo registro se origina en el dictamen médico legal de la causa de muerte. Dado que ambas fuentes capturan el mismo fenómeno por vías institucionales distintas, su comparación permite evaluar la confiabilidad de SIEDCO como insumo para el estudio.

Con este propósito se contrastaron los totales anuales de homicidios de SIEDCO frente a los reportados por Forensis durante el periodo 2015 a 2024, único intervalo en el que ambas fuentes se encuentran disponibles. Para que la comparación sea válida, en ambos casos se emplea el número de víctimas, dado que Forensis contabiliza personas fallecidas; los totales de SIEDCO corresponden, por tanto, a la suma del campo Cantidad y no al conteo de registros. Los resultados se presentan en la Tabla 5.

Tabla 5
Comparación de cantidad de homicidios según SIEDCO y Forensis

Año	SIEDCO	Forensis	Diferencia absoluta	Diferencia relativa
2015	12.334	11.585	+749	+6,5%
2016	12.045	11.532	+513	+4,4%
2017	11.957	11.373	+584	+5,1%
2018	12.573	12.130	+443	+3,7%
2019	12.547	11.880	+667	+5,6%
2020	12.009	11.326	+683	+6,0%
2021	13.685	13.238	+447	+3,4%
2022	13.536	13.939	−403	−2,9%
2023	13.555	14.260	−705	−4,9%
2024	13.496	14.022	−526	−3,8%

Nota. Elaboración propia a partir de los registros de SIEDCO (Policía Nacional) y de los informes Forensis del Instituto Nacional de Medicina Legal y Ciencias Forenses (2017, 2019, 2021, 2022, 2023, 2024).

El contraste evidencia una fuerte concordancia entre las dos fuentes. La correlación entre las series anuales es de 0,955, lo que indica que ambas reproducen prácticamente la misma dinámica temporal del fenómeno. La diferencia promedio entre ellas es de 572 casos por año, equivalente al 4,6% del total, una magnitud reducida y estable que resulta esperable entre dos sistemas de registro con orígenes institucionales diferentes.

Se observa, además, un cambio en el signo de la diferencia a partir de 2022: hasta 2021 SIEDCO reporta cifras ligeramente superiores a las de Forensis, mientras que desde 2022 ocurre lo contrario. Esta observación no altera la conclusión general sobre la consistencia de las fuentes, pero sugiere posibles diferencias en los tiempos de consolidación del dato entre ambos sistemas en los años más recientes.

La alta concordancia entre SIEDCO y una fuente independiente de base forense respalda el uso de SIEDCO como insumo confiable para la construcción de la serie objetivo. Dado que la validación se realiza sobre el periodo común disponible (2015 a 2024) y arroja resultados consistentes, se considera que la confiabilidad establecida se extiende razonablemente a los años 2025 y 2026, para los cuales solo se dispone del registro de SIEDCO.

7.3 Población y delimitación temporal

El estudio no se apoya en un procedimiento de muestreo, sino que trabaja con la totalidad de los registros de homicidio intencional disponibles en la fuente para el periodo analizado; se trata, por tanto, de un censo y no de una muestra. La población de estudio corresponde a la totalidad de las víctimas de homicidio intencional registradas en SIEDCO a nivel nacional entre enero de 2015 y diciembre de 2025.

La unidad de análisis es el agregado mensual nacional: cada observación de la serie corresponde al número total de víctimas de homicidio intencional registradas en el país durante un mes. Bajo esta definición, el periodo de estudio comprende 132 observaciones mensuales consecutivas, sin interrupciones, entre enero de 2015 y diciembre de 2025.

El periodo de análisis de los modelos estadísticos se inicia en 2015. Esta delimitación responde a la estabilidad del patrón estacional y del nivel de la serie en la última década, documentada en el análisis exploratorio (Sección 8.2), condición que favorece la identificación de una única especificación SARIMA mediante la metodología Box-Jenkins. Las series exógenas y la serie de homicidios se encuentran disponibles con periodicidad mensual desde enero de 2005, y los modelos de aprendizaje automático, cuya capacidad para capturar cambios de régimen no depende de la estabilidad de una estructura paramétrica única, se entrenaron sobre ese histórico más largo (Sección 7.4).

Esta asimetría en las ventanas de entrenamiento se declara como una limitación de la comparación en la Sección 7.7.

Adicionalmente, se reservó un conjunto de datos correspondiente a los meses de enero a mayo de 2026, constituido por los valores efectivamente observados de homicidios en ese periodo, que se emplea como prueba final en un escenario de operación real. La partición de las 132 observaciones entre entrenamiento y validación, así como el papel de este conjunto de prueba, se detallan en la Sección 7.7.

7.4 Preparación de datos e ingeniería de variables

Depuración y construcción de la serie

La construcción de la variable objetivo parte de la base de eventos de SIEDCO. Dado que el campo Temática es un valor categórico controlado por el sistema, la identificación de los registros de homicidio intencional se realiza por selección directa de la categoría correspondiente, sin que se requiera una etapa previa de normalización de texto sobre este campo. Una vez seleccionados los registros de la temática de interés, se agregan por mes sumando el campo Cantidad, con lo que se obtiene la serie mensual nacional de víctimas de homicidio.

Sobre la serie resultante se verificó la continuidad temporal, comprobando que no existieran meses ausentes en el rango de estudio. Para ello, la serie se indexó con frecuencia mensual explícita, lo que permite detectar cualquier interrupción en la secuencia de periodos. La verificación confirmó que la serie de homicidios se encuentra completa entre enero de 2015 y diciembre de 2025, sin huecos temporales. Las series exógenas se integraron a esta estructura por fecha, conformando un conjunto de datos mensual alineado en el que cada periodo dispone del valor de la variable objetivo y de los predictores.

Ingeniería de variables

A partir de la serie mensual se construyó el conjunto de variables predictoras empleado por los modelos de aprendizaje automático. Este conjunto se compone de tres tipos de variables:

a. Los rezagos de la variable objetivo, correspondientes a los doce meses anteriores (de un mes hasta doce meses de retraso). La inclusión de doce rezagos permite a los modelos capturar tanto la dependencia de corto plazo como el ciclo anual característico de la serie.
b. El rezago de un mes de cada variable exógena, de modo que el valor de cada predictor socioeconómico y delictivo que entra al modelo corresponde al mes inmediatamente anterior al que se pronostica.
c. El mes del año, incorporado como variable numérica, con el fin de representar el patrón estacional.

La decisión de incorporar las variables exógenas con un rezago de un mes, y no con su valor contemporáneo, constituye el mecanismo central de control de fuga de información: el modelo pronostica el número de homicidios de un mes determinado utilizando exclusivamente información disponible hasta el mes anterior, lo que reproduce las condiciones de un escenario de pronóstico real, en el que el valor de los predictores del mes en curso todavía no se conoce.

Tratamiento de las observaciones iniciales

La construcción de rezagos de hasta doce meses implica que las primeras doce observaciones de la serie no disponen de un valor válido para todos los predictores, dado que carecen de historia previa suficiente. Estas observaciones se excluyeron del conjunto supervisado empleado por los modelos de aprendizaje automático. Dado que estos

modelos se alimentaron con las series mensuales disponibles desde enero de 2005, el conjunto supervisado comprende 240 observaciones (enero de 2006 a diciembre de 2025), de las cuales 216 corresponden al periodo de entrenamiento (hasta diciembre de 2023) y 24 al de validación (2024-2025). Los modelos estadísticos de series temporales (SARIMA, SARIMAX y VAR) operan sobre la serie 2015-2025 de 132 observaciones, ya que incorporan la estructura de dependencia temporal en su propia formulación y no requieren la construcción de rezagos como columnas independientes.

Escalamiento

En el conjunto supervisado empleado por los modelos de aprendizaje automático las variables predictoras no fueron sometidas a escalamiento ni estandarización. Esta decisión responde a la naturaleza de los modelos empleados: los modelos basados en árboles (XGBoost, LightGBM y Random Forest), que constituyen el núcleo de la aproximación de aprendizaje automático, son invariantes a transformaciones monótonas de las variables y, por tanto, insensibles a su escala. Los modelos estadísticos multivariados siguen un tratamiento distinto: en SARIMAX la serie de incautación de estupefacientes se transformó a logaritmo natural, para atenuar su asimetría, y las restantes variables exógenas se estandarizaron a media cero y desviación estándar uno con los parámetros del histórico completo (Sección 8.3.2). Conviene advertir que, al estimarse esos parámetros sobre la totalidad de la serie, la etapa de selección de especificaciones opera con una estandarización que incorpora información del tramo de validación; el efecto es menor, porque afecta a la escala y no al valor de la variable objetivo, pero constituye una salvedad al control de fuga descrito en la Sección 7.7. Las redes recurrentes (LSTM y GRU), en cambio, recibieron las variables sin escalar, condición que se retoma al interpretar su desempeño en la Sección 8.4.5.

La Figura 4 detalla la secuencia completa del proceso de preparación de los datos, desde la base de eventos original hasta la conformación del conjunto de datos empleado en el modelamiento.

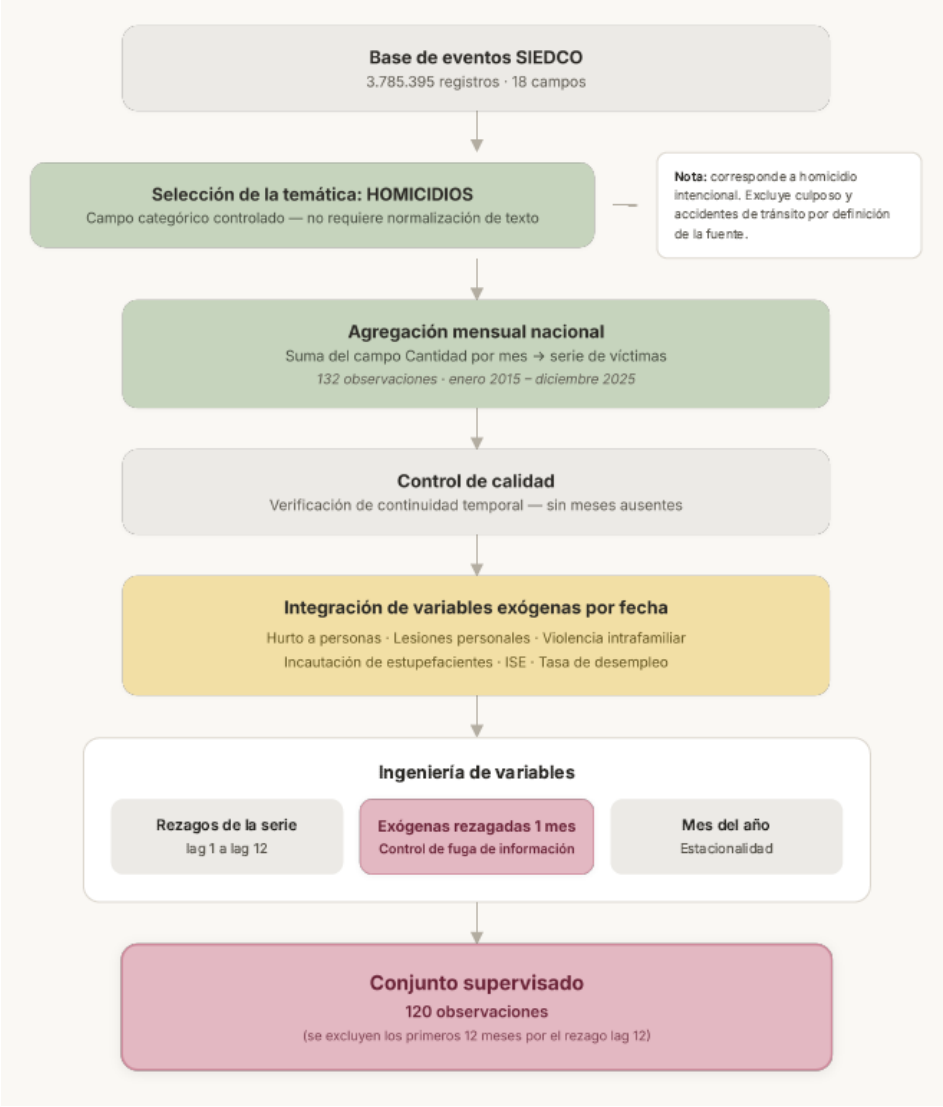
7.5 Especificación e implementación de los modelos

El estudio compara nueve modelos pertenecientes a tres familias: modelos estadísticos de series temporales (SARIMA, SARIMAX y VAR), modelos de aprendizaje automático basados en árboles y en redes neuronales (XGBoost, LightGBM, Random Forest, LSTM y GRU), y un modelo aditivo estructural (Prophet). Todos los modelos se ajustan sobre la misma serie y el mismo periodo, y se evalúan bajo el mismo esquema de validación, con el fin de garantizar la comparabilidad de los resultados.

Estrategia común de selección

Con excepción de los modelos estadísticos, cuya especificación se describe más adelante, todos los modelos de aprendizaje automático se ajustaron mediante un procedimiento común de selección que opera en dos niveles simultáneos.

Figura 4
Proceso de preparación de datos e ingeniería de variables



Nota. Elaboración propia.

En el primero se evalúan combinaciones del conjunto de variables exógenas, desde su ausencia total hasta su inclusión completa. En los modelos basados en árboles esta búsqueda es exhaustiva y cubre las 64 combinaciones posibles de las seis variables candidatas; en LSTM, GRU y Prophet, cuyo costo de entrenamiento es sustancialmente

mayor, se recorre un subconjunto informado que incluye la ausencia de exógenas, el conjunto completo y las combinaciones seleccionadas por los modelos de árboles. En el segundo, para cada combinación de exógenas, se recorre una malla de hiperparámetros específica del modelo. Cada configuración resultante se entrena y se evalúa según el esquema de validación temporal descrito en la Sección 7.7: el ajuste se realiza sobre el periodo de entrenamiento y la comparación entre configuraciones sobre el periodo de validación, midiendo el error cuadrático medio (RMSE).

La configuración con menor RMSE de validación se selecciona como definitiva y, una vez elegida, se reajusta sobre la totalidad de la serie histórica antes de generar el pronóstico del periodo de prueba. Este procedimiento asegura que la selección de variables y de hiperparámetros se realice exclusivamente sobre datos de validación, sin que el periodo de prueba intervenga en ninguna decisión de modelamiento, lo que preserva la validez de la evaluación final.

Mallas de hiperparámetros

Las mallas de búsqueda de cada modelo se presentan en la Tabla 6. Su amplitud se definió buscando un equilibrio entre la cobertura del espacio de configuraciones y el costo computacional, considerando el reducido tamaño de la serie.

Tabla 6
Mallas de hiperparámetros por modelo

Modelo	Hiperparámetros explorados
XGBoost	n_estimators {200, 400}; max_depth {3, 5}; learning_rate {0,05, 0,1}; subsample {0,8, 1,0}
LightGBM	n_estimators {200, 400}; max_depth {3, 5}; learning_rate {0,05, 0,1}; num_leaves {15, 31}

Random Forest	n_estimators {300, 500}; max_depth {5, 10, sin límite}; min_samples_leaf {1, 3}
LSTM / GRU	unidades {16, 32}; dropout {0,0, 0,2}; epochs {60}; batch_size {8}; learning_rate {0,01}
Prophet	changepoint_prior_scale {0,01, 0,1, 0,5}; seasonality_prior_scale {1,0, 10,0}

Nota. Elaboración propia.

Modelos basados en árboles

XGBoost, LightGBM y Random Forest reciben como entrada el conjunto supervisado descrito en la Sección 7.4 (los doce rezagos de la variable objetivo, los rezagos de las exógenas y el mes) y se entrenan mediante la estrategia común de selección. La semilla aleatoria se fijó en todos ellos para garantizar la reproducibilidad de los resultados.

Redes neuronales recurrentes (LSTM y GRU)

Ambos modelos comparten una misma arquitectura de dos ramas: una rama recurrente que procesa la secuencia de los doce rezagos de la variable objetivo, y una rama densa que procesa las variables exógenas y el indicador de mes. Las salidas de ambas ramas se combinan para producir el pronóstico. Esta arquitectura permite que la red trate los rezagos como una secuencia temporal ordenada, aprovechando la capacidad de las celdas recurrentes para modelar dependencias en el tiempo, y simultáneamente incorpore la información contextual de los predictores.

Prophet

El modelo Prophet se especificó con estacionalidad anual activada y sin estacionalidad semanal ni diaria, acordes con la periodicidad mensual de la serie. Las variables exógenas se incorporaron como regresores adicionales, y los parámetros de

regularización de los puntos de cambio y de la estacionalidad se ajustaron mediante la malla correspondiente.

Modelos estadísticos de series temporales

Los modelos SARIMA, SARIMAX y VAR siguen un procedimiento de especificación distinto, apropiado a su naturaleza. La identificación de los órdenes partió de la inspección de las funciones de autocorrelación simple y parcial de la serie diferenciada, que delimitan los rangos plausibles de los componentes autorregresivo y de media móvil, tanto en su parte regular como estacional, y sobre esos rangos se definió la malla de órdenes efectivamente recorrida. Las especificaciones de referencia derivadas de este análisis se presentan en la Tabla 7. A partir de ellas se recorrió una malla de 54 combinaciones de órdenes para SARIMA, todas con componente estacional de periodo doce. Para SARIMAX se retuvieron los diez órdenes con menor AIC de esa malla y cada uno se cruzó con las 64 combinaciones posibles de las seis variables exógenas candidatas, lo que arrojó 640 modelos ajustados por máxima verosimilitud.

Tabla 7
Especificaciones de referencia para la identificación de órdenes SARIMA y SARIMAX

Especificación	Orden (p,d,q)	Componente estacional (P,D,Q)[s]	Evaluada en
1	(1,0,0)	(2,1,2)[12]	SARIMA y SARIMAX
2	(0,1,1)	(1,1,1)[12]	SARIMA y SARIMAX
3	(1,1,0)	(1,1,1)[12]	SARIMA y SARIMAX
4	(2,0,0)	(1,1,1)[12]	SARIMA
5	(1,0,1)	(2,1,2)[12]	SARIMA
6	(2,1,2)	(1,1,1)[12]	SARIMA

Nota. Las especificaciones aquí relacionadas corresponden a los órdenes de partida sugeridos por la inspección de la ACF y la PACF, sobre los cuales se definieron los

rangos de la malla efectivamente recorrida; no constituyen el conjunto completo de modelos evaluados ni el listado de finalistas, que se presenta en la Tabla 10. La notación (p,d,q) corresponde a los órdenes autorregresivo, de diferenciación y de media móvil no estacionales; (P,D,Q)[s] a sus equivalentes estacionales con periodo estacional s = 12 meses.

La selección entre las especificaciones candidatas se rige por un criterio jerárquico que antepone la validez estadística al ajuste predictivo. En primer lugar, cada candidato se evalúa según el número de supuestos que satisface entre tres pruebas diagnósticas: independencia de los residuos (Ljung-Box), normalidad (Jarque-Bera) y homocedasticidad (ARCH). Solo en caso de empate en el número de supuestos cumplidos se recurre, como criterio de desempate, al menor valor de los criterios de información de Akaike (AIC) y bayesiano (BIC). Este orden de prioridad responde a que un modelo estadístico cuyos residuos no satisfacen los supuestos sobre los que se construye produce inferencias e intervalos de confianza no confiables, independientemente de su error puntual; se privilegia, por tanto, la solidez estadística del modelo antes que su ajuste medido por criterios de información. Cuando ninguna de las especificaciones candidatas satisface las pruebas de independencia y de normalidad (situación que se presentó efectivamente en los tres modelos estadísticos), la selección se resuelve por completo mediante AIC y BIC, y las limitaciones inferenciales que ello implica se advierten al presentar cada modelo.

En el caso del modelo VAR se fijó un orden de doce rezagos en todas las especificaciones, con el fin de cubrir un ciclo estacional completo, y se ajustaron las siete combinaciones posibles de los homicidios con una, dos o las tres variables sometidas a la prueba de causalidad de Granger (Índice de Seguimiento a la Economía, violencia intrafamiliar e incautación de estupefacientes), de las cuales dos resultaron significativas; la selección entre ellas se realizó mediante AIC y BIC, y la especificación final quedó conformada por el par homicidios e Índice de Seguimiento a la Economía.

Generación de pronósticos

El periodo de prueba abarca un horizonte de cinco meses (enero a mayo de 2026). El pronóstico no se evalúa únicamente al final de ese horizonte, sino de forma acumulativa a horizontes crecientes, de uno a cinco meses, lo que permite observar cómo se degrada la precisión de cada modelo a medida que se pronostica más lejos en el tiempo. Para los modelos de aprendizaje automático, la predicción se genera de forma recursiva: la predicción de cada mes se incorpora como insumo para construir los rezagos necesarios en la predicción del mes siguiente. En cuanto a las variables exógenas, dado que sus valores futuros no se conocen en un escenario de pronóstico real, se adoptó el supuesto de persistencia, empleando el último valor observado de cada predictor. Este supuesto, coherente con el control de fuga de información descrito en la Sección 7.4, condiciona la interpretación del desempeño de los modelos que dependen de variables exógenas y debe tenerse presente al analizar los resultados.

Las tres familias de modelos difieren en la forma de cuantificar la incertidumbre del pronóstico. SARIMA, SARIMAX y VAR derivan sus intervalos del 95 % de la estructura estocástica del modelo, propagando la varianza del error de innovación a lo largo del horizonte, lo que produce bandas que se ensanchan a medida que el pronóstico se aleja del último dato observado; su validez depende de los supuestos de normalidad e independencia de los residuales, que, como se documenta en la Sección 8.3, no se satisfacen plenamente. Prophet obtiene sus intervalos por simulación de la incertidumbre de la tendencia y del término de error. Los modelos basados en árboles y las redes recurrentes no postulan una distribución del error ni generan intervalos nativos; para ellos se construyó un intervalo aproximado del 95 % como el pronóstico puntual más o menos 1,96 veces la desviación estándar de los residuales del conjunto de validación (24 meses), constante para los cinco horizontes. Esta aproximación captura la dispersión histórica del

modelo, pero no incorpora la acumulación de error propia del pronóstico recursivo, la incertidumbre del supuesto de persistencia de las exógenas ni el sesgo del modelo, por lo que está centrada en el pronóstico puntual y resulta mal calibrada cuando el modelo presenta un sesgo sistemático. Su cobertura empírica sobre el periodo de prueba se reporta en la Sección 8.5 (Tabla 47).

7.6 Entorno de reproducibilidad

El desarrollo se realizó en su totalidad en lenguaje Python, en un entorno de cuaderno interactivo (Jupyter). Con el fin de garantizar la reproducibilidad de los resultados, se fijó una semilla aleatoria común (valor 42) sobre los generadores de números aleatorios de NumPy y de TensorFlow, y se propagó dicha semilla a cada uno de los modelos que incorporan componentes estocásticos, a través de su parámetro de estado aleatorio. Esta medida asegura que la partición de los datos, la inicialización de las redes neuronales y los procesos de muestreo de los modelos basados en árboles produzcan resultados idénticos en ejecuciones sucesivas.

Las librerías empleadas, junto con su función dentro del estudio, se relacionan en la Tabla 8.

Tabla 8
Entorno de ejecución y librerías

Componente	Versión	Función en el estudio
Python	3.12.13	Lenguaje de desarrollo
NumPy	2.0.2	Operaciones numéricas y manejo de arreglos
pandas	2.2.2	Manipulación de datos y series temporales
statsmodels	0.14.6	Modelos SARIMA, SARIMAX y VAR; pruebas diagnósticas (Ljung-Box, Jarque-Bera, ARCH)
scikit-learn	1.6.1	Random Forest, búsqueda de hiperparámetros y métricas de evaluación
XGBoost	3.3.0	Modelo de potenciación de gradiente XGBoost

Componente	Versión	Función en el estudio
LightGBM	4.6.0	Modelo de potenciación de gradiente LightGBM
TensorFlow / Keras	2.20.0	Redes neuronales recurrentes LSTM y GRU
Prophet	1.3.0	Modelo aditivo estructural Prophet
SHAP	0.52.0	Interpretabilidad de los modelos

Nota. Elaboración propia.

7.7 Esquema de validación

La evaluación de los modelos se apoya en una partición temporal de los datos en tres conjuntos disjuntos y ordenados en el tiempo, respetando en todo momento la naturaleza secuencial de la serie. A diferencia de los esquemas de validación cruzada empleados en problemas sin estructura temporal, en los que las observaciones se reordenan de forma aleatoria, en este estudio la partición preserva el orden cronológico, de modo que ningún modelo se entrena con información posterior al periodo que debe pronosticar.

Los tres conjuntos se definen de la siguiente manera. El conjunto de entrenamiento comprende el periodo de enero de 2015 a diciembre de 2023 y se emplea para el ajuste de los modelos. El conjunto de validación abarca los años 2024 y 2025 y se utiliza para la selección de las variables exógenas y de los hiperparámetros: todas las configuraciones candidatas se comparan por su desempeño sobre este periodo, y se elige la de menor error. El conjunto de prueba corresponde a los meses de enero a mayo de 2026 y constituye una evaluación en condiciones de operación real; sus valores no intervienen en ninguna decisión de ajuste ni de selección, y se reservan exclusivamente para medir el desempeño final de los modelos ya definidos.

La separación mediante un único corte temporal, en lugar de una validación cruzada de tipo deslizante o expansivo, responde al tamaño reducido de la serie:

fragmentar 132 observaciones mensuales en múltiples particiones dejaría conjuntos de entrenamiento demasiado pequeños para estimar de forma estable modelos con la complejidad de los aquí considerados, en particular las redes neuronales. El corte único en 2024 garantiza un periodo de entrenamiento suficiente y un periodo de validación de dos años completos, que cubre dos ciclos estacionales.

Una vez seleccionada la configuración definitiva de cada modelo sobre el conjunto de validación, el modelo se reajusta empleando la totalidad de los datos disponibles hasta diciembre de 2025 (es decir, entrenamiento y validación combinados) antes de generar el pronóstico del conjunto de prueba. De este modo, el pronóstico final aprovecha toda la información histórica disponible, sin haber utilizado en ningún momento los datos de 2026 para tomar decisiones de modelamiento.

El control de fuga de información, planteado como parte del primer objetivo de la investigación, se materializa en dos mecanismos complementarios. El primero es la incorporación de las variables exógenas con un rezago de un mes, descrita en la Sección 7.4, que impide que un pronóstico utilice información del mismo periodo que pretende predecir. El segundo es la separación estricta del conjunto de prueba, que no participa en la selección de variables, la búsqueda de hiperparámetros ni el ajuste de los modelos. La combinación de ambos mecanismos asegura que el desempeño reportado sobre el periodo de prueba refleje la capacidad predictiva real de los modelos en un escenario de uso, y no un resultado inflado por el acceso indebido a información futura.

Conviene declarar una asimetría del diseño: los modelos estadísticos se ajustaron sobre la serie 2015-2025 y los modelos de aprendizaje automático sobre las series disponibles desde 2005 (Secciones 7.3 y 7.4). Ambas familias comparten el mismo corte de validación (enero de 2024), el mismo reajuste hasta diciembre de 2025 y el mismo conjunto de prueba, de modo que la evaluación final es comparable en cuanto a la

información posterior a 2015 y al periodo pronosticado; sin embargo, los modelos de aprendizaje automático dispusieron de diez años adicionales de historia para el ajuste. Esta diferencia debe tenerse presente al interpretar la ventaja de LightGBM sobre SARIMAX en la Sección 8.5, y su eliminación, reentrenando ambas familias sobre una ventana común, se plantea como parte del trabajo futuro.

7.8 Métricas de evaluación

El desempeño de todos los modelos se evalúa mediante seis métricas de error, aplicables por igual a las tres familias por tratarse de un problema de regresión sobre una variable continua. Su cálculo se realiza sobre los mismos conjuntos de validación y prueba para todos los modelos, lo que garantiza la comparabilidad directa de los resultados. Las métricas se definen en la Tabla 9.

Tabla 9
Métricas de evaluación

Métrica	Definición
MAE (Error Absoluto Medio)	Promedio del valor absoluto de los errores
RMSE (Raíz del Error Cuadrático Medio)	Raíz cuadrada del promedio de los errores al cuadrado
MAPE (Error Porcentual Absoluto Medio)	Promedio del error absoluto expresado como porcentaje del valor observado
ME (Error Medio)	Promedio de los errores con signo (Observado menos pronosticado; positivo es igual a subestimación promedio y negativo es igual a sobreestimación promedio)

Métrica	Definición
MASE (Error Absoluto Escalado Medio)	Cociente entre el MAE del modelo y el MAE de un pronóstico ingenuo estacional
R^2 (Coeficiente de determinación)	Proporción de la varianza de la variable observada explicada por el modelo

Nota. Elaboración propia.

Cada métrica aporta una perspectiva complementaria. El MAE y el RMSE cuantifican el error en la unidad original de la serie, con la diferencia de que el segundo penaliza más los errores de gran magnitud; su comparación permite detectar la presencia de errores atípicos. El MAPE expresa el error en términos relativos, lo que facilita su interpretación. El ME, al conservar el signo del error, revela si un modelo tiende sistemáticamente a quedarse por debajo o por encima de los valores reales, información que las métricas de error absoluto no capturan.

El MASE cumple un papel particular en la evaluación, pues sitúa el desempeño de cada modelo frente a un referente objetivo: el pronóstico ingenuo estacional, que consiste en tomar como predicción el valor observado doce meses atrás. Un MASE inferior a uno indica que el modelo aporta valor sobre esta regla trivial, mientras que un valor igual o superior a uno revela que el modelo no mejora una predicción que no requiere ajuste alguno. Esta métrica resulta especialmente informativa en una serie con estacionalidad marcada como la de homicidios.

Finalmente, el coeficiente de determinación (R^2) se reporta como medida de ajuste global, si bien su interpretación sobre el conjunto de prueba debe realizarse con atención:

al estar constituido este por solo cinco observaciones (enero a mayo de 2026), el R^2 depende de la varianza de esos cinco valores y resulta sensible a la exclusión de cualquiera de ellos, como se cuantifica en la Sección 8.5. Por ello se privilegia como criterio principal de comparación el conjunto de métricas de error, en especial el RMSE y el MASE, que no dependen de la varianza del periodo evaluado. Conviene precisar que el R^2 y el MASE se miden contra referentes distintos: el primero toma como referencia la media del periodo evaluado y el segundo el pronóstico ingenuo estacional, de modo que un modelo puede superar uno y no el otro. Por la misma razón, el tamaño del conjunto de prueba limita la aplicación de pruebas formales de igualdad de precisión predictiva entre modelos: la prueba de Diebold y Mariano (1995) es de naturaleza asintótica, y con cinco observaciones su potencia es mínima aun con la corrección de muestra pequeña de Harvey, Leybourne y Newbold (1997). En la Sección 8.5 se reporta esta prueba con carácter indicativo, y sus resultados se interpretan a la luz de esa limitación.

7.9 Interpretabilidad (SHAP)

Además de las predicciones, el estudio busca identificar qué variables influyen de manera determinante en el comportamiento del homicidio, en correspondencia con el tercer objetivo de la investigación. Dado que las familias de modelos difieren en su estructura interna, la interpretabilidad se aborda con la técnica apropiada a cada una, evitando forzar un método único sobre modelos de naturaleza distinta.

Para los modelos de aprendizaje automático se emplearon los valores SHAP (SHapley Additive exPlanations), un marco basado en la teoría de juegos que cuantifica la contribución de cada variable a cada predicción individual, permitiendo estimar tanto la importancia global de los predictores como la dirección de su efecto. La implementación se ajustó al tipo de modelo: para los modelos basados en árboles (XGBoost, LightGBM y Random Forest) se utilizó el explicador específico para estructuras de árbol, que calcula

las contribuciones de forma exacta y eficiente; para las redes neuronales recurrentes (LSTM y GRU) se empleó un explicador agnóstico al modelo, aplicado sobre una muestra representativa de datos de referencia, apropiado para arquitecturas en las que no existe un explicador especializado.

Para el modelo Prophet, la interpretación se realiza a través de los coeficientes estimados de los regresores externos, que expresan de forma directa el aporte de cada variable exógena a la predicción, coherente con la naturaleza aditiva del modelo. De manera análoga, en los modelos estadísticos (SARIMA, SARIMAX y VAR) la contribución de cada componente se interpreta a partir de sus coeficientes estimados y de la significancia estadística de estos, evaluada mediante sus intervalos de confianza.

La combinación de estas técnicas permite contrastar, entre familias de modelos con lógicas de funcionamiento diferentes, si las variables identificadas como más influyentes son consistentes, lo que otorga mayor robustez a las conclusiones sobre los determinantes del fenómeno.

8. Modelamiento y Evaluación de Resultados

Este capítulo presenta el desarrollo metodológico y los resultados obtenidos para el pronóstico mensual del homicidio intencional en Colombia, con el objetivo central de comparar el desempeño de modelos estadísticos clásicos y de aprendizaje automático bajo un mismo esquema riguroso de evaluación. El capítulo se organiza en seis secciones. La Sección 8.1 describe el procesamiento de los datos y la construcción de la serie mensual utilizada como variable objetivo. La Sección 8.2 presenta el análisis exploratorio del fenómeno, su comportamiento temporal, estacionalidad y distribución por características de las víctimas, como insumo para interpretar los resultados posteriores. Las secciones 8.3 y 8.4 desarrollan, respectivamente, tres modelos estadísticos clásicos (SARIMA, SARIMAX y VAR) y seis de aprendizaje automático (XGBoost, LightGBM, Random Forest, LSTM, GRU y Prophet), todos seleccionados bajo el mismo esquema (entrenamiento 2015-2023 y validación 2024-2025), reajustados después sobre la serie completa hasta diciembre de 2025 y evaluados finalmente contra el homicidio real y externo observado entre enero y mayo de 2026, un conjunto de prueba que ningún modelo utilizó durante su ajuste o selección de hiperparámetros. La Sección 8.5 consolida los resultados de los nueve modelos, contrasta los hallazgos frente a la hipótesis de trabajo planteada, y discute la idoneidad de cada modelo más allá de sus métricas de error, con miras a su eventual aplicación práctica en el seguimiento y anticipación del fenómeno.

8.1 Procesamiento de los datos

La fuente principal de información fue la base consolidada db_delitos_norm, proveniente del Sistema de Información Estadístico, Delincuencial, Contravencional y Operativo de la Policía Nacional (SIEDCO), integrada por 3.785.395 registros de distintas

temáticas delictivas y 18 variables (Policía Nacional de Colombia, s.f.). Para construir la serie nacional de homicidio intencional del periodo 2015–2025, se seleccionaron 110.381 registros de la temática HOMICIDIOS, correspondientes a 141.774 víctimas. La agregación mensual se realizó mediante la suma del campo Cantidad, de modo que la variable objetivo representa el número de víctimas por mes y no el número de registros.

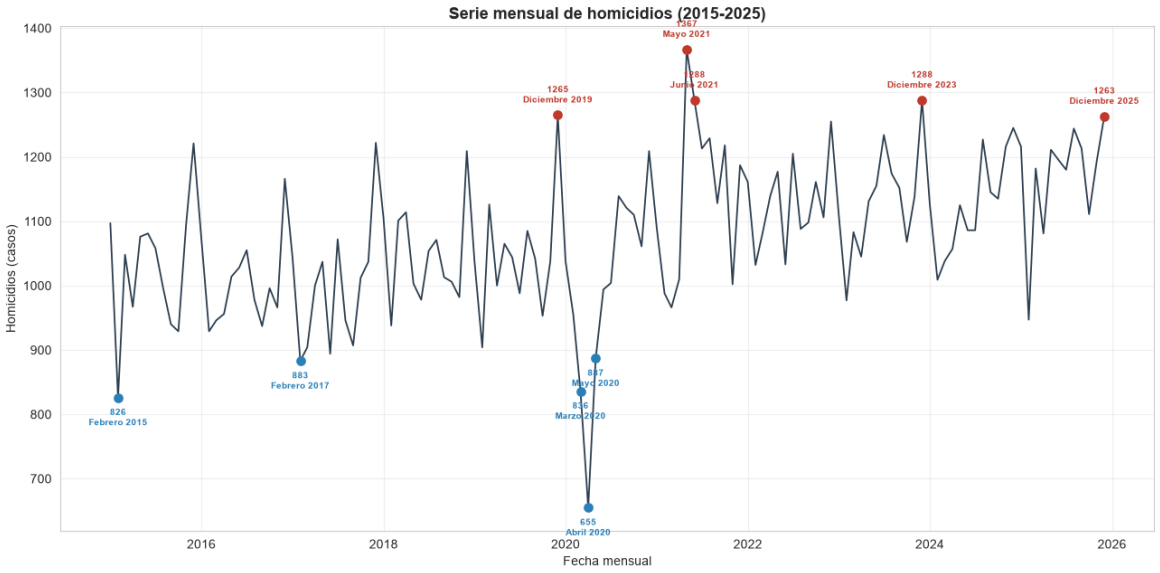
Para el ejercicio de modelamiento comparativo desarrollado en las secciones 8.3 y 8.4, la serie se acotó al periodo enero de 2015 a diciembre de 2025 (132 observaciones mensuales), reservando enero-mayo de 2026 como conjunto de prueba real y externo a todo el proceso de ajuste y selección de modelos. Adicionalmente, la serie de homicidios se enriqueció con variables mensuales potencialmente asociadas al fenómeno (hurto a personas, lesiones personales, violencia intrafamiliar, incautación de estupefacientes (kg), el Índice de Seguimiento a la Economía (ISE) y la tasa de desempleo), que se emplearon como variables exógenas candidatas en los modelos multivariados de las secciones siguientes.

8.2 Análisis exploratorio

Con el fin de caracterizar el comportamiento del homicidio intencional en Colombia antes de ajustar cualquier modelo, se realizó un análisis exploratorio de los homicidios registrados en el periodo 2015-2025. En este periodo se registraron 141.774 víctimas de homicidio en el país, cuyo conteo mensual dio lugar a la serie analizada en este capítulo. Véase Figura 5.

Figura 5

Serie mensual de homicidios en Colombia



Nota. Serie mensual de homicidios en Colombia, enero de 2015 a diciembre de 2025, con los cinco meses de mayor y los cinco de menor incidencia señalados directamente sobre la curva. Elaboración propia.

La serie mensual evidencia fluctuaciones importantes alrededor de una banda relativamente estable, se puede observar una tendencia de crecimiento, aunque no de forma sostenida a lo largo del periodo analizado. Los picos más altos se concentran en mayo y junio de 2021 (1367 y 1265 homicidios respectivamente), en diciembre de 2023 (1288 homicidios), diciembre de 2019 (1265 homicidios) y diciembre de 2025 (1263 homicidios); mientras que los valles más pronunciados aparecen en 2020 y comienzos de 2021, coincidiendo con las restricciones de movilidad asociadas a la pandemia por COVID-19. La Figura 6 permite observar en detalle la descomposición de la serie en cada uno de sus componentes: tendencia, estacionalidad y residuales.

Figura 6

Descomposición aditiva de la serie de homicidios (2015-2025)



Nota. Elaboración propia.

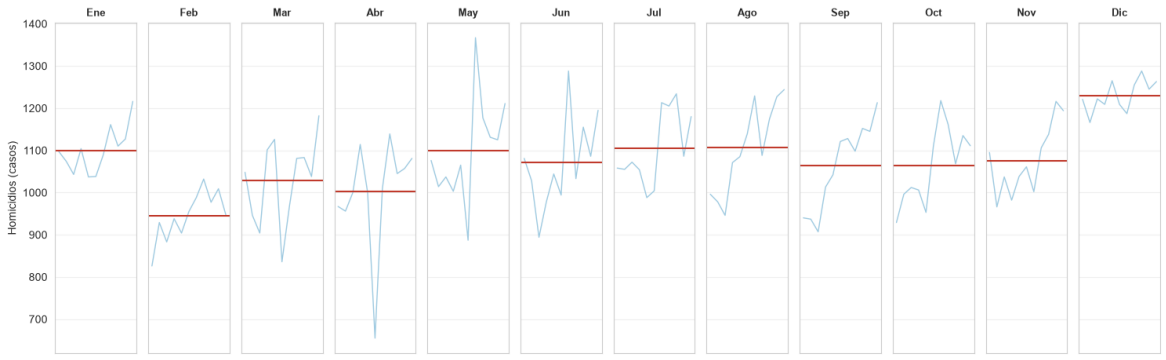
El primer componente permite observar la tendencia de crecimiento del fenómeno, con un descenso moderado entre 2015 y 2020, un mínimo pronunciado por el efecto del COVID-19, luego evidencia un repunte con el mayor pico en 2021 y luego una tendencia sostenida hasta diciembre de 2025. Por su parte, el componente estacional muestra un patrón que se repite año tras año, con un pico marcado de aproximadamente 150 casos sobre la tendencia y un cercano a los 100 casos, esto confirma la presencia de una estacionalidad anual y estable en el fenómeno. Finalmente, el componente residual se

mantiene acotado en alrededor de 100 casos, salvo por el atípico de 2020 (efecto pandemia), lo que indica que la serie es estable alrededor de un promedio.

Este comportamiento sugiere que, más que una tendencia estructural de largo plazo, el homicidio en Colombia exhibe una dinámica de oscilación en torno a un nivel promedio, interrumpida por choques externos puntuales, un patrón que, como se discute en la Sección 8.3, puede resultar difícil de capturar con modelos que asumen estructuras de tendencia y estacionalidad estables.

Al observar en mayor detalle la estacionalidad en la Figura 7, se observa el comportamiento de los homicidios mes a mes durante el periodo 2015-2025 comparado con el promedio mensual (línea roja). El mayor promedio histórico lo tiene el mes de diciembre seguido de enero, mayo, julio y agosto, esto confirma la presencia de una estacionalidad anual definida y consistente con lo observado en la descomposición.

Figura 7
Estacionalidad de la serie de tiempo

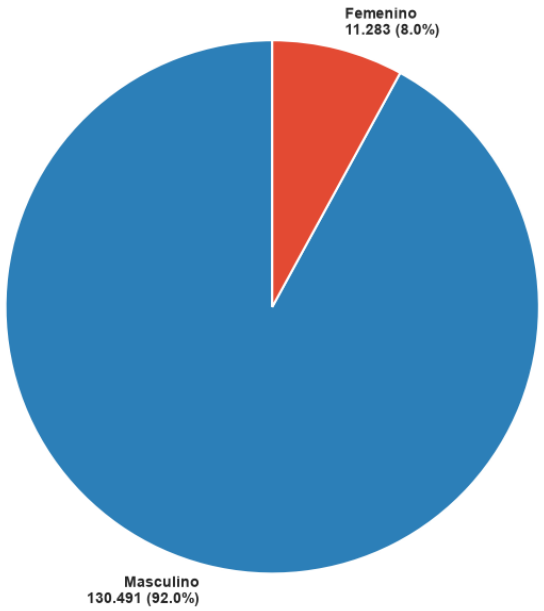


Nota. Elaboración propia.

Dentro del análisis exploratorio también se tuvieron en cuenta las distribuciones por características de las víctimas de homicidio, tales como género, zona y armas empleadas. En este sentido la Figura 8 muestra la distribución de los homicidios por género.

Figura 8

Distribución de los homicidios por género de la víctima (2015-2025)

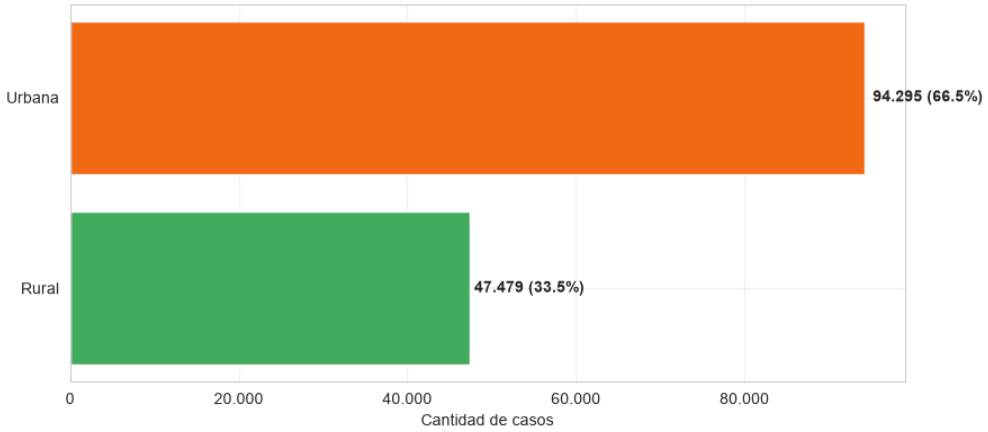


Nota. Elaboración propia.

La distribución por género confirma un patrón ampliamente documentado en la criminología latinoamericana y es que la inmensa mayoría de las víctimas de homicidio son hombres. Esta marcada disparidad sugiere que los factores de riesgo asociados al homicidio en Colombia, vínculos con economías ilegales, conflictos interpersonales armados, dinámicas de crimen organizado, afectan de forma muy distinta a hombres y mujeres, lo que resulta relevante al momento de diseñar estrategias de prevención diferenciadas por género.

La Figura 9 muestra la comparación entre zona urbana y rural, evidenciando que la mayoría de los homicidios del periodo se concentra en cabeceras urbanas, aunque la zona rural conserva una participación considerable.

Figura 9
Distribución de los homicidios por zona (2015-2025)



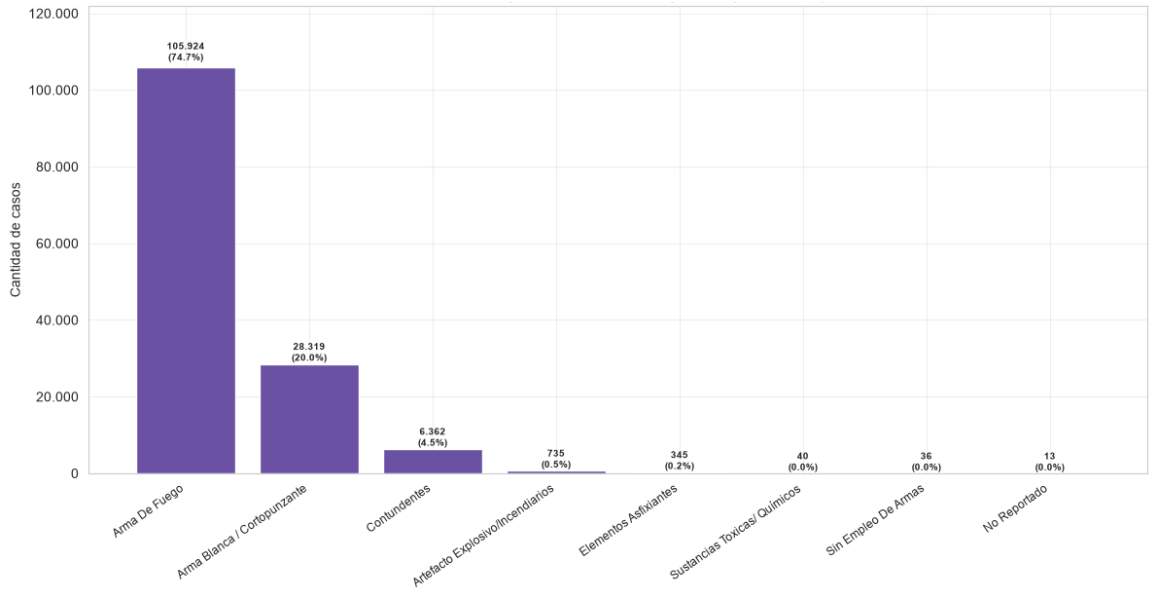
Nota. Elaboración propia.

Esta coexistencia de dos dinámicas territoriales distintas es un antecedente importante para entender el homicidio en Colombia y su distribución espacial; sin embargo, al modelarse la serie agregada a nivel nacional, dichos modelos no incorporan factores espaciales, dado que se centran en identificar determinantes explicativos del fenómeno.

La Figura 10 muestra que el arma de fuego es el medio más utilizado para cometer homicidio en Colombia durante la última década, muy por encima del arma blanca y de las demás categorías.

Figura 10

Distribución de los homicidios por arma o medio empleado (2015-2025)



Nota. Elaboración propia.

Este hallazgo es consistente con la hipótesis de que buena parte del homicidio en el país está asociado a dinámicas de criminalidad organizada y a la disponibilidad de armas de fuego, más que a la violencia interpersonal ocasional, y es un elemento de contexto relevante para interpretar por qué variables macroeconómicas y sociales, más que puramente temporales, pueden aportar poder explicativo adicional a los modelos multivariados presentados en las secciones posteriores.

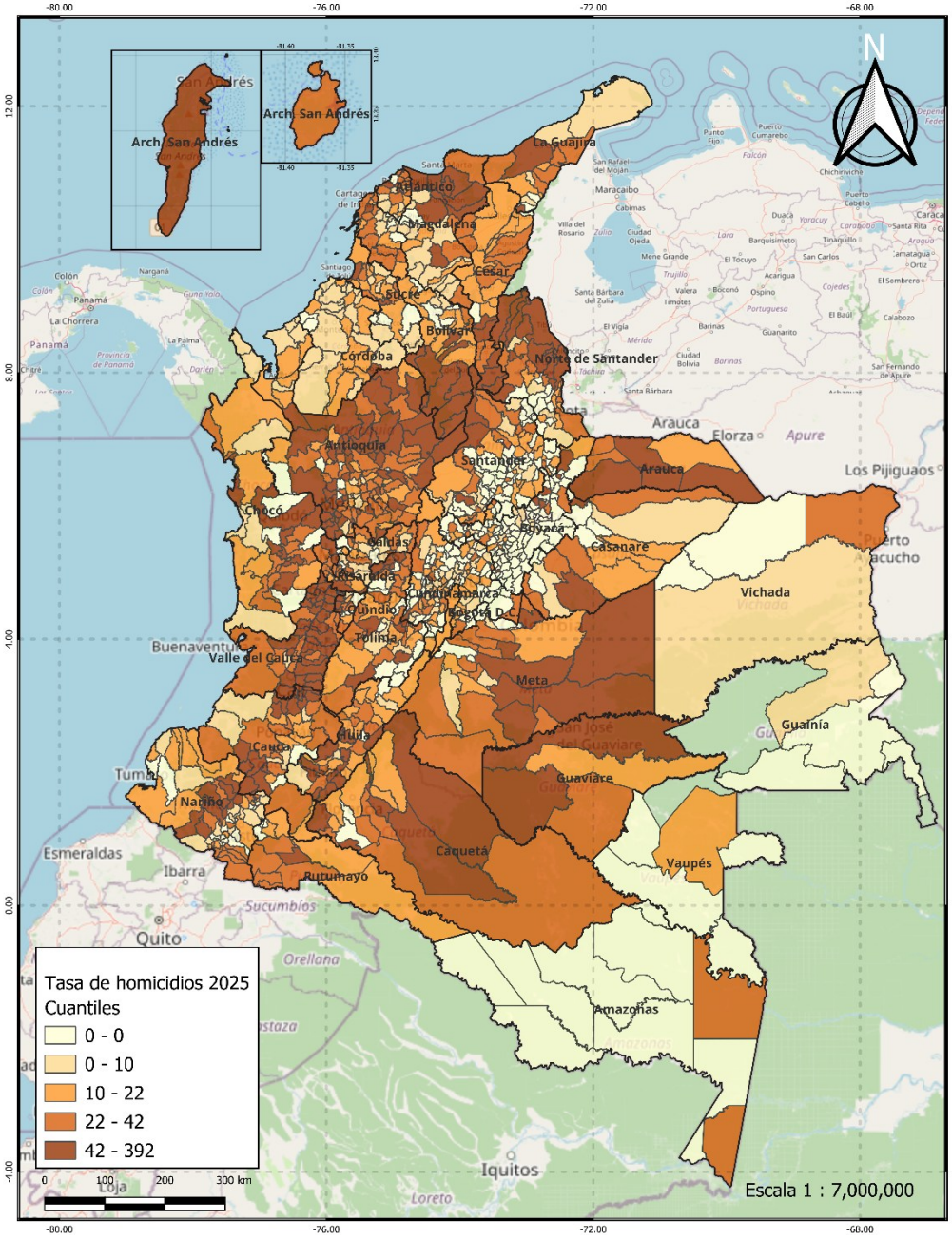
Por otra parte, al revisar la distribución espacial del homicidio, tomando como base la tasa de homicidios por municipio para 2025, ver la Figura 11, se evidencia una heterogeneidad territorial que la serie mensual agregada a nivel nacional, presentada en las figuras anteriores de esta sección, no permite apreciar. Los municipios con las tasas más altas (categoría 42-392 por cada 100.000 habitantes, en tono café oscuro) se concentran en tres corredores geográficos: el Pacífico y Andén Pacífico (Chocó, Cauca, Nariño), la región del Catatumbo y zonas fronterizas de Norte de Santander y Arauca, y

buena parte de la Orinoquia y la Amazonia noroccidental (Meta, Guaviare, Caquetá, Putumayo). Esta distribución espacial coincide, en términos generales, con zonas de presencia histórica de economías ilegales y de disputa territorial entre grupos armados, lo que resulta coherente con el hallazgo ya documentado en esta misma sección de que el arma de fuego es el medio ampliamente predominante en el homicidio nacional: un indicio adicional de que buena parte del fenómeno está asociado a dinámicas de criminalidad organizada más que a la violencia interpersonal ocasional.

En contraste, un número considerable de municipios, particularmente en el altiplano cundiboyacense, el eje cafetero y buena parte de la región Caribe interior, presenta tasas nulas o muy bajas (categorías 0-0 y 0-10, en tonos claros), lo que refleja tanto una menor incidencia real del fenómeno como la posibilidad de subregistro en municipios pequeños o de difícil acceso estadístico. Esta disparidad territorial tan marcada, con municipios que superan las 42 víctimas por 100.000 habitantes conviviendo con otros de tasa cero dentro del mismo departamento, es consistente con la coexistencia de dinámicas urbana y rural ya señalada al analizar la distribución por zona, y constituye una limitación relevante a tener presente en los capítulos siguientes: los modelos de pronóstico desarrollados en este trabajo se ajustan sobre la serie agregada a nivel nacional, por lo que sus resultados deben interpretarse como una aproximación al comportamiento promedio del país, sin capturar la variabilidad espacial que este mapa hace evidente.

Figura 11

Distribución espacial de la tasa de homicidios (2025)



Nota. Elaboración propia.

8.3 Modelos estadísticos

Para el pronóstico mensual del homicidio se evaluaron tres familias de modelos estadísticos clásicos: SARIMA (univariado), SARIMAX (con variables exógenas) y VAR (vectores autorregresivos, con una variable acompañante). En los tres casos, en lugar de fijar de antemano una única especificación, se implementó una búsqueda automática entre varias combinaciones candidatas de orden y de variables exógenas, evaluando y priorizando, en ese orden, (i) el número de supuestos del modelo que se cumplen (independencia, normalidad y homocedasticidad de los residuales) y (ii) los criterios de información de Akaike y Bayesiano.

8.3.1 Modelo SARIMA

La especificación de los modelos SARIMA y SARIMAX de este trabajo se enmarca en la metodología Box-Jenkins, propuesta originalmente por Box y Jenkins (1970), que constituye el enfoque clásico para el modelamiento de series de tiempo univariadas y multivariadas mediante procesos autorregresivos integrados de medias móviles (ARIMA/SARIMA). En este sentido, se debe determinar si la serie es estacionaria, mediante pruebas formales como la Dickey-Fuller Aumentada (ADF) o KPSS y, en caso de no serlo, establecer el orden de diferenciación regular (d) y estacional (D) para estabilizarla, apoyándose además en el análisis de las funciones de autocorrelación (ACF) y autocorrelación parcial (PACF) para sugerir los órdenes autorregresivos (p, P) y de medias móviles (q, Q).

La prueba de Dickey-Fuller Aumentada o ADF (Dickey y Fuller, 1979; Said y Dickey, 1984) contrasta la presencia de una raíz unitaria en una serie de tiempo, es decir, evalúa si la serie sigue un proceso no estacionario cuyos choques tienen efectos permanentes. Las hipótesis son:

H_0: γ = 0, la serie posee raíz unitaria (es no estacionaria).

H_1: γ < 0, la serie no posee raíz unitaria (es estacionaria).

La regla de decisión es la siguiente: si el p-valor es inferior al nivel de significancia establecido (α = 0,05), se rechaza H_0 y se concluye que la serie es estacionaria; en caso contrario, no se rechaza H_0 y no se cuenta con evidencia suficiente para descartar la presencia de raíz unitaria.

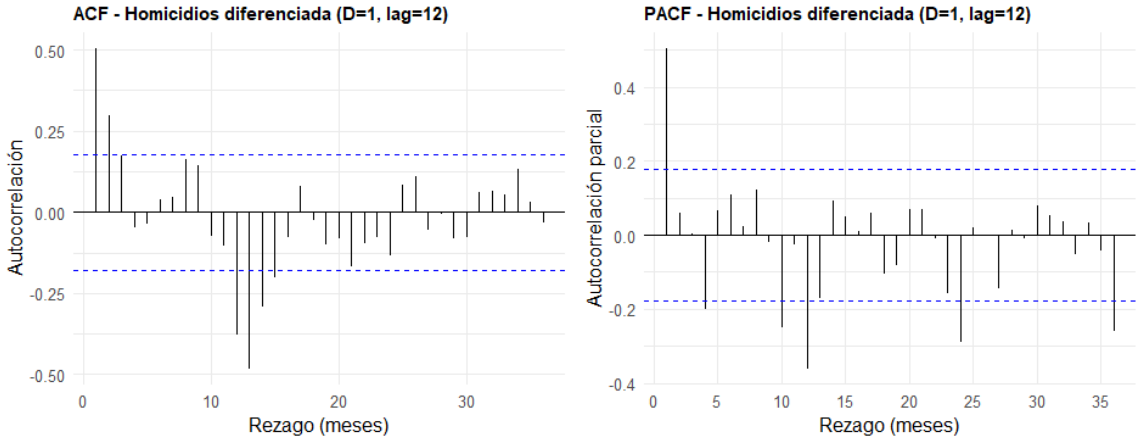
La prueba ADF aplicada a la serie mensual de homicidios en niveles (132 observaciones, enero de 2015 a diciembre de 2025), incorporando 12 rezagos para controlar la estacionalidad anual identificada en la descomposición aditiva, arrojó un estadístico Dickey-Fuller de -3,2 con un p-valor de 0,091. Dado que este valor supera el umbral de significancia convencional (α = 0,05), no se rechaza la hipótesis nula de raíz unitaria, por lo que no existe evidencia estadística suficiente para afirmar que la serie de homicidios en niveles es estacionaria. Este resultado es consistente con el comportamiento observado previamente en la descomposición de la serie (Sección 8.2.1) y en el gráfico de subseries (Sección 8.2.2), donde se identificó una estacionalidad anual marcada y estable, factor que plausiblemente explica la no estacionariedad detectada en la serie original y que justifica metodológicamente la inclusión de una diferencia estacional (D = 1, periodo 12) en la especificación del modelo SARIMA seleccionado en la Sección 8.3.1.

Considerando que fue necesario aplicar una diferencia estacional, de orden 12 a la serie de homicidios, la prueba ADF arrojó un estadístico de -5,0 con un p-valor de 0,01, lo que permite rechazar la hipótesis nula de raíz unitaria y concluir que la serie es estacionaria. Este resultado se corroboró mediante la prueba KPSS (Kwiatkowski et al., 1992), cuya hipótesis nula es la inversa (estacionariedad): con un estadístico de 0,09 y un p-valor de 0,1, tampoco se rechaza dicha hipótesis. La coincidencia de ambas pruebas

ADF rechazando no estacionariedad y KPSS sin rechazar estacionariedad, constituye evidencia robusta y cruzada de que la diferenciación estacional de orden 12 es suficiente para estabilizar la serie de homicidios.

El análisis de las funciones de autocorrelación (ACF) y autocorrelación parcial (PACF) Figura 12 de la serie de homicidios diferenciada estacionalmente (D=1, periodo 12) permite orientar la identificación de los órdenes del modelo.

Figura 12
ACF y PACF



Nota. Elaboración propia.

En la parte regular, el corte abrupto de la PACF después del primer rezago, junto con el decaimiento gradual de la ACF, es consistente con un componente autorregresivo de orden 1, AR(1). En la parte estacional, el pico negativo pronunciado de la ACF en el rezago 12, que no se repite con la misma magnitud en los rezagos 24 y 36, junto con una PACF que decae de forma más persistente a través de dichos múltiplos, sugiere un componente de media móvil estacional de orden 1, SMA(1). En este sentido se identificaron 54 combinaciones de modelos diferentes SARIMA, la Tabla 10 resume los

modelos con mejores criterios de información de Akaike y Bayesiano, así como la evaluación de los supuestos de residuales.

Tabla 10
Mejores modelos SARIMA considerados

Especificación	AIC	BIC	p1	Indep.	p2	Norm.	p3	Homocedast.
SARIMA(2,0,1)(0,1,2)[12]	1401	1418	0,0395	NO	6.94e-12	NO	0,267	SÍ
SARIMA(2,0,1)(1,1,1)[12]	1403	1419	0,0253	NO	6.99e-12	NO	0,307	SÍ
SARIMA(1,0,1)(0,1,2)[12]	1403	1417	0,0193	NO	1.10e-10	NO	0,538	SÍ
SARIMA(2,0,0)(0,1,2)[12]	1403	1417	0,0220	NO	1.97e-10	NO	0,569	SÍ
SARIMA(1,0,1)(2,1,1)[12]	1403	1420	0,0186	NO	1.37e-10	NO	0,679	SÍ
SARIMA(2,0,1)(1,1,2)[12]	1403	1423	0,0238	NO	4.49e-12	NO	0,320	SÍ
SARIMA(2,0,0)(2,1,1)[12]	1404	1420	0,0234	NO	1.66e-10	NO	0,719	SÍ
SARIMA(1,0,0)(0,1,2)[12]	1404	1415	0,0257	NO	1.86e-07	NO	0,544	SÍ
SARIMA(1,0,0)(2,1,1)[12]	1404	1418	0,0384	NO	4.49e-08	NO	0,688	SÍ
SARIMA(2,0,1)(0,1,1)[12]	1404	1418	0,0209	NO	1.26e-13	NO	0,503	SÍ

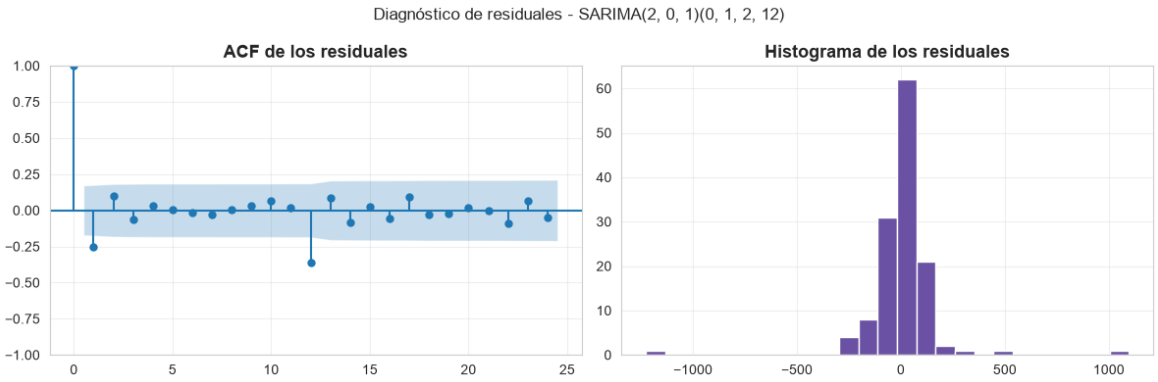
Nota. p1 representa Ljung-Box p2 Jarque-Bera y p3 es p (ARCH). Elaboración propia.

La evaluación de los residuales de cada especificación candidata se realizó mediante tres pruebas de diagnóstico complementarias. La prueba de Ljung-Box contrasta la hipótesis nula de que los residuales no presentan autocorrelación serial, es decir, que se comportan como ruido blanco; un p-valor inferior a 0,05 lleva a rechazar esta hipótesis y concluir que persiste dependencia temporal sin capturar por el modelo, lo que compromete su capacidad de generalización fuera de muestra. La prueba de Jarque-Bera contrasta la hipótesis nula de que los residuales siguen una distribución normal, evaluando

conjuntamente su asimetría y curtosis; su cumplimiento es relevante porque la validez de los intervalos de confianza del pronóstico depende, en buena medida, de este supuesto. La prueba ARCH (implementada como un Ljung-Box sobre los residuales al cuadrado) contrasta la hipótesis nula de homocedasticidad, es decir, de que la varianza de los residuales se mantiene constante en el tiempo; su rechazo indicaría heterocedasticidad condicional, para la cual sería necesario complementar el modelo con una estructura de varianza tipo GARCH. En los tres casos, un p-valor mayor o igual a 0,05 se interpreta como que no hay evidencia suficiente para rechazar el supuesto correspondiente, por lo que se considera "cumplido" para efectos de la comparación de especificaciones.

Considerando que los residuales del modelo SARIMA no superan las pruebas de independencia ni de normalidad, el modelo deja sin explicar parte de la estructura de dependencia temporal de la serie. La Figura 13 muestra en el gráfico de ACF, que la gran mayoría de los rezagos se mantiene dentro de la banda de confianza, con la excepción de un rezago negativo puntual cercano al mes 12 (≈ -0,35), que sugiere un remanente menor de estructura estacional no completamente capturado por el modelo. El histograma evidencia que los residuales se concentran fuertemente alrededor de cero, pero con una cola marcadamente pesada hacia ambos extremos, con valores atípicos

Figura 13
Diagnóstico de los residuales del modelo SARIMA: función de autocorrelación (ACF) e histograma.



Nota. Elaboración propia.

Esta limitación de los supuestos es una primera señal de alerta sobre la capacidad de este modelo para generalizar su pronóstico fuera de muestra, sin embargo, se procedió a calcular el pronóstico para los meses de enero a mayo de 2026 (Tabla 11 y Figura 14), así como el respectivo cálculo de métricas de desempeño del modelo (Tabla 12).

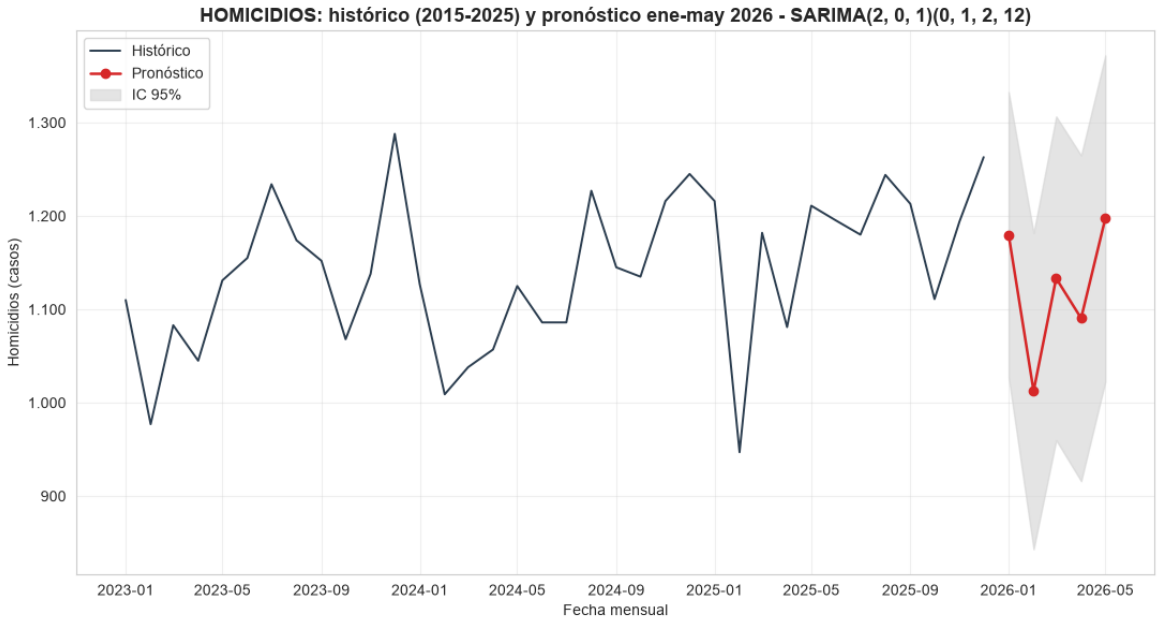
Tabla 11
Pronóstico puntual e intervalo de confianza del 95% del modelo SARIMA, enero-mayo de 2026.

Modelo	Fecha mensual	Pronóstico	Límite inf. (95%)	Límite sup. (95%)
SARIMA(2, 0, 1)(0, 1, 2, 12)	2026-01-01	1.179,9	1.027,2	1.332,6
SARIMA(2, 0, 1)(0, 1, 2, 12)	2026-02-01	1.012,5	843,1	1.181,9
SARIMA(2, 0, 1)(0, 1, 2, 12)	2026-03-01	1.133,2	959,8	1.306,6
SARIMA(2, 0, 1)(0, 1, 2, 12)	2026-04-01	1.090,5	916,0	1.264,9
SARIMA(2, 0, 1)(0, 1, 2, 12)	2026-05-01	1.197,3	1.022,5	1.372,0

Nota. Elaboración propia.

Figura 14

Serie histórica de homicidios (2015-2025) y pronóstico del modelo SARIMA para enero-mayo de 2026, con intervalo de confianza del 95%.



Nota. Elaboración propia.

Tabla 12

Métricas de desempeño del modelo SARIMA frente al valor observado real de enero-mayo de 2026.

Modelo	MAE	RMSE	MASE	ME	MAPE	R2
SARIMA(2, 0, 1)(0, 1, 2, 12)	41,038	47,415	0,543	37,331	3,605	-0,295

Nota. Elaboración propia.

Al contrastarse contra el homicidio efectivamente observado entre enero y mayo de 2026, el modelo SARIMA(2,0,1)(0,1,2)[12] presenta un error absoluto medio (MAE) de 41,0 homicidios por mes, equivalente a un error porcentual medio (MAPE) de 3,6% sobre un nivel mensual que ronda los 1.100-1.200 casos. El error medio (ME = 37,3), muy cercano al MAE, indica que el modelo subestimó de forma consistente el número de homicidios en

cuatro de los cinco meses del horizonte de prueba, más que compensar meses de subestimación con meses de sobreestimación. El R^2 de -0,295 indica que el error cuadrático del modelo supera al obtenido al utilizar como referencia la media de los valores observados en el periodo de prueba: dado que esta métrica compara el error del modelo contra el error que se obtendría al pronosticar simplemente el promedio de los cinco valores observados en 2026 (un promedio cercano a 1.160 casos, con muy poca variabilidad entre esos cinco meses), basta con que el modelo se desvíe de forma constante hacia un mismo lado (en este caso, subestimando) para que su error cuadrático supere al de esa media de referencia y el R^2 caiga en terreno negativo. En otras palabras, el problema no es que el modelo carezca de precisión en términos absolutos (un error de 41 casos sobre ~1.150 es razonable), sino que su patrón de error es más sesgado que disperso, un rasgo que penaliza fuertemente al R^2 cuando el conjunto de prueba es corto (cinco meses) y poco variable.

8.3.2 Modelo SARIMAX

Con el fin de identificar la mejor especificación SARIMAX, se evaluaron todas las combinaciones posibles entre diez órdenes candidatos de la parte SARIMA, seleccionados por ser los de menor AIC en el análisis de identificación Box-Jenkins presentado en la sección anterior, y las 64 combinaciones posibles de las seis variables exógenas candidatas (incautación de estupefacientes, ISE, violencia intrafamiliar, lesiones personales, tasa de desempleo y hurto a personas). En total se ajustaron 640 modelos SARIMAX por máxima verosimilitud, cada uno evaluado mediante los criterios de información AIC y BIC, y sometido a las pruebas de independencia (Ljung-Box), normalidad (Jarque-Bera) y homocedasticidad (ARCH) sobre sus residuales. Para la implementación

de los modelos, las variables de conteo con distribución fuertemente asimétrica, incautación de estupefacientes y lesiones personales, se transformaron mediante logaritmo natural, y las restantes variables exógenas se estandarizaron (media cero, desviación estándar uno) con los parámetros del histórico 2015-2025. La Tabla 13 presenta las diez especificaciones con mejor desempeño, ordenadas primero por el número de supuestos de diagnóstico que satisfacen y, en caso de empate, por el menor AIC.

Tabla 13
Mejores modelos SARIMAX considerados.

Modelo	Variables	AIC	BIC	p (LB)	Indep.	p (JB)	Norm.	p (ARCH)	Homoc.
SARIMA(2,0,1)(0,1,2)[12]	LP + TD	1381	1403	9.43E−75	No	1.03E−71	No	0,0794	Sí
SARIMA(2,0,1)(0,1,2)[12]	LP + TD + HP	1381	1406	9.41E−76	No	1.29E−72	No	0,0693	Sí
SARIMA(2,0,1)(0,1,2)[12]	VI + LP + TD	1383	1408	1.31E−74	No	1.22E−71	No	0,0666	Sí
SARIMA(2,0,1)(0,1,2)[12]	VI + LP + TD + HP	1383	1411	1.10E−75	No	1.49E−72	No	0,0644	Sí
SARIMA(2,0,1)(0,1,2)[12]	VI + TD + HP	1387	1412	7.91E−79	No	1.28E−72	No	0,7271	Sí
SARIMA(2,0,1)(0,1,2)[12]	IK + VI + TD	1387	1415	2.53E−79	No	3.08E−73	No	0,7498	Sí
SARIMA(2,0,1)(0,1,2)[12]	IK + TD + HP	1389	1414	1.29E−79	No	4.69E−72	No	0,8073	Sí
SARIMA(2,0,1)(0,1,2)[12]	ISE + VI + TD	1389	1417	1.69E−78	No	9.75E−73	No	0,7035	Sí
SARIMA(2,0,1)(0,1,2)[12]	TD + HP	1389	1412	5.48E−79	No	3.29E−71	No	0,8157	Sí
SARIMA(2,0,1)(0,1,2)[12]	VI + TD	1391	1413	5.87E−77	No	1.59E−70	No	0,8115	Sí

Nota. LP: Lesiones personales; TD: Tasa de desempleo; HP: Hurto a personas; VI: Violencia intrafamiliar; IK: Incautación de drogas (kg); ISE: Índice de Seguimiento a la Economía; LB: prueba de Ljung-Box; JB: prueba de Jarque-Bera. La notación SARIMA(p,d,q)(P,D,Q)[12] identifica el orden de la parte SARIMA de cada modelo SARIMAX; las variables exógenas incorporadas se indican en la segunda columna. Elaboración propia.

El modelo seleccionado corresponde al primer lugar de esta tabla: SARIMA(2,0,1)(0,1,2)[12] con las variables exógenas, lesiones personales y tasa de desempleo, que alcanza el menor AIC (1380.9) y BIC (1403.2) de todos los candidatos evaluados, ningún modelo pasa pruebas de independencia y normalidad. Es decir, al

revisar los supuestos sobre la distribución de residuales y comparar con el modelo SARIMA univariado, no existen evidencia estadística que permita concluir que la incorporación conjunta de variables exógenas relacionadas con la actividad delictiva conexa, el entorno de mercados ilegales o las condiciones económicas mejoren el ajuste del modelo frente a la especificación puramente autorregresiva.

A partir de este modelo se realizó el pronóstico de enero a mayo de 2026 (Tabla 14 y Figura 15), así como el respectivo cálculo de métricas de desempeño del modelo (Tabla 15).

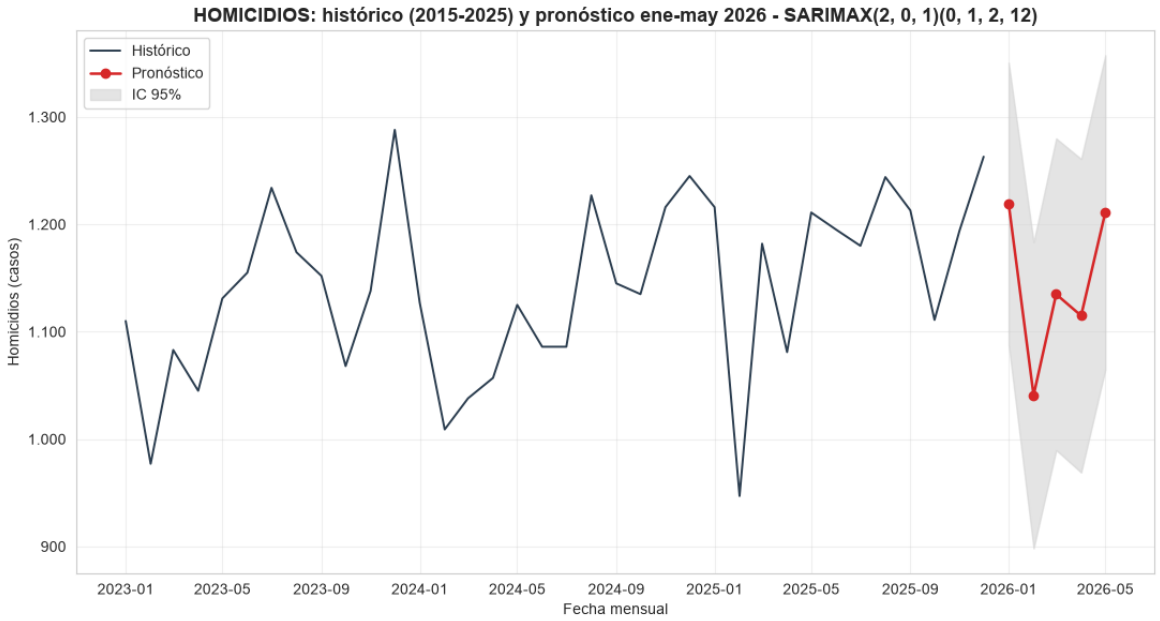
Tabla 14
Pronóstico puntual e intervalo de confianza del 95% del modelo SARIMAX, enero-mayo de 2026.

Fecha	Pronóstico	LI (95%)	LS (95%)
Ene. 2026	1.219,3	1.088,2	1.350,5
Feb. 2026	1.040,8	898,3	1.183,4
Mar. 2026	1.134,9	989,8	1.280,1
Abr. 2026	1.114,9	968,9	1.260,9
May. 2026	1.211,1	1.064,7	1.357,4

Nota. Elaboración propia.

Figura 15

Serie histórica de homicidios (2015-2025) y pronóstico del modelo SARIMAX para enero-mayo de 2026, con intervalo de confianza del 95%.



Nota. Elaboración propia.

Tabla 15

Métricas de desempeño del modelo SARIMAX frente al valor observado real de enero-mayo de 2026.

MAE	RMSE	MASE	ME	MAPE (%)	R^2
31,551	33,158	0,418	15,789	2,749	0,367

Nota. Elaboración propia.

De acuerdo con las métricas de desempeño, el modelo SARIMAX(2,0,1)(0,1,2)[12] resulta más eficiente que el SARIMA univariado en las seis métricas: un MAE de 31,6 frente a 41,0, un RMSE de 33,2 frente a 47,4, un MASE de 0,418 frente a 0,543, un MAPE de 2,7% frente a 3,6%, y un R^2 positivo (0,367) frente al -0,295 del SARIMA. El error medio (ME = 15,8), si bien continúa siendo positivo, lo que indica que el modelo aún subestima

en promedio el nivel de homicidios en el horizonte de prueba, aunque alternando el signo del error entre meses, es menos de la mitad del sesgo observado en SARIMA (37,3), lo que muestra que incorporar las variables exógenas no solo redujo el error absoluto, sino también el sesgo sistemático del pronóstico. Este resultado confirma la expectativa metodológica inicial: enriquecer el modelo autorregresivo con información estructural adicional efectivamente mejora su capacidad predictiva fuera de muestra, incluso considerando que el valor futuro de esas variables exógenas (al no contar con proyecciones reales para 2026) se aproximó mediante persistencia del último valor observado (diciembre de 2025), un supuesto que introduce una fuente de incertidumbre adicional pero que, en este caso, no impidió que el modelo superara a su contraparte univariada.

8.3.3 Modelo VAR

Previo a la selección del modelo VAR fue necesario abordar el número de rezagos o meses hacia atrás que debe tener en cuenta el modelo para explicar el fenómeno en el presente, para esto se probaron entre 1 y 12 rezagos (meses hacia atrás), y se compararon cuatro criterios estadísticos distintos: AIC y FPE señalaron que la mejor opción era usar los 12 meses completos, mientras que los otros dos BIC y HQIC, que penalizan con más fuerza a los modelos con muchos parámetros, sugirieron usar solo 1 o 2 meses. Se optó por adoptar 12 rezagos, siguiendo el criterio AIC/FPE, esto considerando la estacionalidad de la serie.

El modelo VAR, al ser multivariado, en lugar de tratar al homicidio como una serie aislada que solo depende de su propio pasado, modela al mismo tiempo los homicidios y otras variables que puedan explicar o aportar información útil para anticipar el homicidio. En este sentido, se aplicó la prueba de causalidad de Granger, la cual busca comprobar

(desde una perspectiva estadística) si una variable exógena observada mejora la capacidad de predecir los homicidios. Los resultados (Tabla 16) muestran que tanto el Índice de Seguimiento a la Economía, ISE, (p = 0,0009) como la incautación de estupefacientes (p = 0,0115) sí aportan esa capacidad predictiva adicional, mientras que la violencia intrafamiliar, evaluada de forma individual, no lo hace de manera estadísticamente significativa (p = 0,1786). Este resultado justifica construir un modelo VAR, dado que, si ninguna variable hubiera pasado esta prueba, no habría sustento estadístico para preferir un modelo multivariado sobre un modelo univariado como el SARIMA explicado previamente.

Tabla 16
Prueba de causalidad de Granger

Variable	F	p-valor	Causa a HOMICIDIOS (p<0,05)
ISE	3,1218	0,0009	SÍ
VIOLENCIA INTRAFAMILIAR	1,4019	0,1786	NO
INCAUTACION_KG	2,3323	0,0115	SÍ

Nota. Elaboración propia.

Con la evidencia de causalidad ya establecida, se ajustaron en total siete modelos VAR distintos con 12 rezagos cada uno, correspondientes a todas las combinaciones posibles de acompañar al homicidio con una, dos o las tres variables candidatas. Se evaluó además si los residuos de cada modelo cumplían los supuestos de independencia, normalidad y homocedasticidad sobre la ecuación que predice homicidios (Tabla 17).

Tabla 17
Modelos VAR considerados.

Especificación	AIC	BIC	p (LB)	Indep.	p (JB)	Norm.	p (ARCH)	Homoc.
H + ISE + IK	31	33	0,9363	Sí	0,1458	Sí	0,1562	Sí
H + VI + IK	42	45	0,6600	Sí	0,0760	Sí	0,2969	Sí

Especificación	AIC	BIC	p (LB)	Indep.	p (JB)	Norm.	p (ARCH)	Homoc.
H + ISE + VI + IK	45	49	0,4161	Sí	0,3120	Sí	0,3637	Sí
H + ISE + VI	26	28	0,8966	Sí	0,0000	No	0,1538	Sí
H + IK	28	29	0,9762	Sí	0,0000	No	0,8132	Sí
H + ISE	11	13	0,9982	Sí	0,0000	No	0,0406	No
H + VI	23	24	0,9443	Sí	2.00E−04	No	0,0099	No

Nota. H: Homicidios; ISE: Índice de Seguimiento a la Economía; VI: Violencia intrafamiliar; IK: Incautación (kg); LB: prueba de Ljung-Box; JB: prueba de Jarque-Bera.

El modelo seleccionado es el que tiene menor AIC y BIC, es decir Homicidios explicado por el Indicador de Seguimiento a la Economía ISE, si bien no cumple todos los supuestos, al tener menor AIC y BIC cuenta con buena capacidad predictiva relativa frente a las demás especificaciones. A partir de este modelo se realizó el pronóstico de enero a mayo de 2026 (Tabla 18 y Figura 16).

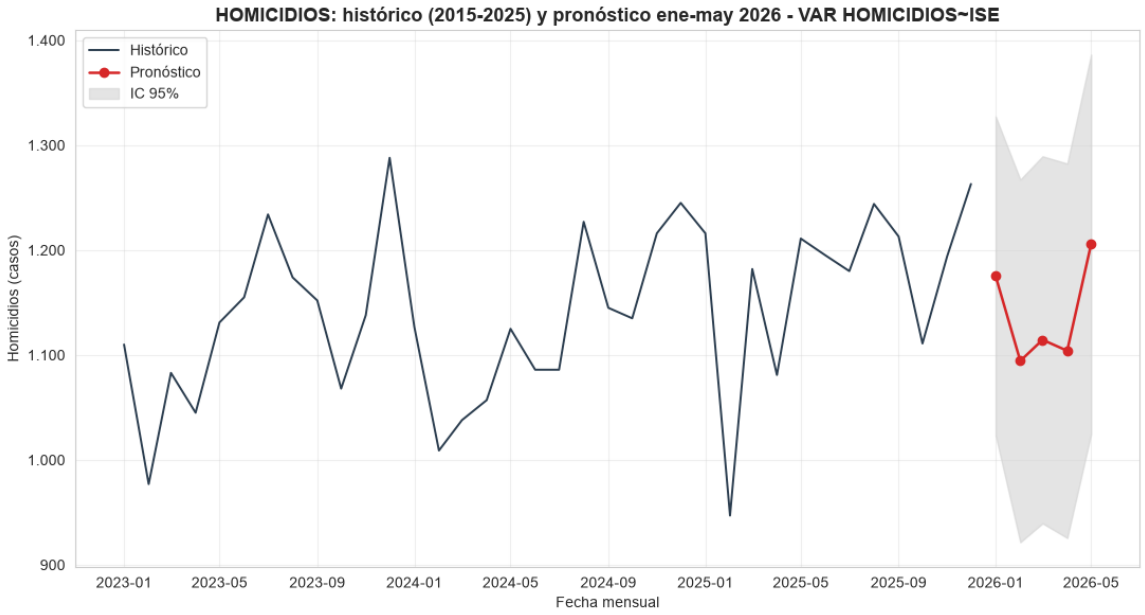
Tabla 18
Pronóstico puntual e intervalo de confianza del 95% del modelo VAR HOMICIDIOS~ISE, enero-mayo de 2026.

Fecha	Pronóstico	LI (95%)	LS (95%)
Ene. 2026	1.175,3	1.023,3	1.327,4
Feb. 2026	1.094,4	921,6	1.267,2
Mar. 2026	1.114,5	939,6	1.289,5
Abr. 2026	1.104,1	925,7	1.282,5
May. 2026	1.205,6	1.024,8	1.386,3

Nota. Elaboración propia.

Figura 16

Serie histórica de homicidios (2015-2025) y pronóstico del modelo VAR HOMICIDIOS~ISE para enero-mayo de 2026, con intervalo de confianza del 95%.



Nota. Elaboración propia.

Los resultados muestran que el nivel de homicidios de un mes está relacionado, de forma estadísticamente confiable, con el nivel del mes inmediatamente anterior, pero también con el comportamiento del ISE de uno, dos, cuatro y doce meses atrás. Es decir, los cambios en la actividad económica no se reflejan en el homicidio de inmediato, sino con distintos rezagos, y de manera particularmente marcada un año después, un patrón que coincide con la fuerte estacionalidad anual ya identificada en el análisis exploratorio del fenómeno (Sección 8.2).

El modelo VAR resulta ser el único de los tres modelos estadísticos clásicos que supera la prueba de independencia de residuales, sin embargo, las métricas de desempeño del modelo SARIMAX tienen mejor desempeño que las del modelo VAR especificado (Tabla 19).

Tabla 19
Métricas de desempeño del modelo VAR HOMICIDIOS~ISE frente al valor observado real de enero-mayo de 2026.

MAE	RMSE	MASE	ME	MAPE (%)	R^2
32,806	37,356	0,434	21,221	2,813	0,196

Nota. Elaboración propia.

8.3.4 Síntesis de los modelos estadísticos

Al contrastar los tres modelos estadísticos entre enero y mayo de 2026 (Tabla 20 y Tabla 21), SARIMAX(2,0,1)(0,1,2)[12] resulta ser el de mejor desempeño: un error absoluto medio (MAE) de 31,6 homicidios por mes (equivalente a un error porcentual (MAPE) de apenas 2,7% sobre un nivel mensual que ronda los 1.100-1.200 casos), y el coeficiente de determinación más alto ($R^2 = 0,37$), es decir, el modelo explica aproximadamente el 37% de la variabilidad que hay en los cinco valores de homicidios observados en 2026. El modelo VAR HOMICIDIOS~ISE ocupa el segundo lugar, con un desempeño cercano al de SARIMAX (MAE = 32,8; $R^2 = 0,20$), mientras que el SARIMA univariado, al no incorporar ninguna variable estructural adicional, presenta el error más alto de los tres (MAE = 41,0) y es el único con R^2 negativo (-0,30), es decir, el único cuyo error cuadrático supera al de la media de los cinco meses observados, que es la referencia sobre la que se calcula el R^2 y no debe confundirse con el pronóstico ingenuo estacional del MASE, al que SARIMA sí supera (MASE = 0,54).

Tabla 20
Pronósticos de enero a mayo de 2026 de los tres modelos estadísticos (SARIMA, SARIMAX y VAR).

Modelo	Ene. 2026	Feb. 2026	Mar. 2026	Abr. 2026	May. 2026
SARIMA(2,0,1)(0,1,2)[12]	1.179,91	1.012,48	1.133,24	1.090,45	1.197,27
SARIMAX(2,0,1)(0,1,2)[12]	1.219,33	1.040,81	1.134,91	1.114,92	1.211,08
VAR HOMICIDIOS~ISE	1.175,33	1.094,41	1.114,52	1.104,08	1.205,55

Modelo	Ene. 2026	Feb. 2026	Mar. 2026	Abr. 2026	May. 2026
Observado (2026)	1.203,00	1.083,00	1.170,00	1.156,00	1.188,00

Nota. Elaboración propia.

Tabla 21
Métricas de desempeño de los tres modelos estadísticos.

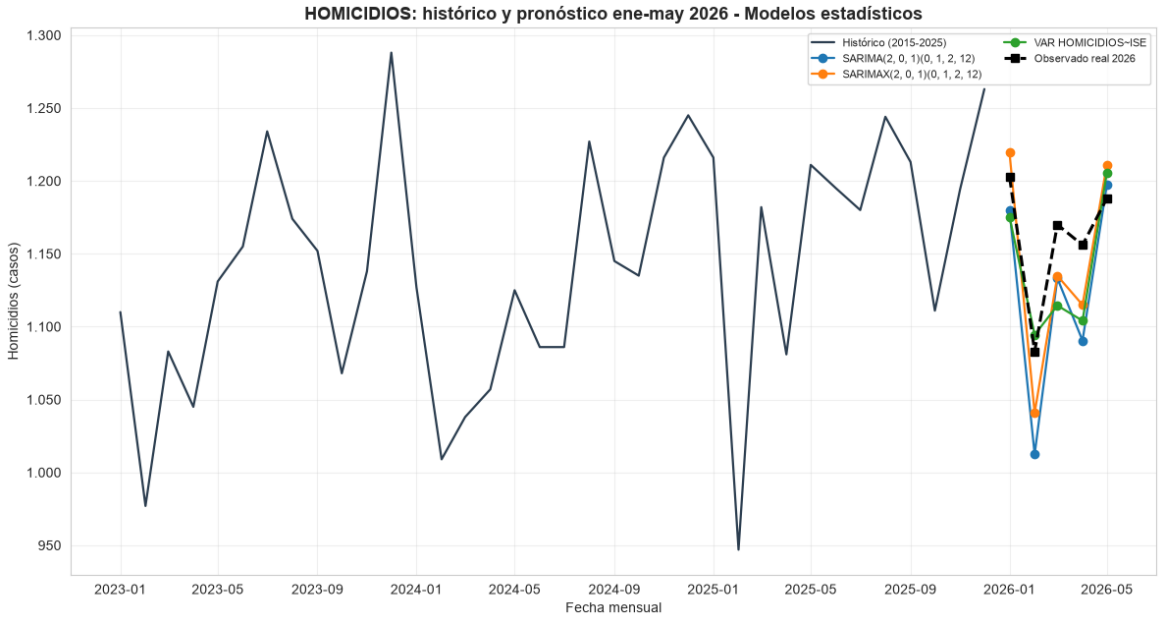
Modelo	MAE	RMSE	MASE	ME	MAPE (%)	R^2
SARIMA(2,0,1)(0,1,2)[12]	41,04	47,41	0,54	37,33	3,60	-0,30
SARIMAX(2,0,1)(0,1,2)[12]	31,55	33,16	0,42	15,79	2,75	0,37
VAR HOMICIDIOS~ISE	32,81	37,36	0,43	21,22	2,81	0,20

Nota. Elaboración propia.

El gráfico de la Figura 17 permite observar que los tres modelos logran capturar razonablemente bien la forma general de la trayectoria observada, el descenso pronunciado de enero a febrero y la recuperación hacia marzo, pero difieren notablemente en su capacidad de anticipar la magnitud de ese descenso. Febrero de 2026 es, con diferencia, el mes más difícil de pronosticar para los tres modelos: mientras el valor real fue de 1.083 homicidios, SARIMA subestimó fuertemente ese mínimo (1.012), SARIMAX se acercó más (1.041) y VAR fue el más preciso de los tres en ese punto específico (1.094). Este patrón es coherente con lo ya documentado sobre el error medio de cada modelo: SARIMA tiende a subestimar de forma sistemática el nivel de homicidios (ME = 37,3), SARIMAX lo hace en promedio pero con signo alternante entre meses (ME = 15,8), mientras que VAR, al capturar de forma explícita la retroalimentación dinámica entre homicidios y actividad económica, se ubica en un sesgo intermedio (ME = 21,2) sin dejar de ser, en el balance global de errores, ligeramente menos preciso que SARIMAX.

Figura 17

Comparación de los pronósticos de enero-mayo de 2026 de los tres modelos estadísticos frente a la serie histórica de homicidios.



Nota. Elaboración propia.

En conjunto, estos resultados sugieren que la incorporación de información estructural adicional, ya sea a través de variables exógenas (SARIMAX) o de una relación dinámica bidireccional con una variable macroeconómica (VAR), efectivamente mejora la capacidad de pronóstico del homicidio en Colombia frente a un enfoque puramente autorregresivo (SARIMA). La cercanía en el desempeño entre SARIMAX y VAR, pese a partir de estrategias de modelamiento distintas, el primero trata la variable exógena de forma unidireccional, el segundo permite una relación de doble vía, refuerza la conclusión de que el valor agregado no proviene de una técnica específica, sino de la incorporación misma de contexto económico y social al modelo, un hallazgo que se retoma con mayor profundidad en la comparación final frente a los modelos de aprendizaje automático.

8.4 Modelos de aprendizaje automático

La incorporación de modelos de aprendizaje automático (*Machine Learning*) como complemento a los modelos estadísticos clásicos (SARIMA, SARIMAX y VAR) responde a una limitación estructural de estos últimos: todos ellos asumen una forma funcional lineal predefinida entre la variable objetivo y sus rezagos o sus variables exógenas. Esta linealidad, si bien facilita la interpretación de los coeficientes y la evaluación formal de supuestos estadísticos, puede resultar restrictiva frente a un fenómeno social como el homicidio, cuyos determinantes (economía, conflictividad conexa, estacionalidad, choques externos como la pandemia) identificados en el análisis exploratorio (Sección 8.2), probablemente interactúan entre sí de formas no lineales y cambiantes en el tiempo, que un modelo lineal solo puede aproximar de manera parcial.

Los algoritmos de aprendizaje automático seleccionados para este trabajo, XGBoost, LightGBM, Random Forest, LSTM, GRU y Prophet, abordan esta limitación desde distintos ángulos, sin exigir al investigador especificar de antemano la forma exacta de la relación entre las variables. Los modelos de ensamble de árboles (XGBoost, LightGBM, Random Forest) pueden capturar automáticamente interacciones y umbrales no lineales entre los rezagos de homicidios y las variables exógenas, sin que ello implique la explosión en el número de parámetros que sufriría un VAR o un SARIMAX de complejidad equivalente. Las redes recurrentes (LSTM, GRU) están diseñadas específicamente para aprender dependencias secuenciales de largo alcance en datos ordenados en el tiempo, un supuesto de diseño directamente relevante para una serie con estacionalidad marcada como la de homicidios. Prophet, por su parte, ofrece un punto intermedio: mantiene la interpretabilidad de una descomposición aditiva explícita en tendencia y estacionalidad, como SARIMA, pero con mayor flexibilidad para incorporar variables exógenas y cambios estructurales en la tendencia.

Más allá de la capacidad técnica de estos algoritmos, existe una justificación metodológica adicional para incluirlos: la pregunta de si los modelos de aprendizaje automático superan a los modelos estadísticos clásicos en el pronóstico de fenómenos criminológicos no tiene, en la literatura, una respuesta única ni universal, el desempeño relativo depende fuertemente del tamaño de la muestra, la frecuencia de los datos y la naturaleza específica del fenómeno. Comparar ambas familias de modelos bajo un mismo esquema de evaluación (mismo periodo de entrenamiento, mismo conjunto de prueba real y externo, mismos criterios de error) permite generar evidencia empírica directamente aplicable al caso colombiano, en lugar de asumir a priori la superioridad de un enfoque sobre el otro. Este ejercicio comparativo constituye, en sí mismo, un aporte metodológico del presente trabajo: no busca reemplazar el análisis estadístico clásico, sino contrastarlo de forma rigurosa frente a alternativas de mayor flexibilidad funcional, documentando honestamente en qué escenarios cada familia de modelos resulta más adecuada, una discusión que se retoma con los resultados empíricos en la Sección 8.5 y en el Capítulo 9.

8.4.1 Modelo XGBoost

Para el modelo XGBoost se probaron cuatro hiperparámetros, cada uno con dos valores posibles: *n_estimators* (200 o 400) es la cantidad de "árboles de decisión" que el modelo combina para hacer su pronóstico final (más árboles suelen capturar patrones más finos, pero también aumentan el riesgo de memorizar el pasado en vez de generalizar); *max_depth* (3 o 5) controla qué tan profundo, es decir, cuántas preguntas encadenadas, puede hacer cada árbol individual antes de dar una respuesta (árboles más profundos son más flexibles, pero también más propensos a sobreajustarse); *learning_rate* (0,05 o 0,1) determina qué tan grande es el "paso de aprendizaje" que da el modelo en cada nuevo árbol que agrega (un paso más pequeño aprende de forma más cautelosa y estable, uno

más grande aprende más rápido pero con más riesgo de pasarse de la meta); y *subsample* (0,8 o 1,0) es la proporción de los datos de entrenamiento que se usa al azar para construir cada árbol, usar menos del 100% (0,8) introduce algo de variedad entre árboles, lo que suele ayudar a que el modelo generalice mejor.

Combinar estos cuatro hiperparámetros con sus dos valores posibles da 16 configuraciones distintas, y cada una se cruzó además con las 64 combinaciones posibles de variables exógenas candidatas, para un total de 1.024 modelos evaluados. Cada uno de esos 1.024 modelos se entrenó únicamente con datos de 2015 a 2023 y se puso a prueba pronosticando 2024-2025, un periodo que el modelo no vio durante su entrenamiento, precisamente para evitar quedarse con una configuración que simplemente memoriza el pasado en lugar de aprender un patrón que se sostenga hacia adelante. Esta partición rige únicamente la etapa de selección: una vez elegida la configuración ganadora, el modelo final se reajustó sobre la serie completa hasta diciembre de 2025, entrenamiento y validación combinados, antes de generar el pronóstico de 2026 (Sección 7.7).

Entre las 1.024 combinaciones evaluadas, la que menor error de validación obtuvo, medido como RMSE sobre el periodo 2024-2025, fue la que usa las variables exógenas lesiones personales, tasa de desempleo y violencia intrafamiliar, con 400 árboles (*n_estimators*=400), una profundidad máxima de 5 (*max_depth*=5), una tasa de aprendizaje de 0,1 (*learning_rate*=0.1) y un submuestreo del 80% de los datos por árbol (*subsample*=0.8), alcanzando un RMSE de validación de 51,52 homicidios (Tabla 22). Es "la mejor" en un sentido preciso y acotado: la que, de entre todas las combinaciones efectivamente probadas, logró el menor error al pronosticar un periodo que el modelo no usó para entrenarse, no una garantía de que sea la óptima entre todas las combinaciones matemáticamente posibles, ni de que necesariamente sea la que mejor se comporte frente

al observado real de 2026, ya que validación (2024-2025) y prueba final (2026) son dos ventanas de tiempo distintas.

Tabla 22
Mejores 10 combinaciones de variables exógenas e hiperparámetros identificados para XGBoost.

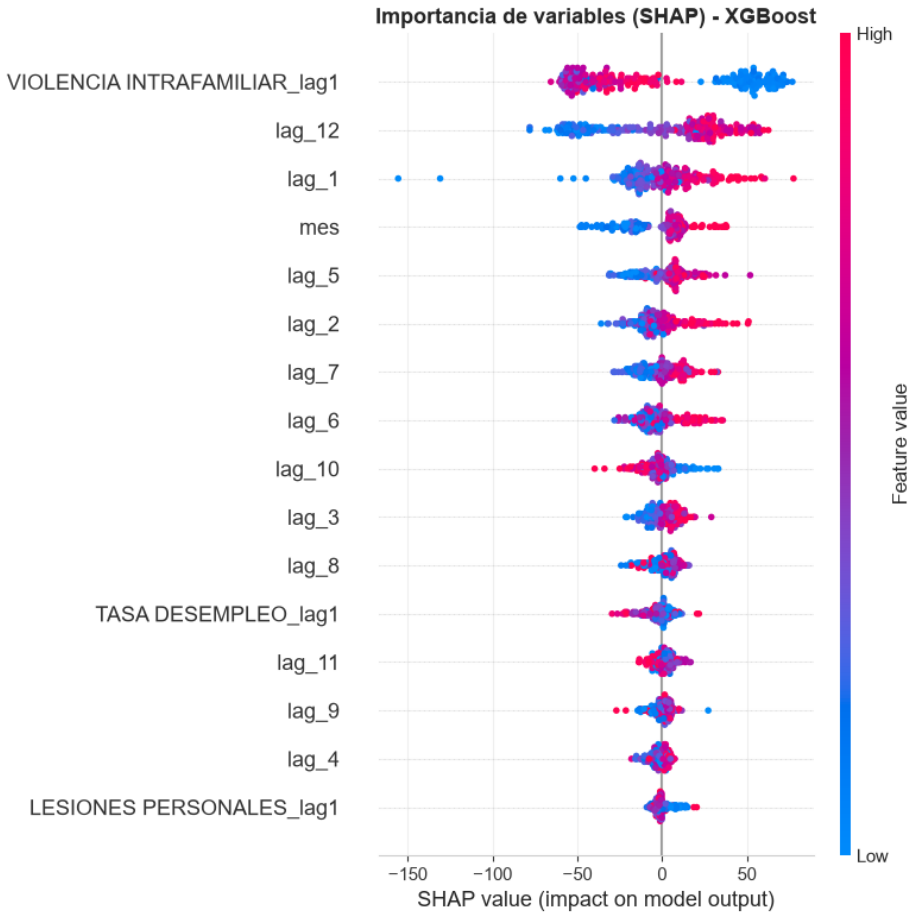
Variables	LR	Prof.	Árb.	Subm.	RMSE val.	SD resid.	n
LP + TD + VI	0,10	5	400	0,8	51,52	49,95	24
LP + TD + VI	0,10	5	200	0,8	51,53	49,93	24
ISE + TD + VI	0,10	3	400	0,8	55,93	55,88	24
ISE + TD + VI	0,10	3	200	0,8	56,92	56,91	24
LP + VI	0,10	5	200	0,8	57,38	56,57	24
LP + VI	0,10	5	400	0,8	57,46	56,64	24
HP + LP + ISE + TD + VI	0,10	3	400	1,0	57,84	57,75	24
LP + ISE + TD + VI	0,05	3	400	1,0	58,19	57,87	24
HP + TD + VI	0,10	5	400	0,8	58,71	57,40	24
HP + TD + VI	0,10	5	200	0,8	58,71	57,40	24

Nota. LP: Lesiones personales; TD: Tasa de desempleo; VI: Violencia intrafamiliar; HP: Hurto a personas; ISE: Índice de Seguimiento a la Economía; **LR**: learning rate; **Prof.**: profundidad máxima (max_depth); **Árb.**: número de árboles (n_estimators); **Subm.**: fracción de muestreo (subsample); **RMSE val.**: RMSE de validación; **SD resid.**: desviación estándar de los residuos de validación.

El análisis de importancia de variables mediante valores SHAP (*Shapley Additive explanations) (*Figura 18*)* permite identificar qué información utiliza el modelo XGBoost para construir su pronóstico y en qué dirección influye cada variable. Las cuatro variables más influyentes son, en orden, la violencia intrafamiliar del mes anterior, el rezago de doce meses de la propia serie de homicidios, el rezago de un mes y el número de mes calendario. El rezago de doce meses muestra una relación positiva y consistente, niveles altos de homicidios un año atrás empujan el pronóstico hacia arriba, y niveles bajos lo empujan hacia abajo, lo que confirma dentro del propio modelo la estacionalidad anual ya

identificada en la descomposición de la serie (Sección 8.2). El rezago de un mes exhibe un patrón similar, reflejando la persistencia de corto plazo del fenómeno.

Figura 18
Importancia de variables (valores SHAP) del modelo XGBoost.



Nota. Elaboración propia.

La violencia intrafamiliar, en cambio, presenta una relación negativa: meses con niveles altos de este delito el mes anterior se asocian, según el modelo, con un pronóstico menor de homicidios, mientras que niveles bajos lo empujan hacia arriba. Este resultado es contraintuitivo frente a la expectativa de que ambos fenómenos de violencia se muevan en la misma dirección, y debe interpretarse con cuidado: podría reflejar una dinámica estadística real, por ejemplo, patrones estacionales opuestos entre ambos delitos, o ser

una asociación espuria derivada del tamaño limitado de la muestra (132 observaciones mensuales), por lo que no debe leerse como evidencia de una relación causal. El resto de los rezagos y las dos variables exógenas restantes (tasa de desempleo y lesiones personales, ambas con un mes de rezago) muestran una dispersión considerablemente menor, es decir, aportan al pronóstico con un peso individual más modesto que las cuatro variables principales.

El modelo XGBoost basa su pronóstico en la violencia intrafamiliar del mes anterior, la variable individual de mayor peso, y en la memoria de la propia serie (rezago de un año y del mes anterior) y en la estacionalidad calendario, aunque con una dirección que merece discusión cualitativa, no solo reportarse como un dato más, mientras que la tasa de desempleo y las lesiones personales aportan una señal secundaria.

A partir de este modelo se realizó el pronóstico de enero a mayo de 2026 (Tabla 23 y Figura 19).

Tabla 23
Pronóstico puntual e intervalo aproximado del 95% de XGBoost, enero-mayo de 2026.

Modelo	Fecha mensual	Pronóstico	Límite inf. (95%)	Límite sup. (95%)
XGBoost	2026-01-01	1.146,0	1.048,1	1.243,9
XGBoost	2026-02-01	1.055,2	957,3	1.153,1
XGBoost	2026-03-01	1.120,7	1.022,8	1.218,6
XGBoost	2026-04-01	1.084,0	986,1	1.181,9
XGBoost	2026-05-01	1.127,7	1.029,8	1.225,6

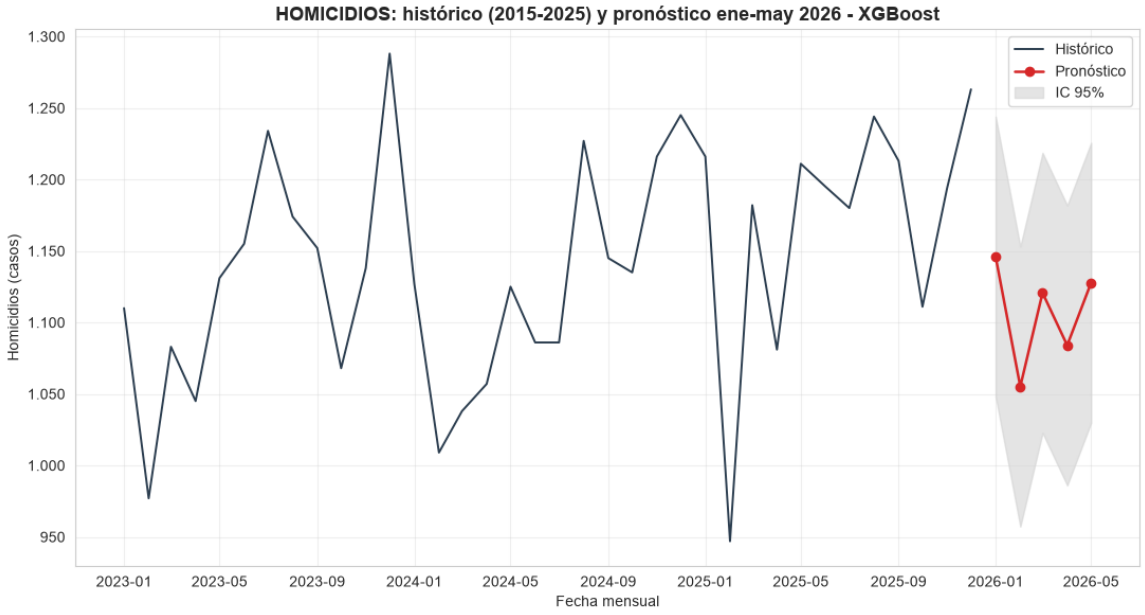
Nota. Elaboración propia.

El pronóstico de XGBoost para enero-mayo de 2026 mantiene un nivel general consistente con el rango histórico reciente de la serie, pero con una oscilación pronunciada

mes a mes: parte de 1.146 casos en enero, cae abruptamente a 1.056 en febrero, una caída que, aunque en la dirección correcta, subestima la magnitud de la caída real observada en 2026, repunta a 1.121 en marzo, vuelve a bajar a 1.084 en abril y cierra en 1.128 en mayo. Este patrón de "zigzag" pronunciado, en contraste con la trayectoria históricamente más suave del último tramo de la serie (2025), es característico del pronóstico recursivo mes a mes de los modelos de árboles: cada predicción se construye sobre la anterior, lo que puede amplificar pequeñas oscilaciones en vez de suavizarlas.

Figura 19

Serie histórica de homicidios (2015-2025) y pronóstico de XGBoost para enero-mayo de 2026 con intervalo aproximado del 95 %.



Nota. Elaboración propia.

El intervalo de confianza aproximado (construido a partir de la desviación estándar de los residuales de validación, dado que XGBoost no genera intervalos de predicción nativos) es considerablemente amplio, de alrededor de 950 a 1.350 casos en varios meses,

lo que refleja la incertidumbre real de este pronóstico y es coherente con el desempeño débil ya documentado en las métricas de la tabla anterior (R^2 = -0,76).

Al revisar las métricas de desempeño del modelo contrastando contra el homicidio observado en enero-mayo de 2026 (Tabla 24), XGBoost presenta un error absoluto medio (MAE) de 53,3 homicidios por mes, equivalente a un error porcentual (MAPE) de 4,6% sobre un nivel mensual que ronda los 1.100-1.200 casos. El error medio (ME = 53,3), idéntico al MAE, indica que el modelo subestimó de forma sistemática y sin compensación el nivel de homicidios en todo el horizonte de prueba: nunca sobreestimó, siempre se quedó por debajo del valor real. El MASE de 0,705, aunque por debajo de 1, lo que indica que el modelo aún supera a un pronóstico ingenuo estacional (rezago de 12 meses), es considerablemente peor que el de los modelos estadísticos ya presentados (SARIMAX = 0,42; VAR = 0,43), y el R^2 negativo (-0,76) confirma que, en este horizonte de prueba específico, XGBoost se equivoca más que si simplemente se hubiera pronosticado el promedio de los cinco meses observados. Este resultado es consistente con lo señalado en el marco metodológico de la Sección 8.4: pese a la búsqueda exhaustiva de hiperparámetros (1.024 combinaciones evaluadas), el modelo parece sobreajustarse a patrones del histórico de entrenamiento (2015-2025) que no se repitieron con la misma intensidad en el periodo de prueba, una limitación típica de los modelos de árboles cuando se entrenan con series de tamaño reducido.

Tabla 24
Métricas de desempeño de XGBoost frente al valor observado real de enero-mayo de 2026.

Modelo	MAE	RMSE	MASE	ME	MAPE	R2
XGBoost	53,278	55,262	0,705	53,278	4,564	-0,760

Nota. Elaboración propia.

Pese a la búsqueda exhaustiva de hiperparámetros, XGBoost obtiene un desempeño modesto frente al observado real: MAE = 53,3, MAPE = 4,6% y un R^2 negativo (-0,76). Este resultado, inferior al de LightGBM y Random Forest, entrenados sobre la misma familia de variables, sugiere que, en este escenario de datos limitados (132 observaciones mensuales), XGBoost tiende a sobreajustarse a patrones del periodo de entrenamiento que no se repiten con la misma intensidad en el horizonte de prueba.

8.4.2 Modelo LightGBM

Al igual que con XGBoost, en LightGBM se probaron varios hiperparámetros con dos valores posibles cada uno: *n_estimators* (200 o 400) es la cantidad de árboles de decisión pequeños que el modelo combina para producir su pronóstico final; *max_depth* (3 o 5) limita qué tan profundo puede crecer cada árbol, es decir, cuántas divisiones sucesivas puede hacer antes de llegar a una respuesta; y *learning_rate* (0,05 o 0,1) controla qué tan grande es el ajuste que aporta cada nuevo árbol (valores más bajos aprenden de forma más cautelosa y estable, a costa de necesitar más árboles para lograr el mismo efecto).

La diferencia respecto a XGBoost está en *num_leaves* (15 o 31), un hiperparámetro propio de LightGBM: en vez de limitar la profundidad del árbol de forma pareja en todas sus ramas, limita directamente el número total de "hojas" (los puntos finales del árbol, donde efectivamente se produce un número como salida), permitiendo que el árbol crezca de forma más irregular, más ramificado en las zonas donde eso ayuda a explicar mejor los datos, y menos en las que no. Es, en la práctica, la manera que tiene LightGBM de controlar la complejidad del modelo con un poco más de flexibilidad que limitar solo la profundidad, lo cual explica en parte por qué suele entrenar más rápido que otros algoritmos de árboles con un desempeño comparable.

Al igual que con XGBoost, estos cuatro hiperparámetros (16 combinaciones) se cruzaron con las 64 combinaciones posibles de variables exógenas, para un total de 1.024 modelos LightGBM evaluados, cada uno entrenado con datos de 2015 a 2023 y puesto a prueba pronosticando el periodo 2024-2025. Entre todos ellos, la combinación con menor error de validación (RMSE = 60,43) fue la que usa como variables exógenas hurto a personas, ISE, tasa de desempleo y violencia intrafamiliar (Tabla 25), es decir, las cuatro candidatas con mayor contenido estructural, excluyendo únicamente incautación de estupefacientes, con una tasa de aprendizaje conservadora de 0,05, una profundidad máxima de 3, 200 árboles y un máximo de 15 hojas por árbol. Es "la mejor" en el mismo sentido acotado que ya se señaló para XGBoost: la de menor error al pronosticar un periodo que el modelo no vio durante su entrenamiento, entre las combinaciones efectivamente evaluadas, no una garantía de optimalidad absoluta ni de que sea, necesariamente, la que mejor se comporte frente al observado real de 2026, dado que validación (2024-2025) y prueba final (2026) son ventanas de tiempo distintas. Esta partición rige únicamente la etapa de selección: una vez elegida la configuración ganadora, el modelo final se reajustó sobre la serie completa hasta diciembre de 2025, entrenamiento y validación combinados, antes de generar el pronóstico de 2026 (Sección 7.7).

Tabla 25
Mejores 10 combinaciones de variables exógenas e hiperparámetros identificados para LightGBM.

Variables	LR	Prof.	Árb.	Hojas	RMSE val.	SD resid.	n
HP + ISE + TD + VI	0,05	3	200	31	60,43	60,37	24
HP + ISE + TD + VI	0,05	3	200	15	60,43	60,37	24
HP + ISE + TD + VI	0,10	3	200	15	60,48	60,05	24
HP + ISE + TD + VI	0,10	3	200	31	60,48	60,05	24
LP + IK + TD + VI	0,05	3	200	31	60,49	60,47	24
LP + IK + TD + VI	0,05	3	200	15	60,49	60,47	24

Variables	LR	Prof.	Árb.	Hojas	RMSE val.	SD resid.	n
HP + ISE + TD	0,10	3	400	31	60,90	60,75	24
HP + ISE + TD	0,10	3	400	15	60,90	60,75	24
HP + LP + ISE + TD + VI	0,05	3	200	31	60,95	60,82	24
HP + LP + ISE + TD + VI	0,05	3	200	15	60,95	60,82	24

Nota. HP = Hurto a personas, ISE = ISE, TD = Tasa desempleo, VI = Violencia intrafamiliar, LP = Lesiones personales, IK = Incautación_kg. Elaboración propia.

Por su parte, el análisis SHAP del modelo LightGBM con la mejor combinación de hiperparámetros (Figura 20) confirma el mismo patrón estructural ya observado en XGBoost: la variable más influyente es, nuevamente, la violencia intrafamiliar del mes anterior, seguida del rezago de doce meses de la propia serie de homicidios y del número de mes calendario. El rezago de doce meses muestra una relación positiva y consistente, niveles altos hace un año empujan el pronóstico hacia arriba, niveles bajos lo empujan hacia abajo, reforzando la estacionalidad anual ya documentada en la descomposición de la serie (Sección 8.2). El mes calendario exhibe el mismo comportamiento: meses de "valor alto" (identificados internamente por el modelo) desplazan el pronóstico hacia arriba, mientras que los de "valor bajo" lo hacen hacia abajo. El rezago de un mes (lag_1), en cambio, muestra una dispersión más amplia y menos ordenada por color, lo que indica que su efecto sobre el pronóstico es más variable de un mes a otro, sin un patrón único y consistente.

Figura 20
Importancia de variables (valores SHAP) del modelo LightGBM.



Nota. Elaboración propia.

La violencia intrafamiliar repite aquí la misma relación negativa y contraintuitiva que ya se señaló para XGBoost: niveles altos el mes anterior se asocian con un pronóstico menor de homicidios, y niveles bajos con uno mayor. Es un resultado que, dada su recurrencia en dos modelos independientes entrenados con la misma información, merece discutirse explícitamente como un hallazgo consistente del ejercicio, sin que ello implique necesariamente una relación causal, dado el tamaño limitado de la muestra (132 meses). De las variables exógenas que integran esta especificación (hurto a personas, ISE, tasa de desempleo y violencia intrafamiliar), el hurto a personas es la segunda con mayor peso

individual después de la violencia intrafamiliar, mientras que la tasa de desempleo y el ISE aportan una influencia considerablemente menor, similar en magnitud a la de los rezagos intermedios de la propia serie (lag_3 a lag_8).

El pronóstico de LightGBM para enero-mayo de 2026 (Tabla 26 y Figura 21) replica de forma notablemente cercana el patrón de altibajos observado en la serie real: parte de 1.183 casos en enero, cae a 1.075 en febrero, capturando correctamente tanto la dirección como buena parte de la magnitud de la caída mensual característica de ese mes, repunta con fuerza a 1.197 en marzo, retrocede levemente a 1.144 en abril y vuelve a subir a 1.194 en mayo.

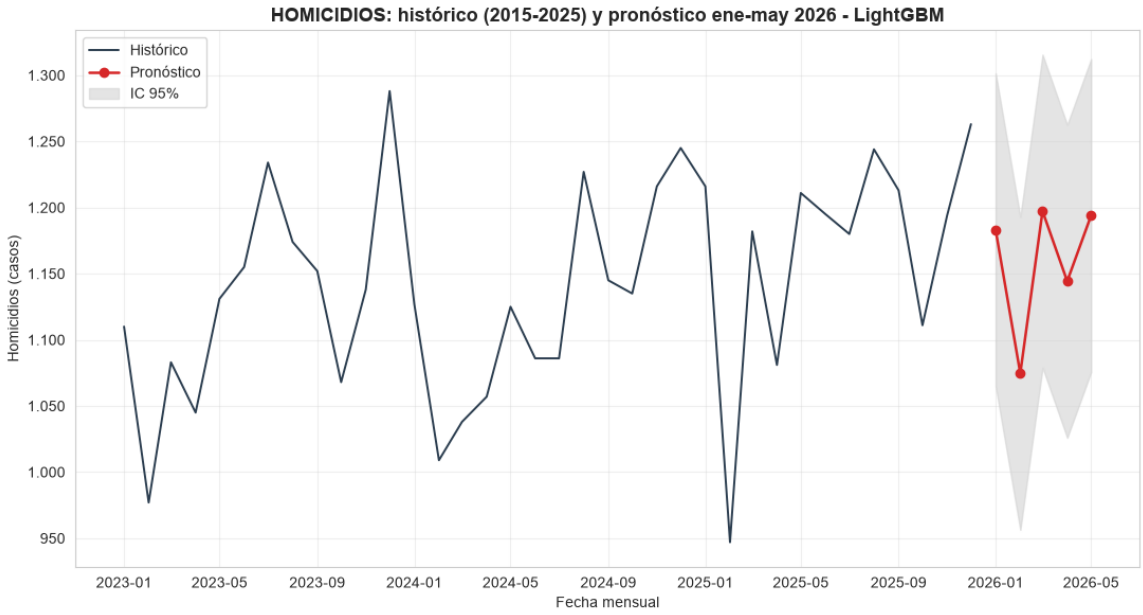
Tabla 26
Pronóstico puntual e intervalo aproximado del 95% de LightGBM, enero-mayo de 2026.

Modelo	Fecha mensual	Pronóstico	Límite inf. (95%)	Límite sup. (95%)
LightGBM	2026-01-01	1.183,2	1.064,8	1.301,5
LightGBM	2026-02-01	1.074,6	956,3	1.193,0
LightGBM	2026-03-01	1.197,3	1.078,9	1.315,6
LightGBM	2026-04-01	1.144,3	1.026,0	1.262,6
LightGBM	2026-05-01	1.193,9	1.075,6	1.312,2

Nota. Elaboración propia.

Figura 21

Serie histórica de homicidios (2015-2025) y pronóstico de LightGBM para enero-mayo de 2026 con intervalo aproximado del 95 %.



Nota. Elaboración propia.

El intervalo de confianza aproximado, aunque igual de amplio que en los demás aprendizaje automático, dado que se construye a partir de la desviación estándar de los residuales de validación, al no contar estos algoritmos con intervalos de predicción nativos, contiene en todos los casos el valor real observado, lo que refuerza la solidez del pronóstico puntual de este modelo frente al resto de especificaciones evaluadas hasta el momento.

Al revisar las métricas de desempeño de modelo (Tabla 27), LightGBM obtiene, por un margen amplio, el mejor desempeño de todos los modelos evaluados hasta el momento: un error absoluto medio (MAE) de apenas 14,6 homicidios por mes, equivalente a un error porcentual (MAPE) de 1,25% sobre un nivel mensual que ronda los 1.100-1.200 casos. El MASE de 0,194 (muy por debajo de 1) indica que el modelo supera ampliamente a un pronóstico ingenuo estacional, y es el valor más bajo registrado en el estudio hasta ahora.

Tabla 27
Métricas de desempeño de LightGBM frente al valor observado real de enero-mayo de 2026.

Modelo	MAE	RMSE	MASE	ME	MAPE	R2
LightGBM	14,615	16,605	0,194	1,348	1,252	0,841

Nota. Elaboración propia.

El error medio (ME = 1,35), prácticamente nulo y mucho menor que el MAE, revela que el modelo no presenta un sesgo sistemático relevante: sus errores se distribuyen de forma casi equilibrada entre sobreestimaciones y subestimaciones mes a mes, en contraste con el sesgo marcado y unidireccional observado en SARIMA, SARIMAX y XGBoost. El R^2 de 0,841 confirma este desempeño superior, indicando que el modelo explica más del 84% de la variabilidad del homicidio observado en el periodo de prueba y convierte a LightGBM en el modelo de referencia de todo el ejercicio comparativo hasta este punto del capítulo.

8.4.3 Modelo Random Forest

Random Forest funciona de forma distinta a XGBoost y LightGBM: en vez de construir los árboles uno tras otro, cada uno corrigiendo los errores del anterior, entrena muchos árboles de forma independiente y simultánea, cada uno sobre una muestra distinta de los datos, y al final promedia sus pronósticos. Por eso su grilla de hiperparámetros es más simple: *n_estimators* (300 o 500) es la cantidad de árboles independientes que se promedian, más árboles suelen dar un pronóstico más estable, a cambio de mayor tiempo de cómputo, *max_depth* (5, 10, o sin límite) controla qué tan profundo puede crecer cada árbol individual, es decir, cuántas divisiones sucesivas puede hacer antes de dar una respuesta.

El hiperparámetro *min_samples_leaf* (1 o 3) es propio de este tipo de bosque: define el número mínimo de observaciones que debe tener cada "hoja" final del árbol (el punto donde efectivamente se produce el pronóstico). Con *min_samples_leaf*=1, un árbol puede llegar a aislar un solo mes específico en su propia hoja, el árbol más flexible posible; con *min_samples_leaf*=3, se le exige que cada hoja agrupe al menos tres meses antes de dar una respuesta, lo que produce árboles más simples y menos propensos a memorizar casos aislados del pasado.

Con tres hiperparámetros de dos o tres valores cada uno (12 combinaciones) cruzados con las 64 combinaciones de variables exógenas, se evaluaron 768 modelos Random Forest en total, cada uno entrenado con datos de 2015 a 2023 y puesto a prueba pronosticando el periodo 2024-2025 (Tabla 28). La combinación con menor error de validación (RMSE = 66,8) fue la que usa como variables exógenas incautación de estupefacientes, ISE y violencia intrafamiliar, con una profundidad máxima de 10, un mínimo de una sola observación por hoja y 300 árboles. Al igual que en los modelos de árboles anteriores, es "la mejor" en un sentido acotado: la de menor error al pronosticar un periodo que el modelo no vio durante el entrenamiento, entre las combinaciones efectivamente evaluadas, no una garantía de optimalidad absoluta, y tampoco una certeza de que sea la que mejor se comporte frente al observado real de 2026, dado que validación (2024-2025) y prueba final (2026) son ventanas de tiempo distintas. Esta partición rige únicamente la etapa de selección: una vez elegida la configuración ganadora, el modelo final se reajustó sobre la serie completa hasta diciembre de 2025, entrenamiento y validación combinados, antes de generar el pronóstico de 2026 (Sección 7.7).

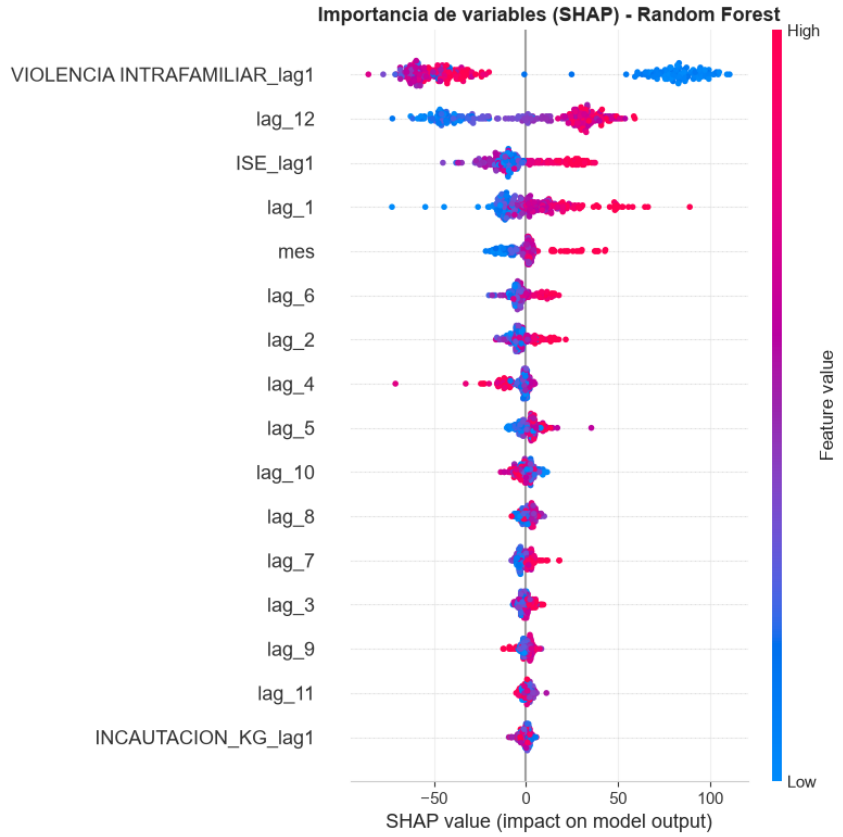
Tabla 28
Mejores 10 combinaciones de variables exógenas e hiperparámetros identificados para Random Forest.

Variables	Prof.	Hojas mín.	Árb.	RMSE val.	SD resid.	n
IK + ISE + VI	10	1	300	66,80	64,75	24
IK + ISE + VI	10	1	500	66,94	64,53	24
ISE + TD + VI	Sin límite	1	300	66,95	64,36	24
IK + ISE + TD + VI	Sin límite	1	300	66,98	64,44	24
IK + ISE + VI	Sin límite	1	300	67,06	64,99	24
IK + ISE + TD + VI	10	1	300	67,32	64,73	24
IK + ISE + VI	Sin límite	1	500	67,36	64,98	24
ISE + VI	10	1	500	67,37	64,96	24
ISE + VI	Sin límite	1	500	67,38	64,92	24
ISE + VI	10	1	300	67,41	65,27	24

Nota. IK = Incautación_kg, ISE = ISE, TD = Tasa desempleo, VI = Violencia intrafamiliar. Elaboración propia.

El análisis SHAP del modelo Random Forest con la mejor combinación de hiperparámetros (Figura 22) confirma nuevamente el mismo patrón estructural: la violencia intrafamiliar del mes anterior es la variable más influyente, seguida del rezago de doce meses de la propia serie de homicidios. El rezago de doce meses vuelve a mostrar una relación positiva y consistente, niveles altos hace un año empujan el pronóstico hacia arriba, niveles bajos lo empujan hacia abajo, reforzando por tercera vez consecutiva (tras XGBoost y LightGBM) la estacionalidad anual ya documentada en el análisis exploratorio. A diferencia de los dos modelos anteriores, aquí aparece con peso propio el ISE del mes anterior como tercera variable más importante, con una relación mayormente positiva: valores altos de actividad económica (rojo) tienden a desplazar el pronóstico hacia arriba, aunque con más dispersión y casos mixtos que en los rezagos autorregresivos, lo que indica una relación menos determinística.

Figura 22
Importancia de variables (valores SHAP) del modelo Random Forest.



Nota. Elaboración propia.

La violencia intrafamiliar mantiene la misma relación negativa y contraintuitiva observada en los dos modelos de árboles anteriores (niveles altos el mes previo se asocian con un pronóstico menor de homicidios), y su recurrencia en los tres modelos de esta familia (XGBoost, LightGBM y Random Forest), entrenados de forma independiente pero sobre la misma información, refuerza que se trata de un patrón sistemático capturado por los datos y no de una particularidad de un único algoritmo; aun así, se mantiene la misma cautela ya señalada sobre no interpretarlo como evidencia de causalidad. El resto de las variables, incluida incautación de estupefacientes, la tercera exógena de esta especificación, y los rezagos intermedios de la serie (lag_2 a lag_11) muestran una

dispersión considerablemente menor, aportando al pronóstico con un peso individual mucho más modesto que las tres variables principales.

El pronóstico de Random Forest para enero-mayo de 2026 (Tabla 29 y Figura 23) muestra una trayectoria de altibajos más contenida que la de los demás modelos de árboles: parte de 1.161 casos en enero, cae a 1.089 en febrero, repunta a 1.154 en marzo, retrocede a 1.103 en abril y cierra en 1.149 en mayo. Aunque el modelo captura correctamente la dirección de cada movimiento mes a mes (caída en febrero, recuperación en marzo, nueva caída en abril, recuperación en mayo), la magnitud de sus oscilaciones es notablemente menor que la variación real observada en la serie histórica, lo que resulta en un pronóstico algo más "aplanado" que suaviza los picos y valles más pronunciados.

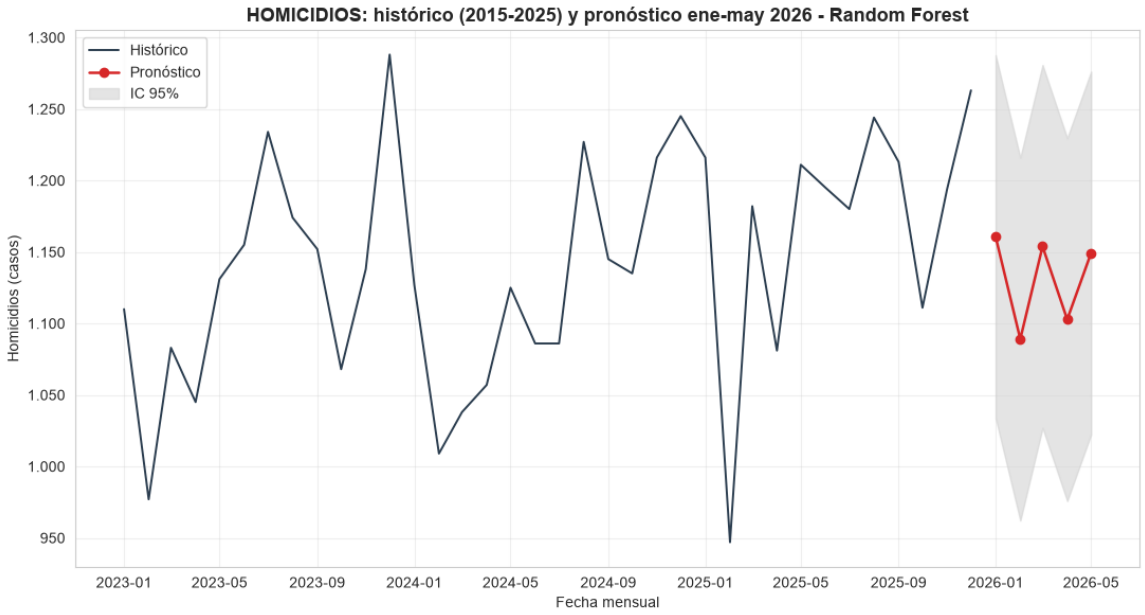
Tabla 29
Pronóstico puntual e intervalo aproximado del 95% de Random Forest, enero-mayo de 2026.

Modelo	Fecha mensual	Pronóstico	Límite inf. (95%)	Límite sup. (95%)
Random Forest	2026-01-01	1.160,7	1.033,8	1.287,6
Random Forest	2026-02-01	1.089,1	962,2	1.216,0
Random Forest	2026-03-01	1.153,9	1.027,0	1.280,8
Random Forest	2026-04-01	1.102,8	975,9	1.229,7
Random Forest	2026-05-01	1.149,2	1.022,3	1.276,1

Nota. Elaboración propia.

Figura 23

Serie histórica de homicidios (2015-2025) y pronóstico de Random Forest para enero-mayo de 2026 con intervalo aproximado del 95 %.



Nota. Elaboración propia.

El intervalo de confianza aproximado, de nuevo construido a partir de la desviación estándar de los residuales de validación, es amplio y prácticamente simétrico en torno al pronóstico puntual en los cinco meses, reflejando la incertidumbre inherente a un modelo que, aunque supera a SARIMA y a XGBoost en las métricas ya presentadas, no alcanza el nivel de precisión logrado por LightGBM.

Al revisar las métricas de desempeño de modelo (Tabla 30), Random Forest presenta un error absoluto medio (MAE) de 31,3 homicidios por mes, equivalente a un error porcentual (MAPE) de 2,67% sobre un nivel mensual que ronda los 1.100-1.200 casos, un desempeño prácticamente idéntico al de VAR HOMICIDIOS~ISE (MAE = 32,8) y notablemente mejor que SARIMA y XGBoost. El error medio (ME = 28,9), muy cercano al MAE, indica que el modelo subestima de forma sistemática y poco compensada el nivel de homicidios en el horizonte de prueba, un sesgo del mismo signo al observado en SARIMA

y SARIMAX, aunque con diferencias en su magnitud y en la alternancia de los errores mensuales, consistente con lo señalado en la interpretación del gráfico de pronóstico: el modelo suaviza los picos y valles de la serie, lo que en este periodo de prueba específico se traduce en un pronóstico por debajo del valor observado en cuatro de los cinco meses, con una subestimación promedio de 28,9 homicidios mensuales.

Tabla 30
Métricas de desempeño de Random Forest frente al valor observado real de enero-mayo de 2026.

Modelo	MAE	RMSE	MASE	ME	MAPE	R2
Random Forest	31,303	35,843	0,414	28,878	2,665	0,260

Nota. Elaboración propia.

El MASE de 0,414 confirma que el modelo supera ampliamente a un pronóstico ingenuo estacional, y el R^2 de 0,26, positivo, aunque considerablemente más bajo que el de LightGBM (0,84), indica que el modelo explica alrededor de una cuarta parte de la variabilidad del homicidio observado en 2026, ubicando a Random Forest como el tercer mejor modelo del estudio hasta este punto del capítulo, por debajo de LightGBM y SARIMAX pero por encima de SARIMA, VAR y XGBoost.

8.4.4 Modelo LSTM

LSTM (Long Short-Term Memory) es una red neuronal diseñada específicamente para aprender patrones en datos ordenados en el tiempo, y su forma de funcionar es distinta a la de los modelos de árboles ya vistos, así que también lo son sus hiperparámetros. *unidades* (16 o 32) es, en términos simples, el "tamaño" de la memoria interna de la red (cuántas piezas de información puede retener y combinar la red para reconocer patrones a lo largo de los meses); más unidades le dan más capacidad de aprendizaje, a costa de necesitar más datos para entrenarse bien sin memorizar de más;

dropout (0,0 o 0,2) es una técnica para evitar que la red memorice el pasado en vez de aprender un patrón general: en cada paso de entrenamiento, "apaga" al azar una fracción de esas unidades (0% o 20%, según el valor), obligando a la red a no depender demasiado de ninguna pieza de información en particular.

El hiperparámetro *epochs* (fijo en 60) es el número de veces que la red revisa el conjunto completo de datos de entrenamiento antes de terminar de ajustarse, cada repaso adicional le da otra oportunidad de mejorar, aunque después de cierto punto ya no aporta demasiado; *batch_size* (fijo en 8) es la cantidad de meses que la red procesa juntos antes de actualizar lo que ha aprendido, un valor más pequeño hace que el aprendizaje sea más gradual y detallado. Y *learning_rate* (fijo en 0,01) es, igual que en los modelos de árboles, qué tan grande es el ajuste que hace la red cada vez que corrige su aprendizaje. A diferencia de XGBoost, LightGBM y Random Forest, aquí solo se varió *unidades* y *dropout* (4 combinaciones), dejando fijos *epochs, batch_size y learning_rate* en valores razonables, una decisión práctica, ya que entrenar una red neuronal es mucho más costoso computacionalmente que entrenar un árbol de decisión, y probar una grilla tan amplia como la de los modelos anteriores habría sido poco viable en un equipo estándar.

Por el mismo motivo de costo computacional, la búsqueda de variables exógenas para LSTM tampoco evaluó las 64 combinaciones completas (Tabla 31): se limitó a un subconjunto informado (sin exógenas, con las seis exógenas completas, y con las combinaciones ya identificadas como mejores por XGBoost, LightGBM y Random Forest), cruzado con las 4 combinaciones de *unidades* y *dropout*.

Tabla 31
Mejores 10 combinaciones de variables exógenas e hiperparámetros identificados para LSTM.

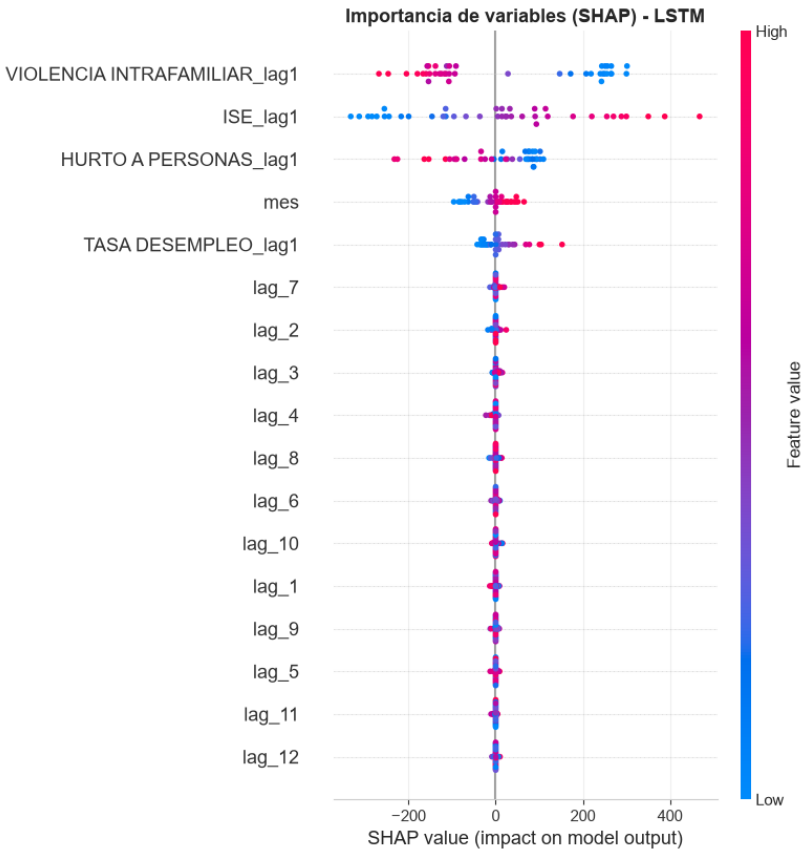
Variables	Batch	Dropout	Epochs	LR	Unid.	RMSE val.	SD resid.	n
HP + ISE + TD + VI	8	0,0	60	0,01	32	117,70	113,56	24
LP + TD + VI	8	0,0	60	0,01	16	133,23	120,63	24
HP + LP + IK + ISE + TD + VI	8	0,0	60	0,01	16	183,00	181,46	24
HP + ISE + VI	8	0,0	60	0,01	16	203,28	100,58	24
HP + ISE + TD + VI	8	0,2	60	0,01	32	207,41	103,83	24
HP + ISE + TD + VI	8	0,2	60	0,01	16	219,02	98,77	24
LP + TD + VI	8	0,2	60	0,01	32	224,06	141,03	24
LP + TD + VI	8	0,0	60	0,01	32	263,44	136,67	24
(sin exógenas)	8	0,0	60	0,01	16	286,50	263,73	24
(sin exógenas)	8	0,2	60	0,01	16	304,42	277,11	24

Nota. HP = Hurto a personas, ISE = ISE, TD = Tasa desempleo, VI = Violencia intrafamiliar, LP = Lesiones personales, IK = Incautación_kg. Elaboración propia.

Entre esas opciones, la de menor error de validación (RMSE = 117,7) fue la que usa como variables exógenas hurto a personas, ISE, tasa de desempleo y violencia intrafamiliar, con 32 unidades de memoria interna, sin dropout (0%), entrenada durante 60 repasos completos de los datos con un tamaño de lote de 8 y una tasa de aprendizaje de 0,01. Nótese que este RMSE de validación (117,7) ya es sustancialmente más alto que el de cualquiera de los modelos de árboles vistos hasta ahora (entre 51 y 67), una primera señal (antes incluso de llegar al pronóstico final contra 2026) de que la red neuronal está teniendo más dificultad para ajustarse bien con la cantidad de datos disponible (240 observaciones supervisadas, tras descontar los doce rezagos iniciales de la serie de la longitud de la serie disponible), consistente con la limitación ya señalada en el marco metodológico sobre el tamaño de muestra reducido para este tipo de arquitecturas.

El análisis SHAP de LSTM (Figura 24) revela un patrón de importancia notablemente distinto al de los modelos de árboles: las cinco variables con mayor influencia son, en este orden, el ISE, la violencia intrafamiliar, el hurto a personas, el mes calendario y la tasa de desempleo (es decir, las cuatro variables exógenas seleccionadas más el mes), mientras que los doce rezagos de la propia serie de homicidios (lag_1 a lag_12) quedan relegados a los últimos lugares, con una dispersión de valores SHAP mínima y prácticamente sin distinción visual entre ellos. Este resultado contrasta de forma marcada con XGBoost, LightGBM y Random Forest, donde los rezagos, en particular lag_1 y lag_12, dominaban claramente la importancia del modelo. Para LSTM, en cambio, la señal predictiva parece concentrarse casi por completo en las variables exógenas y en el mes calendario, sin que la memoria autorregresiva de la serie aporte una contribución individual comparable.

Figura 24
Importancia de variables (valores SHAP) del modelo LSTM.



Nota. Elaboración propia.

Esta diferencia estructural es coherente con el desempeño deficiente de LSTM ya documentado en las métricas de pronóstico: una red recurrente está diseñada precisamente para aprender la dependencia temporal de la propia serie a través de su arquitectura interna (las conexiones recurrentes), no necesariamente a través de los rezagos explícitos que sí se le entregan como variables de entrada; que estos aparezcan con tan poca importancia sugiere que el modelo no logró aprovechar esa información autorregresiva de forma efectiva, apoyándose en cambio de manera casi exclusiva en las variables exógenas (varias de las cuales, como se señaló al interpretar los modelos de árboles, mantienen relaciones cuya dirección requiere lectura cuidadosa). Este hallazgo

refuerza la explicación ya planteada sobre la limitación de las redes recurrentes profundas en escenarios de datos reducidos (132 observaciones mensuales): con tan pocos datos, el modelo parece no haber logrado aprender de forma robusta el componente secuencial que en principio debería ser su mayor fortaleza frente a los demás algoritmos evaluados.

El pronóstico de LSTM para enero-mayo de 2026 (Tabla 32 y Figura 25) muestra el patrón más problemático de todos los modelos presentados hasta ahora: una trayectoria de crecimiento sostenido y prácticamente lineal, de 1.372 casos en enero a 1.428 en mayo, que se desprende por completo del comportamiento oscilante característico de la serie histórica (2015-2025) y del propio observado real de 2026 (que se mantuvo entre 1.083 y 1.203 casos, sin ninguna tendencia ascendente). Todo el horizonte de pronóstico queda muy por encima del rango histórico reciente, y el intervalo de confianza, aunque amplio, entre aproximadamente 1.150 y 1.650 casos hacia el final del periodo, ni siquiera logra acercar su límite inferior al nivel donde realmente se ubicó el homicidio en esos cinco meses. Esta divergencia sistemática es la manifestación visual directa de la falla ya anticipada en el análisis SHAP: al no aprovechar de forma efectiva la memoria autorregresiva de la propia serie, el modelo parece haber aprendido, a partir del pronóstico recursivo mes a mes, una tendencia creciente espuria que no tiene respaldo en los datos reales, un modo de falla característico de las redes recurrentes profundas cuando se entrenan con series de tamaño reducido (132 observaciones mensuales).

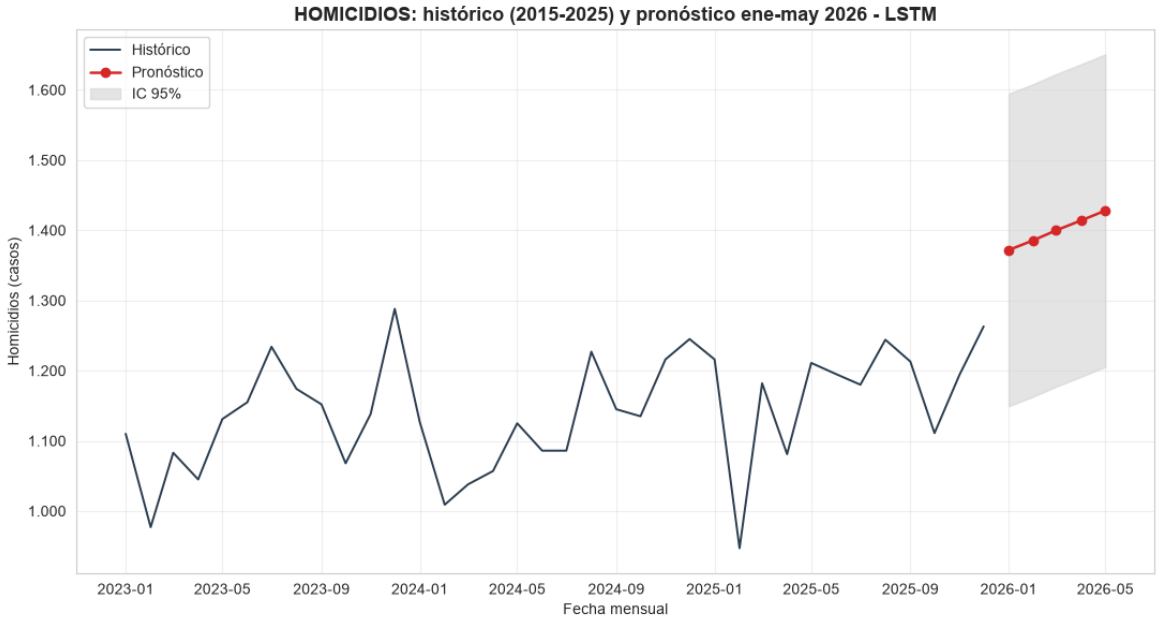
Tabla 32
Pronóstico puntual e intervalo aproximado del 95% de LSTM, enero-mayo de 2026.

Modelo	Fecha mensual	Pronóstico	Límite inf. (95%)	Límite sup. (95%)
LSTM	2026-01-01	1.371,8	1.149,2	1.594,4
LSTM	2026-02-01	1.385,9	1.163,3	1.608,4
LSTM	2026-03-01	1.399,9	1.177,3	1.622,5

Modelo	Fecha mensual	Pronóstico	Límite inf. (95%)	Límite sup. (95%)
LSTM	2026-04-01	1.414,0	1.191,4	1.636,6
LSTM	2026-05-01	1.428,0	1.205,4	1.650,6

Nota. Elaboración propia.

Figura 25
Serie histórica de homicidios (2015-2025) y pronóstico de LSTM para enero-mayo de 2026 con intervalo aproximado del 95 %.



Nota. Elaboración propia.

Al revisar las métricas de desempeño de modelo (Tabla 33), LSTM presenta el peor desempeño de todos los modelos evaluados hasta este punto del capítulo: un error absoluto medio (MAE) de 239,9 homicidios por mes, equivalente a un error porcentual (MAPE) de 20,8%, una magnitud de error muy superior a la de cualquier otro modelo, que hasta ahora no habían superado el 8%. El error medio (ME = -239,92), prácticamente idéntico al MAE en valor absoluto, confirma que el modelo sobreestimó en los cinco meses el nivel de homicidios en el horizonte de prueba: el signo negativo del error medio es la

firma directa de la trayectoria creciente y desprendida de la realidad ya observada en el gráfico de pronóstico.

Tabla 33
Métricas de desempeño de LSTM frente al valor observado real de enero-mayo de 2026.

Modelo	MAE	RMSE	MASE	ME	MAPE	R2
LSTM	239,920	243,824	3,177	-239,920	20,834	-33,253

Nota. Elaboración propia.

El MASE de 3,177, muy por encima de 1, indica que el modelo es más de tres veces peor que un pronóstico ingenuo estacional (simplemente repetir el valor de 12 meses atrás), y el R^2 de -33,25 confirma la magnitud del problema: el modelo se equivoca 33 veces más de lo que se equivocaría con solo pronosticar el promedio de los cinco meses observados. Este resultado extremo es consistente con las dos limitaciones concurrentes ya documentadas: el tamaño reducido de la serie frente a la cantidad de parámetros de una red recurrente, y la ausencia de escalamiento de las entradas (Sección 7.4), factor que se discute al cerrar la Sección 8.4.5; bajo esas condiciones, la red LSTM no logra generalizar de forma confiable y su pronóstico recursivo mes a mes termina amplificando el sesgo en lugar de corregirlo.

8.4.5 Modelo GRU

GRU (Gated Recurrent Unit) es, al igual que LSTM, una red neuronal diseñada para aprender patrones secuenciales en datos de series de tiempo, pero con una arquitectura interna más simple y liviana. Por eso comparte exactamente el mismo tipo de hiperparámetros que LSTM: *unidades* (16 o 32) es el tamaño de la "memoria" interna de la red (cuánta información puede retener y combinar mes a mes), con más unidades dándole mayor capacidad de aprendizaje a costa de necesitar más datos para no memorizar de

más; *dropout* (0,0 o 0,2) es una técnica que, durante el entrenamiento, "apaga" al azar una fracción de esa memoria interna (0% o 20%) para evitar que la red dependa demasiado de un patrón específico y generalice mejor a datos nuevos.

Los hiperparámetros *epochs* (fijo en 60), *batch_size* (fijo en 8) y *learning_rate* (fijo en 0,01) tienen el mismo significado que en LSTM: cuántas veces la red revisa por completo los datos de entrenamiento, cuántos meses procesa juntos antes de actualizar lo aprendido, y qué tan grande es cada ajuste que hace al corregirse, respectivamente. Al igual que con LSTM, solo se variaron *unidades* y *dropout* (4 combinaciones en total), dejando los demás fijos por el mismo motivo práctico: entrenar redes neuronales es considerablemente más costoso en tiempo de cómputo que entrenar modelos de árboles, así que se optó por una grilla más acotada.

Al igual que con LSTM, la búsqueda de variables exógenas para GRU se limitó a un subconjunto informado de combinaciones, cruzado con las 4 combinaciones de *unidades y dropout (*Tabla 34*)*.

Tabla 34
Mejores 10 combinaciones de variables exógenas e hiperparámetros identificados para GRU.

Variables	Batch	Dropout	Epochs	LR	Unid.	RMSE val.	SD resid.	n
HP + LP + IK + ISE + TD + VI	8	0,2	60	0,01	32	141,56	141,47	24
HP + ISE + TD + VI	8	0,0	60	0,01	16	166,60	64,66	24
HP + ISE + TD + VI	8	0,0	60	0,01	32	168,71	105,78	24
LP + TD + VI	8	0,0	60	0,01	32	174,82	124,94	24
LP + TD + VI	8	0,2	60	0,01	32	195,01	129,30	24
HP + ISE + TD + VI	8	0,2	60	0,01	32	206,45	101,65	24
HP + ISE + TD + VI	8	0,2	60	0,01	16	221,22	119,63	24
LP + TD + VI	8	0,0	60	0,01	16	245,53	139,12	24
(sin exógenas)	8	0,0	60	0,01	32	277,61	256,34	24

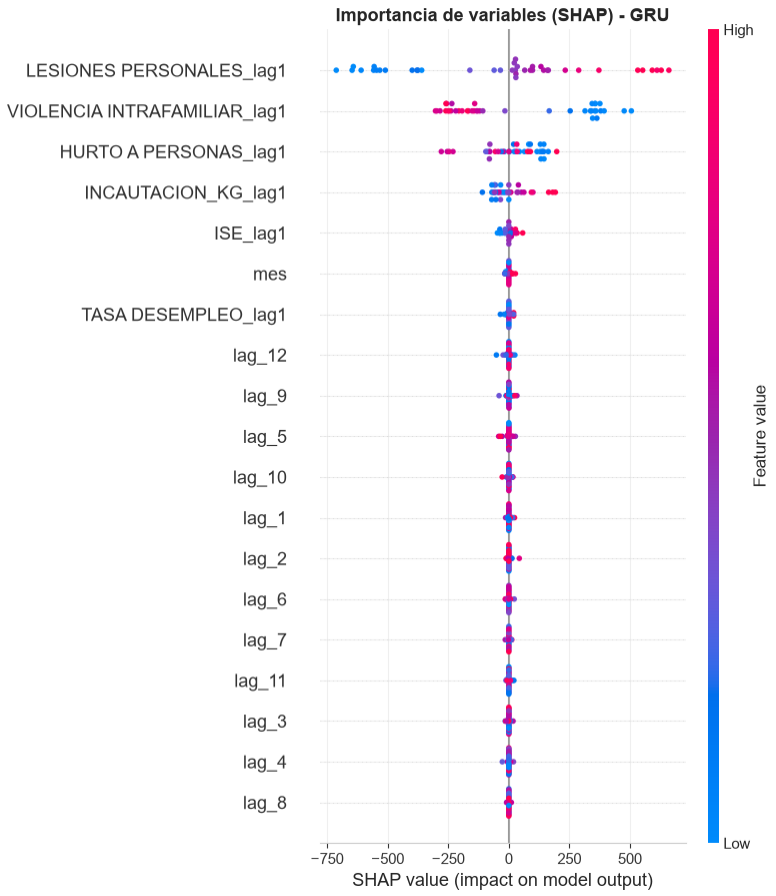
Variables	Batch	Dropout	Epochs	LR	Unid.	RMSE val.	SD resid.	n
HP + LP + IK + ISE + TD + VI	8	0,0	60	0,01	16	283,03	153,22	24

Nota. HP = Hurto a personas, ISE = ISE, TD = Tasa desempleo, VI = Violencia intrafamiliar, LP = Lesiones personales, IK = Incautación_kg. Elaboración propia.

La configuración con menor error de validación (RMSE = 141,56) fue, en este caso, la que usa las seis variables exógenas candidatas completas (hurto a personas, lesiones personales, incautación de estupefacientes, ISE, tasa de desempleo y violencia intrafamiliar), con 32 unidades de memoria interna, un dropout del 20%, entrenada durante 60 repasos completos de los datos con un tamaño de lote de 8 y una tasa de aprendizaje de 0,01. Este RMSE de validación, aunque más alto que el de cualquier modelo de árboles (entre 51 y 67), resulta más alto incluso que el de LSTM (117,7), una primera señal, antes de llegar al pronóstico final contra 2026, de que GRU podría tener más dificultad todavía para ajustarse bien con la cantidad de datos disponibles, algo que conviene verificar con las métricas de desempeño una vez calculadas.

El análisis SHAP de GRU (Figura 26) muestra un patrón similar al observado en LSTM: las cuatro variables con mayor influencia son: lesiones personales, violencia intrafamiliar, hurto a personas e incautación de estupefacientes (cuatro de las seis exógenas incluidas en esta especificación), mientras que los doce rezagos de la propia serie de homicidios (lag_1 a lag_12) y el mes calendario quedan relegados a los últimos lugares, con una dispersión de valores SHAP considerablemente menor.

Figura 26
Importancia de variables (valores SHAP) del modelo GRU.



Nota. Elaboración propia.

Nuevamente se repite el patrón ya observado en LSTM: la señal predictiva del modelo se concentra casi por completo en las variables exógenas, sin que la memoria autorregresiva explícita de la serie (los rezagos) aporte una contribución individual comparable, a pesar de que, al tratarse de una red recurrente, se esperaría que esa dependencia temporal quedara capturada principalmente en la arquitectura interna del modelo y no en los rezagos usados como variables de entrada.

Esta concentración de la importancia en las exógenas, común a ambas redes recurrentes (LSTM y GRU) pero ausente en los modelos de árboles, donde los rezagos,

en particular lag_1 y lag_12, dominaban claramente, sugiere que la limitación no es específica de un algoritmo particular sino de la familia de modelos recurrentes en su conjunto, al menos en las condiciones de este trabajo (240 observaciones supervisadas). Este hallazgo es coherente con el desempeño débil que ya mostró LSTM en el pronóstico contra 2026, y anticipa (antes incluso de revisar las métricas de GRU) que es probable que este modelo comparta una limitación similar: sin aprovechar de forma efectiva su propia memoria secuencial, ambas redes recurrentes parecen depender en exceso de variables exógenas cuya proyección futura, además, se aproximó mediante persistencia del último valor observado, una fuente adicional de incertidumbre que se traslada directamente al pronóstico.

El pronóstico de GRU para enero-mayo de 2026 (Tabla 35 y Figura 27) exhibe el patrón contrario al de LSTM: en lugar de sobreestimar, el modelo se instala en un nivel prácticamente plano de entre 861 y 869 casos mensuales, muy por debajo tanto del rango histórico reciente de la serie (~1.100-1.250) como del observado real de 2026 (~1.083-1.203).

Tabla 35
Pronóstico puntual e intervalo aproximado del 95% de GRU, enero-mayo de 2026.

Modelo	Fecha mensual	Pronóstico	Límite inf. (95%)	Límite sup. (95%)
GRU	2026-01-01	861,2	583,9	1.138,4
GRU	2026-02-01	863,0	585,7	1.140,3
GRU	2026-03-01	864,9	587,6	1.142,1
GRU	2026-04-01	866,7	589,4	1.144,0
GRU	2026-05-01	868,6	591,3	1.145,8

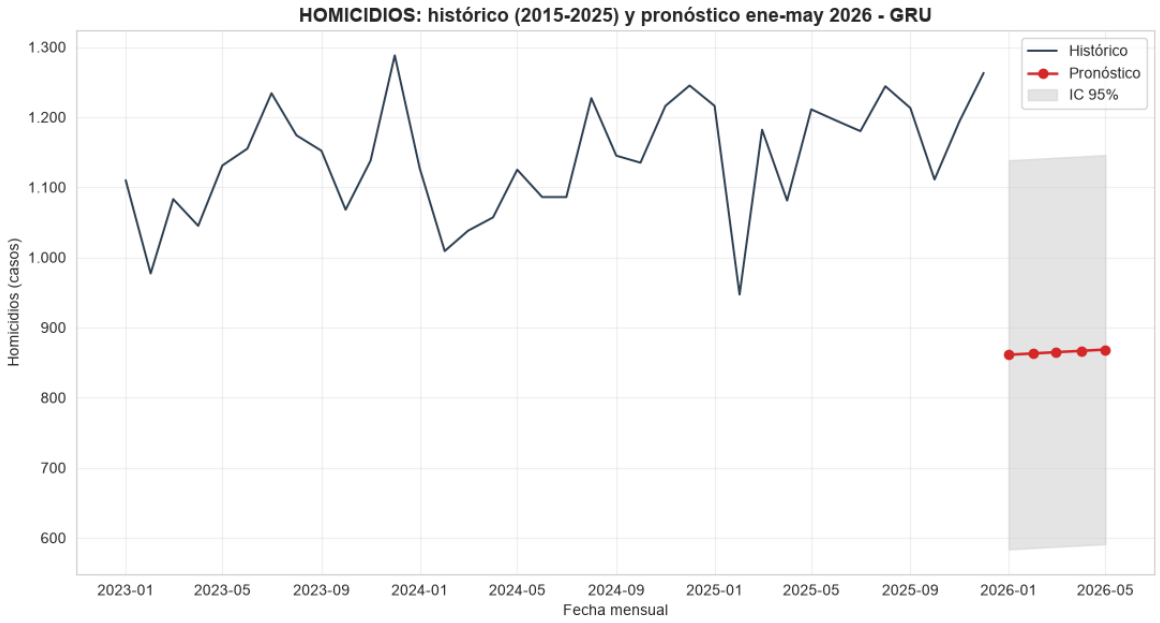
Nota. Elaboración propia.

A diferencia de la trayectoria creciente de LSTM, aquí el pronóstico apenas varía mes a mes, una diferencia de menos de 10 casos entre enero y mayo, lo que sugiere que

el modelo convergió a un nivel casi constante en lugar de aprender la dinámica oscilante propia de la serie.

Figura 27

Serie histórica de homicidios (2015-2025) y pronóstico de GRU para enero-mayo de 2026 con intervalo aproximado del 95 %.



Nota. Elaboración propia.

El intervalo de confianza aproximado, aunque amplio (entre aproximadamente 590 y 1.145 casos), tampoco logra que su límite superior se acerque de forma consistente al nivel donde realmente se ubicó el homicidio en esos cinco meses. Este resultado, en conjunto con el de LSTM, refuerza la conclusión ya planteada a partir del análisis SHAP: ambas redes recurrentes, entrenadas con apenas 240 observaciones supervisadas, no logran anclar su pronóstico de forma confiable al nivel real de la serie (una convergiendo hacia una sobreestimación sistemática, la otra hacia un nivel plano y subestimado), evidenciando dos modos de falla distintos, pero igualmente asociados a la limitación de tamaño de muestra para este tipo de arquitectura.

Al revisar las métricas de desempeño de modelo (Tabla 36), GRU presenta el peor desempeño de todo el estudio hasta este punto del capítulo: un error absoluto medio (MAE) de 295,1 homicidios por mes, equivalente a un error porcentual (MAPE) de 25,3%. El error medio (ME = 295,141), prácticamente idéntico al MAE y de signo positivo, revela un sesgo sistemático de subestimación, el modelo siempre pronosticó por debajo del valor real observado, un patrón opuesto al de LSTM, que sobreestimaba de forma igual de sistemática (ME negativo).

Tabla 36
Métricas de desempeño de GRU frente al valor observado real de enero-mayo de 2026.

Modelo	MAE	RMSE	MASE	ME	MAPE	R2
GRU	295,141	298,025	3,908	295,141	25,345	-50,175

Nota. Elaboración propia.

El MASE de 3,908 indica que el modelo es casi cuatro veces peor que un pronóstico ingenuo estacional, y el R^2 de -50,175 confirma que GRU se equivoca más de 50 veces respecto a lo que se equivocaría con solo pronosticar el promedio de los cinco meses de prueba. En conjunto, LSTM y GRU exhiben los dos peores desempeños del ejercicio comparativo completo, cada una fallando en una dirección opuesta (sobreestimación marcada en un caso, subestimación marcada en el otro), lo que refuerza la conclusión ya documentada en el análisis SHAP de ambos modelos: con apenas 240 observaciones supervisadas, ninguna de las dos arquitecturas recurrentes logró aprovechar de forma efectiva la memoria autorregresiva de la propia serie, apoyándose en cambio casi exclusivamente en variables exógenas cuya proyección futura (al no contar con valores reales para 2026) introduce una fuente adicional de incertidumbre.

A esta limitación de tamaño de muestra conviene añadir un segundo factor, no controlado en el diseño del estudio: las variables predictoras no fueron sometidas a

escalamiento (Sección 7.4), decisión justificada por la invarianza de los modelos basados en árboles a las transformaciones monótonas, propiedad que no se extiende a las redes recurrentes. El entrenamiento de LSTM y GRU es sensible a la escala de sus entradas, y en este conjunto conviven variables de órdenes de magnitud muy distintos (conteos mensuales de homicidios en el orden de las centenas, incautaciones en kilogramos, índices y tasas porcentuales). Con la evidencia disponible no es posible atribuir el fallo de ambas arquitecturas a una única causa; el tamaño reducido de la serie y la ausencia de normalización de las entradas actúan como explicaciones concurrentes, y su separación exigiría repetir el ejercicio con las variables escaladas.

8.4.6 Modelo Prophet

Prophet funciona de forma muy distinta a los modelos anteriores: en vez de aprender a partir de rezagos de la serie o de árboles de decisión, descompone la serie explícitamente en una tendencia de largo plazo, un patrón estacional que se repite cada año, y el efecto de las variables exógenas, de forma similar a la descomposición aditiva de la Sección 8.2. Por eso sus hiperparámetros no controlan "cuántos árboles" o "cuánta memoria", sino qué tan flexible se le permite ser a cada uno de esas dos componentes; *changepoint_prior_scale* (0,01; 0,1 o 0,5) controla qué tan libre está la tendencia para cambiar de dirección a lo largo del tiempo: un valor bajo produce una tendencia más rígida y suave, casi una línea recta; un valor alto le permite a la tendencia doblarse con más frecuencia y más marcadamente, ajustándose más de cerca a las subidas y bajadas históricas, aunque con mayor riesgo de sobreajustarse a fluctuaciones que no son realmente un cambio de tendencia.

Por otro lado, *seasonality_prior_scale* (1,0 o 10,0) cumple un papel equivalente, pero para el componente estacional: un valor bajo obliga al patrón anual que se repite cada año a ser más suave y contenido; un valor alto le permite a ese patrón estacional tener oscilaciones más marcadas entre los distintos meses del año. En conjunto, estos dos hiperparámetros son la forma que tiene Prophet de decidir "qué tanto confiar" en la tendencia de largo plazo versus en el patrón que se repite cada año, sin necesidad de que quien construye el modelo tenga que especificar manualmente esa proporción. A diferencia de los modelos anteriores, Prophet descompone la serie de forma aditiva en tendencia, estacionalidad y regresores exógenos, y entrega intervalos de predicción nativos sin necesidad de aproximaciones. Sobre 30 combinaciones evaluadas (un subconjunto informado, análogo al usado para LSTM y GRU), la búsqueda seleccionó las seis variables exógenas candidatas de forma conjunta, con una escala de flexibilidad de la tendencia (*changepoint_prior_scale*) de 0,1 y una escala de la estacionalidad (*seasonality_prior_scale*) de 10, alcanzando un RMSE de validación de 53,9 homicidios (Tabla 37).

Tabla 37
Mejores 10 combinaciones de variables exógenas e hiperparámetros identificados para Prophet.

Variables	Changep.	Season.	RMSE val.	n
HP + LP + IK + ISE + TD + VI	0,10	10,0	53,88	24
HP + LP + IK + ISE + TD + VI	0,10	1,0	54,08	24
HP + ISE + TD + VI	0,10	1,0	58,43	24
HP + ISE + TD + VI	0,10	10,0	58,72	24
LP + TD + VI	0,10	10,0	59,03	24
LP + TD + VI	0,10	1,0	59,30	24
(sin exógenas)	0,10	10,0	60,19	24
(sin exógenas)	0,10	1,0	60,31	24

Variables	Changep.	Season.	RMSE val.	n
IK + ISE + VI	0,10	10,0	61,07	24
HP + ISE + TD + VI	0,50	10,0	61,62	24

Nota. HP = Hurto a personas, ISE = ISE, TD = Tasa desempleo, VI = Violencia intrafamiliar, LP = Lesiones personales, IK = Incautación_kg. Elaboración propia.

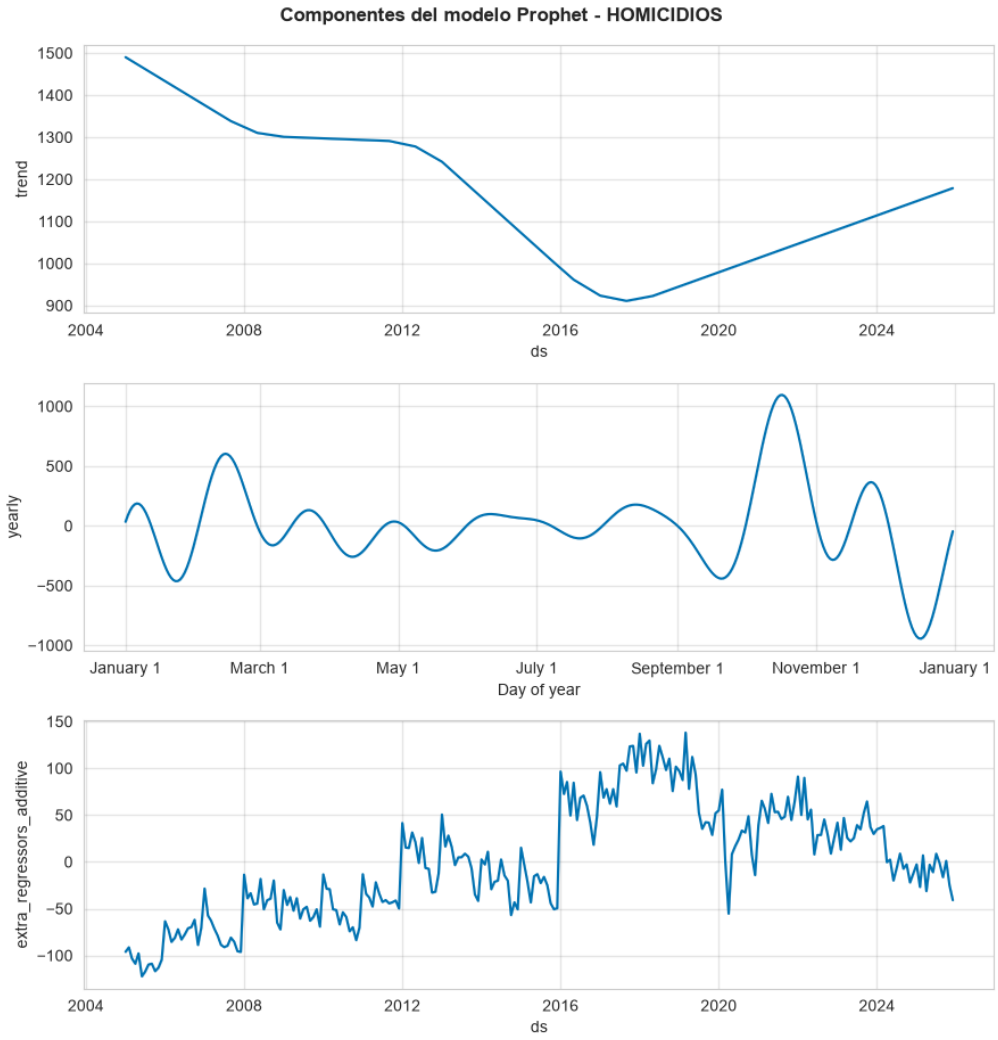
La descomposición aditiva del modelo Prophet muestra tres componentes con dinámicas muy distintas (Figura 28). La tendencia exhibe una fase de descenso sostenido entre 2005 y 2018, con un mínimo pronunciado hacia 2018-2019, seguida de un repunte constante y acelerado hasta el final del periodo estimado (2025), consistente con el aumento del nivel de homicidios ya identificado en la descomposición aditiva de la Sección 8.2. El componente estacional anual (yearly) confirma un patrón que se repite cada año, con un pico marcado alrededor de marzo y otro considerablemente más pronunciado hacia noviembre, este último de magnitud creciente en años recientes, llegando a superar los 1.000 casos por encima de la tendencia, y un valle profundo en enero, coherente con la estacionalidad ya documentada en el gráfico de subseries de la Sección 8.2.2.

El componente de regresores exógenos (*extra_regressors_additive*) (que agrupa el efecto conjunto de las seis variables exógenas incluidas en esta especificación) muestra un patrón particularmente informativo: su contribución es cercana a cero o incluso negativa hasta aproximadamente 2011, luego se estabiliza en un rango moderado (-50 a 50) entre 2012 y 2015, y a partir de 2016 se eleva de forma marcada, alcanzando su punto máximo (más de 100 casos adicionales) entre 2018 y 2020, para finalmente descender de nuevo a valores negativos hacia 2024-2025. Este cambio de régimen alrededor de 2016, coincidente, en la componente de tendencia, con el inicio del repunte sostenido, sugiere que el aporte explicativo de las variables estructurales (economía, desempleo, criminalidad conexa) no ha sido constante a lo largo del tiempo, sino que se intensificó notablemente

en la segunda mitad de la década pasada, un matiz relevante para la interpretación integral del modelo que va más allá de lo que capturan sus coeficientes fijos.

Figura 28

Descomposición aditiva del modelo Prophet: tendencia, estacionalidad anual y efecto de las variables exógenas.

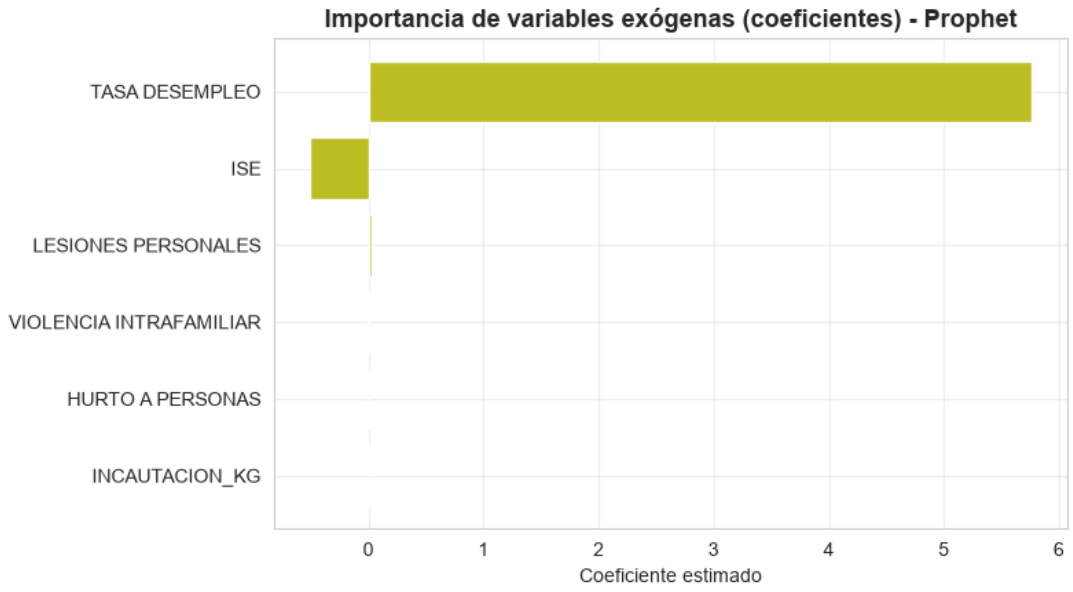


Nota. Elaboración propia.

Por su estructura aditiva y lineal en los regresores, en Prophet la interpretabilidad no se evalúa mediante valores SHAP (que exigen un modelo cuya salida dependa de interacciones no lineales entre variables) sino mediante los coeficientes estimados para

cada regresor, que cuantifican de forma directa su efecto marginal sobre el pronóstico (Figura 29 y Tabla 38).

Figura 29
Coeficientes estimados de las variables exógenas del modelo Prophet.



Nota. Elaboración propia.

Tabla 38
Coeficientes de las seis variables exógenas del modelo Prophet.

Regresor	Modo	Centro	Coef. inf.	Coeficiente	Coef. sup.
Hurto a personas	additive	13.955,46	0,001788	0,001788	0,001788
Lesiones personales	additive	7.044,74	0,027095	0,027095	0,027095
Incautación_kg	additive	40.870,40	0,000854	0,000854	0,000854
ISE	additive	96,38	-0,504869	-0,504869	-0,504869
Tasa desempleo	additive	11,05	5,758552	5,758552	5,758552
Violencia intrafamiliar	additive	5.993,55	-0,007100	-0,007100	-0,007100

Nota. Elaboración propia.

Dos coeficientes de Prophet resultan especialmente coherentes con la teoría criminológica y económica: el ISE presenta un coeficiente negativo (-0,505), es decir, una

mejor actividad económica se asocia con un menor número de homicidios pronosticado, mientras que la tasa de desempleo presenta un coeficiente positivo y de gran magnitud (5,76), consistente con la hipótesis de que el deterioro de las condiciones laborales incrementa la propensión a la violencia homicida. Los coeficientes de hurto a personas, lesiones personales e incautación de estupefacientes son positivos, pero de magnitud reducida, y el de violencia intrafamiliar es prácticamente nulo.

El pronóstico de Prophet para enero-mayo de 2026 (Tabla 39 y Figura 30) replica de forma razonable el patrón de altibajos de la serie: parte de 1.160 casos en enero, cae con fuerza a 1.031 en febrero, capturando correctamente tanto la dirección como buena parte de la magnitud de la caída característica de ese mes, repunta a 1.110 en marzo, retrocede levemente a 1.097 en abril y cierra con un salto notorio a 1.184 en mayo.

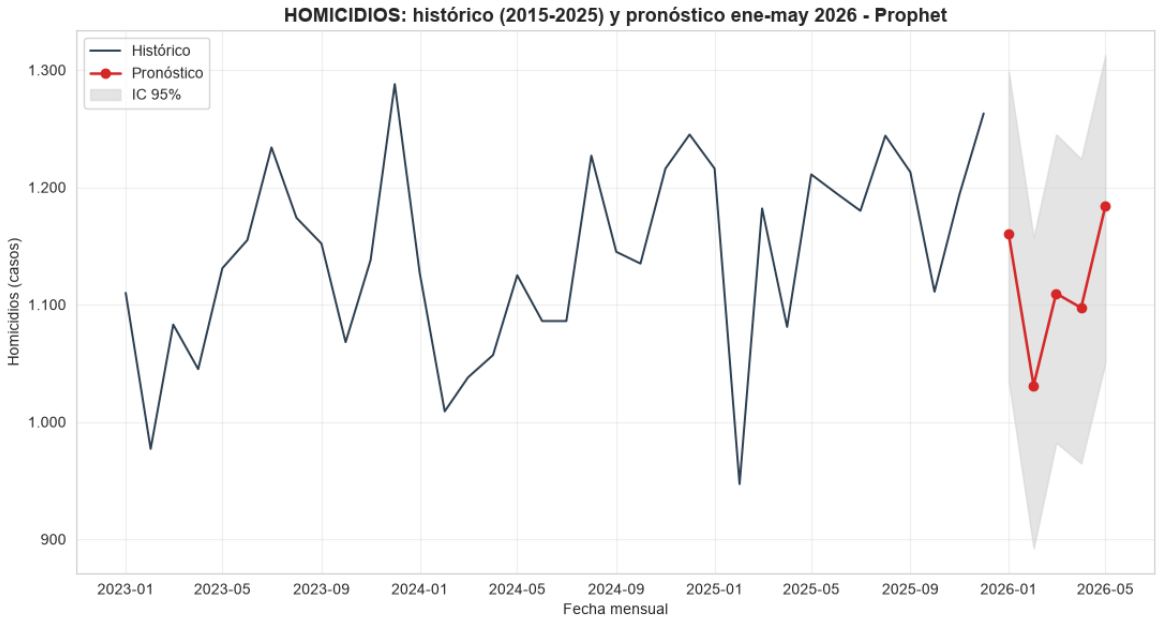
Tabla 39
Pronóstico puntual e intervalo de confianza del 95% de Prophet, enero-mayo de 2026.

Modelo	Fecha mensual	Pronóstico	Límite inf. (95%)	Límite sup. (95%)
Prophet	2026-01-01	1.160,2	1.035,0	1.298,6
Prophet	2026-02-01	1.031,1	892,2	1.157,4
Prophet	2026-03-01	1.109,5	982,3	1.245,2
Prophet	2026-04-01	1.097,3	964,7	1.224,2
Prophet	2026-05-01	1.183,7	1.050,5	1.312,6

Nota. Elaboración propia.

Figura 30

Serie histórica de homicidios (2015-2025) y pronóstico de Prophet para enero-mayo de 2026, con intervalo de confianza del 95%.



Nota. Elaboración propia.

La trayectoria general se mantiene dentro del rango histórico reciente de la serie (2023-2025), sin el desprendimiento hacia niveles irreales que mostraron LSTM y GRU, aunque con una caída de febrero algo más pronunciada de lo que resultó ser el valor real observado. El intervalo de confianza, aquí calculado de forma nativa por el propio modelo, a diferencia de los modelos de aprendizaje automático que requieren una aproximación, es notablemente amplio, llegando a un rango de casi 900 a 1.300 casos hacia el final del horizonte, reflejando la incertidumbre propia de una estructura aditiva que combina tendencia, estacionalidad y seis variables exógenas.

Las métricas de desempeño del modelo Prophet (Tabla 40) presenta un error absoluto medio (MAE) de 43,6 homicidios por mes, equivalente a un error porcentual (MAPE) de 3,79%. El error medio (ME = 43,6), idéntico al MAE, revela que el modelo subestimó de forma sistemática y sin compensación el nivel de homicidios en todo el

horizonte de prueba, una subestimación presente en los cinco meses, mientras que SARIMA subestima en cuatro de ellos, aunque de mayor magnitud. El MASE de 0,578 confirma que el modelo supera a un pronóstico ingenuo estacional, y el R^2 de -0,342, aunque negativo, se ubica en un rango similar al de SARIMA (-0,30), reflejando que, pese a su estructura interpretable y a la ventaja de contar con intervalos de confianza nativos, Prophet no logra en este ejercicio un desempeño competitivo frente a los modelos de mejor resultado del estudio (LightGBM, SARIMAX, Random Forest y VAR), aunque sí se ubica claramente por encima de XGBoost, LSTM y GRU.

Tabla 40
Métricas de desempeño de Prophet frente al valor observado real de enero-mayo de 2026.

Modelo	MAE	RMSE	MASE	ME	MAPE	R2
Prophet	43,633	48,263	0,578	43,633	3,792	-0,342

Nota. Elaboración propia.

8.4.7 Síntesis de los modelos de aprendizaje automático

La interpretabilidad de los modelos de aprendizaje automático se evaluó mediante valores SHAP (SHapley Additive exPlanations) para los cinco modelos de esta familia que lo permiten, XGBoost, LightGBM, Random Forest, LSTM y GRU, presentados en detalle junto con cada modelo en la Sección 8.4. Para Prophet, cuya estructura aditiva y lineal en los regresores no se presta al mismo tipo de explicación, se usó en su lugar el análisis de sus coeficientes estimados, también presentado en la Sección 8.4.6.

Un patrón se repite de forma consistente en los tres modelos basados en árboles de decisión (XGBoost, LightGBM y Random Forest): la variable más influyente es, en los tres casos, la violencia intrafamiliar del mes anterior, una variable exógena, seguida de cerca por el rezago de periodicidad anual de la propia serie de homicidios (lag_12). Este

orden se mantiene incluso en LightGBM, el modelo con mejor desempeño global del estudio, lo que indica que la señal exógena aporta un poder explicativo primario y no meramente secundario frente a la memoria interna de la serie. Véase Tabla 41.

Tabla 41

Síntesis de las variables más influyentes por modelo, según SHAP (modelos de árboles y redes recurrentes) o coeficientes estimados (Prophet).

Modelo	Variables destacadas	Interpretación
XGBoost	Violencia intrafamiliar (t-1), lag_12, lag_1, mes	Violencia intrafamiliar del mes anterior como variable dominante, seguida de la estacionalidad anual (lag_12) y de la memoria de corto plazo (lag_1); lesiones personales y tasa de desempleo aportan señal secundaria
LightGBM	Violencia intrafamiliar (t-1), lag_12, mes, hurto a personas (t-1)	Mismo patrón: violencia intrafamiliar del mes anterior en primer lugar, seguida del rezago anual y del mes calendario; hurto a personas, ISE y tasa de desempleo aportan señal secundaria
Random Forest	Violencia intrafamiliar (t-1), lag_12, ISE (t-1)	Mismo patrón: violencia intrafamiliar del mes anterior en primer lugar, seguida del rezago anual; incautación de estupefacientes e ISE aportan señal secundaria
LSTM / GRU	Variables exógenas (ISE, violencia intrafamiliar, hurto, lesiones) y mes; rezagos de la serie con importancia marginal	Señal concentrada en las exógenas; la memoria autorregresiva explícita no aporta contribución individual apreciable, en contraste con los modelos de árboles
Prophet	ISE, Tasa de desempleo (coeficientes)	ISE con efecto negativo y tasa de desempleo con efecto positivo sobre el homicidio pronosticado

Nota. Elaboración propia.

Este patrón, con la violencia intrafamiliar del mes anterior por encima de los rezagos de la propia serie, matiza lo observado en la descomposición de las secciones 8.2 y 8.3.1: la estacionalidad anual y la memoria de corto plazo siguen siendo determinantes, y el rezago de doce meses es la segunda variable en importancia en los tres modelos de árboles, pero la conflictividad asociada a otros delitos no actúa como un simple ajuste adicional, sino como la señal individual de mayor peso predictivo.

8.5 Verificación de hipótesis y propuestas sobre resultados

Al consolidar y comparar los nueve modelos entrenados, tres estadísticos clásicos y seis de aprendizaje automático, contra el homicidio efectivamente observado entre enero y mayo de 2026 (Tabla 42), se evidenció lo siguiente: Primero, cuatro modelos logran un ajuste razonable fuera de muestra, con coeficiente de determinación positivo: LightGBM ($R^2 = 0,84$) como ganador destacado y por un margen amplio, seguido de SARIMAX ($R^2 = 0,37$), Random Forest ($R^2 = 0,26$) y VAR HOMICIDIOS~ISE ($R^2 = 0,20$) (es decir, tanto modelos de árboles con variables de rezago como los modelos estadísticos multivariados (SARIMAX y VAR) logran superar a la media del periodo de prueba, que es la referencia del R^2). Segundo, SARIMA y Prophet logran capturar el orden de magnitud del fenómeno, pero con un desempeño por debajo de esa misma referencia ($R^2 = -0,30$ y $-0,34$, respectivamente). Tercero, XGBoost y, sobre todo, las dos redes neuronales recurrentes fallan de forma severa: XGBoost con un R^2 de -0,76, LSTM proyectando una trayectoria creciente que se aleja cada vez más del nivel real observado ($R^2 = -33,25$), y GRU, pese a estabilizarse en un nivel plano en vez de una trayectoria creciente, registrando el peor desempeño de todo el estudio ($R^2 = -50,17$).

Tabla 42
Métricas de desempeño de los nueve modelos, ordenadas de mejor a peor por RMSE.

Modelo	MAE	RMSE	MASE	ME	MAPE	R2
LightGBM	14,62	16,61	0,19	1,35	1,25	0,84
SARIMAX(2, 0, 1)(0, 1, 2, 12)	31,55	33,16	0,42	15,79	2,75	0,37
Random Forest	31,30	35,84	0,41	28,88	2,66	0,26
VAR HOMICIDIOS~ISE	32,81	37,36	0,43	21,22	2,81	0,20
SARIMA(2, 0, 1)(0, 1, 2, 12)	41,04	47,41	0,54	37,33	3,60	-0,30
Prophet	43,63	48,26	0,58	43,63	3,79	-0,34
XGBoost	53,28	55,26	0,71	53,28	4,56	-0,76
LSTM	239,92	243,82	3,18	-239,92	20,83	-33,25

Modelo	MAE	RMSE	MASE	ME	MAPE	R2
GRU	295,14	298,03	3,91	295,14	25,34	-50,17

Nota. Elaboración propia. ME: error medio, definido como valor observado menos pronosticado (un valor positivo indica subestimación). para SARIMA, SARIMAX, VAR y Prophet el intervalo es el nativo del modelo, y para los demás es la aproximación descrita en la Sección 7.5

El objetivo central de este trabajo fue comparar el desempeño de modelos estadísticos clásicos aprendizaje automático en el pronóstico mensual del homicidio intencional en Colombia. Los resultados solo respaldan de forma parcial la hipótesis de que los modelos de aprendizaje automático superan de manera uniforme a los modelos estadísticos: esto es cierto para LightGBM (R^2 = 0,84) que efectivamente supera a los tres modelos estadísticos, parcialmente cierto para Random Forest (R^2 = 0,26) que supera dos modelos estadísticos, pero es falso para Prophet (R^2 = -0,34), XGBoost (R^2 = -0,76), LSTM (R^2 = -33,25) y, en particular, GRU (R^2 = -50,17), los cuales quedan por debajo incluso del más débil de los modelos estadísticos evaluados (SARIMA, R^2 = -0,30). La evidencia respalda, en cambio, una hipótesis más matizada: la familia de algoritmo importa más que la etiqueta genérica "Machine Learning" o "modelo estadístico"; específicamente, los métodos de ensamble de árboles con variables de rezago bien construidas (LightGBM, Random Forest) y los modelos dinámicos multivariados bien especificados (VAR, SARIMAX) son los que efectivamente aportan valor sobre el pronóstico de este fenómeno, mientras que las arquitecturas de mayor complejidad nominal (redes neuronales profundas) no lo hacen en este escenario de datos limitados.

La Tabla 43 y la Figura 31 consolidan visualmente lo que ya se documentó en las métricas individuales: los nueve modelos se agrupan en tres bandas claramente diferenciadas frente al valor observado real de 2026. La primera banda, LightGBM, SARIMAX, Random Forest, VAR, SARIMA, Prophet y XGBoost, se mantiene dentro o muy

cerca del rango histórico reciente (aproximadamente 1.050 a 1.250 casos), siguiendo de cerca la trayectoria en "V" de la serie observada (la caída pronunciada hacia febrero y la recuperación posterior). La segunda y tercera banda corresponden a las dos redes recurrentes, que se desprenden por completo de esa zona: LSTM se dispara muy por encima de todos los demás modelos y de la propia serie histórica, con una trayectoria creciente que no guarda ninguna relación con el patrón real, mientras que GRU se ubica muy por debajo, en un nivel casi plano y considerablemente inferior a cualquier valor observado en los últimos dos años.

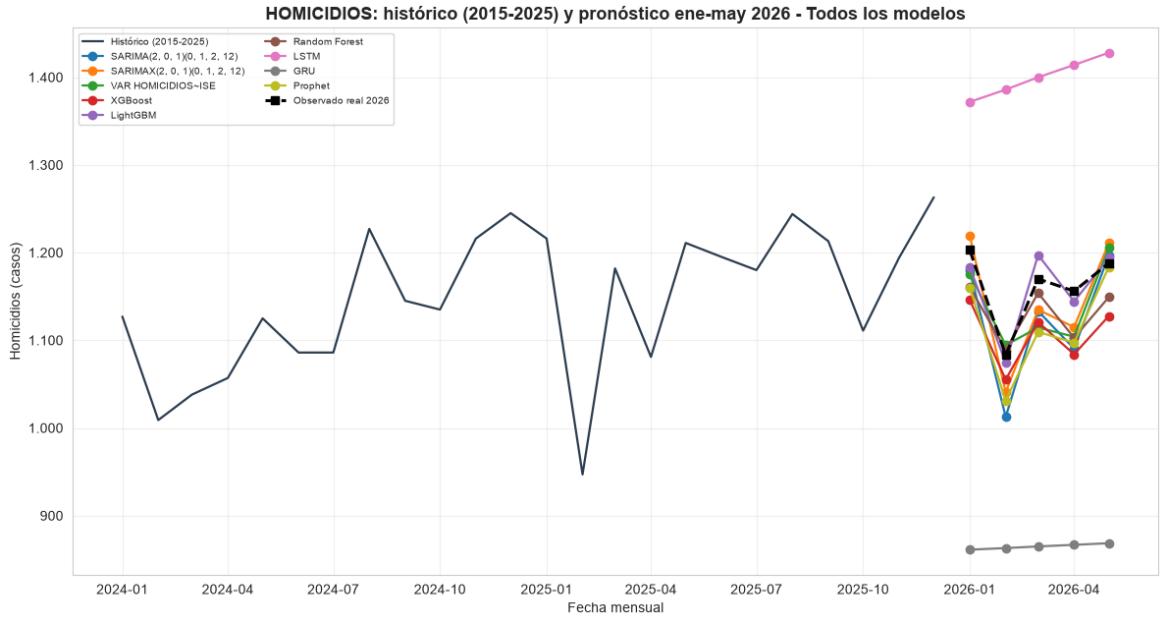
Tabla 43
Pronósticos de enero a mayo de 2026 de los nueve modelos evaluados.

Modelo	Pronóstico 2026-01	Pronóstico 2026-02	Pronóstico 2026-03	Pronóstico 2026-04	Pronóstico 2026-05
LightGBM	1.183,20	1.074,60	1.197,30	1.144,30	1.193,90
SARIMAX(2, 0, 1)(0, 1, 2, 12)	1.219,30	1.040,80	1.134,90	1.114,90	1.211,10
Random Forest	1.160,70	1.089,10	1.153,90	1.102,80	1.149,20
VAR HOMICIDIOS~ISE	1.175,30	1.094,40	1.114,50	1.104,10	1.205,60
SARIMA(2, 0, 1)(0, 1, 2, 12)	1.179,90	1.012,50	1.133,20	1.090,50	1.197,30
Prophet	1.160,20	1.031,10	1.109,50	1.097,30	1.183,70
XGBoost	1.146,00	1.055,20	1.120,70	1.084,00	1.127,70
LSTM	1.371,80	1.385,90	1.399,90	1.414,00	1.428,00
GRU	861,20	863,00	864,90	866,70	868,60
Observado real 2026	1.203,00	1.083,00	1.170,00	1.156,00	1.188,00

Nota. Elaboración propia.

Figura 31

Comparación del pronóstico de enero-mayo de 2026 de los nueve modelos frente al valor realmente observado.



Nota. Elaboración propia.

Dentro de la primera banda, donde se concentran los modelos con desempeño razonable, el gráfico permite apreciar que ninguno reproduce exactamente la forma de la "V" de febrero, pero LightGBM (línea morada) es visualmente el que más se aproxima al observado real (línea negra punteada) en los cinco meses, seguido de cerca por VAR y SARIMAX en marzo y abril. SARIMA (línea azul) es el que se aleja más del valor real en el punto más bajo de febrero, subestimándolo con fuerza (1.012 frente a 1.083 real), consistente con el predominio de errores de subestimación ya documentado en su tabla de errores. En conjunto, ambas imágenes refuerzan visualmente la conclusión central del capítulo: la diferencia de desempeño entre modelos no es sutil ni marginal, sino que separa con claridad a un grupo de modelos "razonables", así tengan distintos niveles de precisión entre sí, de dos modelos con fallas estructurales severas (LSTM y GRU), cuya inclusión

en un sistema real de pronóstico generaría alarmas sistemáticamente equivocadas en una u otra dirección.

Al comparar la consistencia del error entre los nueve modelos (Tabla 44), la evidencia es clara: los modelos estadísticos son, en conjunto, más consistentes que aprendizaje automático, aunque por un motivo específico que conviene precisar. De los tres modelos estadísticos, SARIMAX alterna de signo (sobreestima en enero y mayo, subestima en el resto) y VAR corrige su propio sesgo en dos de cinco meses, de modo que ninguno de los dos acumula error sistemáticamente en una sola dirección a lo largo de todo el horizonte; solo SARIMA se acerca a un patrón unidireccional (subestima en cuatro de cinco meses), pero incluso así con una magnitud moderada (entre -9,3 y +70,5). En contraste, dentro de los modelos de aprendizaje automático aparecen los dos casos más extremos de todo el estudio: XGBoost, LSTM y GRU sostienen el mismo signo de error en los cinco meses sin ninguna excepción, XGBoost y GRU subestimando siempre, LSTM sobreestimando siempre, y en el caso de LSTM y GRU esa consistencia de signo viene acompañada de una magnitud entre 6 y 10 veces mayor que la de cualquier modelo estadístico (hasta +341,8 en GRU, frente a un máximo de +70,5 en SARIMA). Solo LightGBM y, en menor medida, Random Forest logran un comportamiento comparable, o mejor, al de los modelos estadísticos: LightGBM es, de los nueve, el único con signos genuinamente alternados y magnitudes pequeñas (entre -27,3 y +19,8), superando en este aspecto incluso a VAR y SARIMAX.

Tabla 44
Error de pronóstico mensual (real menos predicho) de los nueve modelos frente al valor observado, enero-mayo de 2026.

Modelo	Error 2026-01	Error 2026-02	Error 2026-03	Error 2026-04	Error 2026-05
LightGBM	+19,8	+8,4	-27,3	+11,7	-5,9

Modelo	Error 2026-01	Error 2026-02	Error 2026-03	Error 2026-04	Error 2026-05
SARIMAX(2, 0, 1)(0, 1, 2, 12)	-16,3	+42,2	+35,1	+41,1	-23,1
Random Forest	+42,3	-6,1	+16,1	+53,2	+38,8
VAR HOMICIDIOS~ISE	+27,7	-11,4	+55,5	+51,9	-17,6
SARIMA(2, 0, 1)(0, 1, 2, 12)	+23,1	+70,5	+36,8	+65,5	-9,3
Prophet	+42,8	+51,9	+60,5	+58,7	+4,3
XGBoost	+57,0	+27,8	+49,3	+72,0	+60,3
LSTM	-168,8	-302,9	-239,9	-258,0	-240,0
GRU	+341,8	+220,0	+305,1	+289,3	+319,4

Nota. Elaboración propia

En síntesis: la consistencia no está dividida limpiamente por familia de modelo, sino que dentro de cada familia coexisten comportamientos muy distintos. Los modelos estadísticos, como grupo, muestran una consistencia moderada y relativamente homogénea entre sí (SARIMA, SARIMAX y VAR se mueven todos en un rango de magnitud similar); los modelos de aprendizaje automático, en cambio, presentan la mayor dispersión de comportamiento de todo el estudio, con LightGBM en un extremo (el más consistente y de menor magnitud de los nueve modelos) y LSTM-GRU en el otro (los únicos con sesgo perfectamente unidireccional y errores de magnitud extrema), lo que refuerza la conclusión que se retoma en el Capítulo 9: la familia algorítmica por sí sola no determina la calidad ni la consistencia del pronóstico, sino la especificación particular de cada modelo frente al tamaño de muestra disponible.

Considerando que el conjunto de prueba tiene cinco (5) observaciones, se contrastó formalmente si la diferencia de precisión entre LightGBM y cada uno de los demás modelos es estadísticamente distinguible de cero mediante la prueba de Diebold y Mariano (1995) con pérdida cuadrática y horizonte uno, en su versión asintótica y con la

corrección de muestra pequeña de Harvey, Leybourne y Newbold (1997), que contrasta el estadístico corregido frente a una distribución t de Student con cuatro grados de libertad (Tabla 45). Un estadístico negativo indica menor pérdida de LightGBM.

Tabla 45
Prueba de Diebold-Mariano de LightGBM frente a los demás modelos, enero-mayo de 2026.

Comparación	DM asintótico	p (normal)	DM corregido (HLN)	p (t, 4 g.l.)
LightGBM vs SARIMAX	-2,64	0,008	-2,36	0,078
LightGBM vs Random Forest	-1,97	0,048	-1,77	0,152
LightGBM vs VAR	-2,30	0,021	-2,06	0,109
LightGBM vs SARIMA	-2,09	0,037	-1,87	0,135
LightGBM vs Prophet	-3,80	<0,001	-3,40	0,027
LightGBM vs XGBoost	-4,13	<0,001	-3,69	0,021
LightGBM vs LSTM	-6,42	<0,001	-5,74	0,005
LightGBM vs GRU	-8,66	<0,001	-7,74	0,001

Nota. Calculada a partir de los errores mensuales de la Tabla 44. DM = media de la diferencia de pérdidas cuadráticas dividida por su error estándar; la corrección HLN multiplica el estadístico por la raíz de (n + 1 - 2h + h(h - 1)/n)/n, con n = 5 y h = 1. Los valores p asintóticos se reportan solo con fines comparativos, pues la aproximación normal no es válida con cinco observaciones. Elaboración propia.

Los resultados permiten dos lecturas. La primera es que la ventaja de LightGBM sobre SARIMAX, Random Forest, VAR y SARIMA es de magnitud amplia (la razón entre el error cuadrático medio de SARIMAX y el de LightGBM es de 4,0) pero no alcanza significancia estadística al 5 % con la corrección de muestra pequeña; frente a SARIMAX, LightGBM obtuvo menor error absoluto en cuatro de los cinco meses (prueba de signos, p = 0,19; prueba de Wilcoxon, p = 0,06). El ordenamiento dentro del grupo de modelos con desempeño razonable debe entenderse, por tanto, como descriptivo y propio de la ventana evaluada. La segunda lectura es que la separación entre ese grupo y los modelos de menor desempeño (Prophet, XGBoost, LSTM y GRU) sí es estadísticamente defendible incluso

con cinco observaciones, lo que sustenta la conclusión de que la familia algorítmica, y no la etiqueta genérica de aprendizaje automático, determina la calidad del pronóstico. Una comparación concluyente entre los modelos del primer grupo requiere un conjunto de prueba más largo, lo que motiva el esquema de validación de origen móvil propuesto en la Sección 10.

La elección de un modelo para un escenario de política pública no debería basarse únicamente en la métrica de error más baja. LightGBM logró el mejor desempeño numérico, pero es, por diseño, una "caja negra" relativamente opaca: su valor práctico depende de que el análisis SHAP se mantenga y actualice para sostener la interpretabilidad ante tomadores de decisión. El modelo VAR, en cambio, ofrece una ventaja distinta y complementaria: además de un desempeño competitivo, permite cuantificar de forma explícita y con significancia estadística la relación dinámica bidireccional entre homicidios y actividad económica (Sección 8.3.3), lo que lo hace más adecuado cuando el objetivo no es solo pronosticar sino también entender y comunicar mecanismos. Prophet, pese a no ser el mejor en desempeño, es el modelo más transparente de todos: su descomposición aditiva en tendencia, estacionalidad y regresores (Sección 8.4.6) es directamente interpretable incluso para audiencias no técnicas, lo que lo hace razonable para reportes periódicos de seguimiento, aunque no para el pronóstico de mayor precisión. En el otro extremo, este ejercicio sugiere que las redes neuronales recurrentes (LSTM, GRU) no son adecuadas para este escenario concreto, una serie mensual de poco más de una década de longitud, y que su uso debería reservarse para contextos con series más largas, de mayor frecuencia (semanal o diaria) o con un volumen de datos sustancialmente mayor.

9. Discusión

9.1 Contraste de hipótesis

La comparación de los nueve modelos frente al homicidio efectivamente observado entre enero y mayo de 2026 permite examinar cada una de las hipótesis planteadas en la Sección 5, así como discutir sus implicaciones metodológicas y prácticas. La Hipótesis 1 planteó que los modelos de aprendizaje automático, por su mayor flexibilidad, ofrecerían un desempeño superior al de los modelos estadísticos clásicos. Los resultados solo respaldan esta hipótesis de forma parcial: uno de los seis modelos de aprendizaje automático, LightGBM (R^2 = 0,84) supera efectivamente a los tres modelos estadísticos, Random Forest (R^2 = 0,26) supera dos modelos estadísticos (VAR y SARIMA) mientras que los otros cuatro, Prophet (R^2 = -0,34), XGBoost (R^2 = -0,76), LSTM (R^2 = -33,25) y GRU (R^2 = -50,17), obtienen un desempeño considerablemente peor que el mejor modelo estadístico (SARIMAX, R^2 = 0,37). La Hipótesis 1 no puede, entonces, aceptarse ni rechazarse de forma categórica: la evidencia indica que la superioridad del aprendizaje automático depende del algoritmo específico y no constituye una propiedad uniforme de la familia.

Este resultado plantea una pregunta relevante: si XGBoost, LightGBM y Random Forest son los tres modelos basados en árboles de decisión, ¿por qué solo dos de ellos generalizan razonablemente al periodo de prueba? La evidencia disponible apunta a dos factores. El primero es la complejidad de la configuración seleccionada en validación: la de XGBoost es la más flexible de las tres (profundidad máxima 5, 400 árboles, tasa de aprendizaje 0,1), mientras que la de LightGBM es la más contenida (profundidad 3, 200 árboles, tasa 0,05 y un máximo de 15 hojas), y Random Forest, por su mecanismo de agregación de árboles independientes, reduce la varianza por construcción. Una

configuración más flexible es más propensa a ajustarse a regularidades del histórico que no se repiten en el periodo pronosticado. El segundo factor es la variabilidad propia de una ventana de prueba de cinco meses: en el periodo de validación de 24 meses, XGBoost obtuvo el menor RMSE de los tres modelos de árboles (51,5 frente a 60,4 de LightGBM y 66,8 de Random Forest; Tablas 22, 25 y 28), y el orden se invirtió en el periodo de prueba. Aunque ambas evaluaciones no son estrictamente comparables (la de validación es a un paso con rezagos observados y la de prueba es recursiva a cinco pasos con exógenas por persistencia), esta inversión indica que el ordenamiento entre modelos de desempeño razonable es sensible a la ventana evaluada, en coherencia con la prueba de Diebold-Mariano de la Tabla 45. En síntesis, la pertenencia a una misma familia algorítmica no garantiza un comportamiento similar frente a datos no vistos; lo determinante es la interacción entre la complejidad de la configuración seleccionada, el tamaño de la muestra y la longitud del periodo de prueba.

La Hipótesis 2 planteó que las variables exógenas de naturaleza socioeconómica y delictiva aportarían capacidad predictiva adicional más allá de la historia propia de la serie. La evidencia la confirma de forma parcial, y conviene precisar en qué sentido. En el sentido estrictamente predictivo, que es el que enuncia la hipótesis, la evidencia directa proviene de los modelos estadísticos: SARIMAX, que incorpora lesiones personales y tasa de desempleo, supera a SARIMA en las seis métricas de prueba (Tabla 21), y VAR incorpora el Índice de Seguimiento a la Economía con una relación de causalidad en el sentido de Granger estadísticamente significativa (Tabla 16), superando también al modelo univariado. En el sentido de la importancia de las variables, el análisis SHAP de la Sección 8.4 aporta un hallazgo que no se anticipaba: la violencia intrafamiliar del mes anterior es la variable individual más influyente en los tres modelos basados en árboles, incluido LightGBM, el modelo con mejor desempeño global, por encima incluso del rezago anual

de la propia serie. Esta evidencia encuentra una lectura plausible en la teoría de la desorganización social (Sampson y Groves, 1989), presentada en la Sección 4.1: los factores de debilitamiento del control social y comunitario que elevan la violencia intrafamiliar operan también como determinantes de la violencia letal en el espacio público, de modo que la primera actuaría como un indicador adelantado del deterioro del tejido social, y no como una causa directa del homicidio.

Tres límites impiden considerar la hipótesis confirmada en su totalidad. Primero, la importancia SHAP cuantifica cuánto utiliza un modelo una variable, no cuánto mejora su pronóstico fuera de muestra por incluirla; esa ganancia se documentó de manera directa solo para los modelos estadísticos. Segundo, el hallazgo no es uniforme entre familias: la violencia intrafamiliar no resultó causal en el sentido de Granger dentro del sistema VAR ($p = 0{,}18$) y recibe un coeficiente prácticamente nulo en Prophet, de modo que su protagonismo es específico de los modelos de árboles. Tercero, la relación estimada es de signo negativo, no admite lectura causal y, como toda ganancia atribuible a las exógenas en este trabajo, está condicionada al supuesto de persistencia empleado para proyectarlas. En consecuencia, la Hipótesis 2 se considera confirmada parcialmente: en su sentido predictivo para los modelos estadísticos multivariados y en su sentido de importancia para los modelos de árboles, sin que ello constituya evidencia de causalidad ni se generalice a todas las familias evaluadas.

La Hipótesis 3, en su formulación ajustada al diseño de ventana única (Sección 5), planteó que el error de los modelos estadísticos con estructura estacional explícita crecería con el horizonte de manera más moderada que el de los modelos de aprendizaje automático, y que estos últimos serían relativamente más competitivos en el horizonte más corto. El error absoluto por horizonte de la Tabla 44 no confirma ninguna de las dos partes. En el horizonte de un mes (enero de 2026) el menor error lo obtiene SARIMAX (16,3),

seguido de LightGBM (19,8) y SARIMA (23,1), de modo que los modelos de aprendizaje automático no muestran la ventaja esperada en el corto plazo. Y en ninguno de los tres modelos estadísticos el error crece de forma sostenida conforme el horizonte se aleja del último dato observado: SARIMA registra su error más bajo precisamente en mayo, el mes más lejano (9,3, frente a 70,5 en febrero); SARIMAX presenta en mayo uno de sus errores más bajos (23,1); y VAR tiene su error mínimo en febrero (11,4) y uno de los más altos en marzo (55,5). En los tres, el error máximo se concentra en un punto intermedio del horizonte, coincidiendo con el tramo en que la serie observada cae y se recupera en forma de V. LightGBM, por su parte, es el único modelo que se ubica entre los dos de menor error en los cinco horizontes. Debe advertirse, sin embargo, que con un único origen de pronóstico se dispone de una sola observación por horizonte, de modo que esta evaluación es descriptiva; la comprobación rigurosa de la hipótesis, en particular en su formulación original de dos horizontes, requiere el esquema de validación de origen móvil planteado en la Sección 10.

Más allá de las métricas, estos resultados tienen una lectura práctica directa para un tomador de decisiones sin formación técnica. El modelo con el menor error, LightGBM, es al mismo tiempo el más difícil de auditar: su lógica interna es una "caja negra" que solo resulta interpretable a través de un análisis SHAP que debe mantenerse y actualizarse junto con el modelo. El modelo VAR, aunque menos preciso, ofrece una ventaja complementaria: permite cuantificar de forma explícita y con significancia estadística la relación entre homicidio y actividad económica, lo cual resulta valioso cuando el objetivo no es solo pronosticar sino también explicar y comunicar un mecanismo a audiencias no técnicas. Prophet, pese a no ser el más preciso, es el modelo más transparente de todos gracias a su descomposición aditiva en tendencia, estacionalidad y regresores externos, lo que lo hace razonable para reportes periódicos de seguimiento dirigidos a públicos no

especializados. En el extremo opuesto, este ejercicio sugiere que las redes neuronales recurrentes (LSTM y GRU) no resultaron adecuadas en la configuración evaluada: una sola capa recurrente de 16 o 32 unidades, entradas sin escalar, 60 épocas fijas sin parada temprana y sin regularización distinta del dropout, cuyo incremento, dentro de la malla evaluada, no mejoró el error de validación. Con la evidencia disponible no es posible separar el efecto del tamaño reducido de la serie del de estas decisiones de configuración, que actúan como explicaciones concurrentes (Sección 8.4.5); por ello, la conclusión se limita a este escenario y a esta configuración, y la exploración de variantes con escalamiento de entradas, parada temprana y regularización L2 se plantea como trabajo futuro antes de descartar la familia recurrente para series mensuales de esta longitud.

Respecto a la estacionalidad, el mes calendario se ubica entre las variables más influyentes en los modelos de árboles, aunque con un aporte menor al de la violencia intrafamiliar y el rezago anual. El patrón más claro identificado en el conjunto de prueba es una caída marcada del homicidio en febrero de 2026, seguida de una recuperación hacia marzo (Sección 8.5), consistente con el patrón estacional ya documentado en la descomposición de la serie (Sección 8.2). Una interpretación criminológica más fina de este efecto, por ejemplo su posible asociación con ciclos económicos de inicio de año, excede la evidencia disponible en este estudio y se sugiere como línea de indagación futura.

En conjunto, las tres contribuciones prometidas en la introducción se sostienen con lo encontrado. La comparación sistemática entre modelos estadísticos y de aprendizaje automático se realizó bajo un esquema de validación común y frente a datos reales no utilizados en el entrenamiento, lo que permite una lectura directa de las fortalezas y debilidades relativas de cada familia. El pipeline construido controla explícitamente la fuga de información mediante el rezago de un mes en las variables exógenas y la separación

estricta del conjunto de prueba. Y el análisis SHAP aporta evidencia concreta, y en cierto sentido inesperada, sobre las variables más influyentes en el pronóstico. La brecha identificada en la introducción, la ausencia de un estudio colombiano que compare de forma sistemática y con control de fuga de información modelos estadísticos y de aprendizaje automático sobre una serie nacional de homicidios, queda cerrada de forma parcial: el estudio se circunscribe a la escala nacional agregada y no aborda la heterogeneidad territorial del fenómeno, limitación reconocida desde la Justificación y que se retoma como línea de trabajo futuro.

9.2 Recomendaciones metodológicas para el pronóstico del homicidio intencional en Colombia

De la evidencia anterior se derivan tres recomendaciones metodológicas para futuros ejercicios de pronóstico del homicidio intencional en Colombia. La primera es que la elección del modelo debe subordinarse al propósito del ejercicio y no a la sofisticación del algoritmo. Para la precisión del pronóstico mensual dentro de un horizonte del orden del evaluado en este estudio, se recomienda LightGBM (RMSE = 16,61; R^2 = 0,84). Para establecer con formalidad estadística la relación entre el homicidio y un determinante específico, se recomienda VAR. Para comunicar a públicos no especializados el origen del comportamiento de la serie, Prophet ofrece la descomposición más transparente (Figura 28 y Tabla 38), si bien su R^2 de -0,34 desaconseja emplearlo como generador de la cifra de pronóstico. En todos los casos conviene conservar SARIMAX como línea base, por su bajo costo de implementación y por un error absoluto medio de 31,6 homicidios mensuales equivalente a un MAPE de 2,7 % (Sección 8.3.4), criterio más informativo que su R^2 de 0,37, calculado sobre solo cinco observaciones.

La segunda recomendación es evitar las redes neuronales recurrentes (LSTM y GRU) en escenarios con las características de datos de este estudio, esto es, series

mensuales de poco más de una década y 132 observaciones de entrenamiento y validación; su uso debería reservarse para series más largas o de mayor frecuencia de reporte. Por una razón afín, la interacción entre el mecanismo de regularización de cada algoritmo y el tamaño de la muestra, se recomienda cautela con el boosting poco regularizado, como ilustra el caso de XGBoost. Deben privilegiarse mecanismos explícitos de control de la varianza, como el bagging de Random Forest o la regularización de LightGBM, y evaluarse siempre frente a una ventana prospectiva real y no únicamente frente a una partición de retención histórica (holdout).

La tercera recomendación se refiere a la interpretabilidad y a la selección de variables. Todo modelo de caja negra que se lleve a operación debe acompañarse de un componente de interpretabilidad, como el análisis SHAP empleado en este trabajo, actualizado con la misma periodicidad que el modelo; sin él, el pronóstico no es auditable ni defendible ante un tomador de decisiones. En cuanto a las variables, se recomienda mantener la violencia intrafamiliar del mes anterior como variable de seguimiento prioritaria en modelos no lineales, donde resultó la más influyente por encima incluso del rezago anual de la propia serie (Sampson y Groves, 1989), y someterla a pruebas de causalidad específicas antes de atribuirle un papel explicativo en modelos lineales, pues evaluada de forma individual mediante la prueba de Granger no resultó significativa ($p = 0{,}1786$; Tabla 16). La diferencia no invalida el hallazgo, sino que delimita su alcance: Granger examina la precedencia lineal dentro de un sistema autorregresivo, mientras que SHAP cuantifica la contribución de la variable dentro de modelos no lineales capaces de capturar interacciones. Junto a ella conviene conservar el rezago anual de la propia serie y el mes calendario (Sección 8.4), así como el Índice de Seguimiento a la Economía (Sección 8.3.3), y toda variable exógena debe incorporarse con un rezago mínimo de un mes para preservar el control de fuga de información temporal. En conjunto, estas recomendaciones

refuerzan la conclusión central del estudio: la competitividad de un modelo no deriva de su pertenencia a una familia, sino de la interacción entre el mecanismo del algoritmo, el tamaño de la muestra disponible y el propósito para el cual se construye el pronóstico.

9.3 Consideraciones éticas y de uso responsable del modelo

Un modelo de pronóstico de homicidio intencional, construido con la intención declarada desde la justificación de aportar evidencia cuantitativa a la formulación de política pública en seguridad, no puede cerrarse sin una reflexión sobre las condiciones en que ese pronóstico debería emplearse. Esta sección recoge cuatro consideraciones que se desprenden directamente de los resultados obtenidos en los capítulos anteriores: la manera en que debe leerse la relación entre una variable estadísticamente influyente y un mecanismo social real, las precauciones que exigiría cualquier desagregación territorial futura, la sensibilidad del pronóstico a cambios regulatorios que alteran las condiciones bajo las cuales fue entrenado, y las condiciones mínimas de gobernanza bajo las cuales un pronóstico de esta naturaleza podría incorporarse a la toma de decisiones.

Antes de desarrollarlas, conviene precisar el alcance de esta reflexión: no se ocupa de la protección de datos personales, un aspecto que en este trabajo presenta baja exposición desde su origen mismo. La fuente utilizada, SIEDCO, describe cada evento mediante dieciocho campos de naturaleza temática, temporal, geográfica y demográfica (Sección 7.2), entre los que no figura ningún identificador de la víctima ni del presunto responsable, de modo que la ausencia de datos identificables no depende únicamente de la agregación mensual a escala nacional que este trabajo realiza sobre esos registros, sino que ya es una característica de la base administrativa en su forma original. Ninguna etapa del proceso, desde la fuente hasta el modelo final, requiere entonces manipular registros individuales identificables, condición que mantiene este trabajo por fuera del régimen general de protección de datos personales establecido en la Ley 1581 de 2012. Lo que

sigue se concentra, en cambio, en la interpretación y en el uso del pronóstico una vez este existe.

Una primera consideración, ya anunciada de forma puntual al presentar el análisis SHAP en la Sección 8.4, conviene formularla aquí como principio general de lectura. Que un modelo identifique una variable como estadísticamente influyente, incluso mediante una prueba formal de causalidad como la de Granger o una medida de atribución como los valores SHAP, no equivale a confirmar que exista detrás un mecanismo causal en el sentido social del término, ni habilita por sí sola una decisión de intervención sobre esa variable. Rudin (2019) señala que los modelos de caja negra empleados en decisiones de alto impacto exigen una prudencia interpretativa reforzada, y Kleinberg et al. (2015) formalizan la distinción entre los problemas de política pública que requieren inferencia predictiva y aquellos que requieren inferencia causal, mostrando que un modelo capaz de pronosticar bien un fenómeno no resuelve, con eso, la pregunta de si actuar sobre alguna de sus variables modificaría el resultado que se busca prevenir. El caso de la violencia intrafamiliar, identificada de forma consistente como la variable individual más influyente en los tres modelos de árboles, ilustra bien esta prudencia, pues la relación negativa y contraintuitiva encontrada por el modelo puede reflejar una dinámica estadística real de la serie o puede ser una asociación espuria derivada del tamaño reducido de la muestra, y esa distinción no puede resolverse con más modelamiento sobre los mismos datos administrativos. Corroborar si el patrón detectado corresponde a un mecanismo social real exige un tipo de evidencia distinto, de naturaleza cualitativa e institucional, capaz de triangular el hallazgo estadístico con otras fuentes de conocimiento sobre el fenómeno (Small, 2011), antes de que ese hallazgo se traduzca en una narrativa pública o en un criterio de intervención.

Una segunda consideración, pertinente sobre todo si el trabajo futuro llega a materializar la desagregación territorial propuesta en la Sección 10, tiene que ver con la manera en que un patrón detectado por el modelo puede terminar asociándose a un municipio o una región de forma prematura. La literatura sobre policía predictiva documenta con suficiente claridad el mecanismo que conviene prevenir aquí: cuando el pronóstico se apoya en registros administrativos que dependen de la denuncia y el registro policial efectivos, como ocurre con SIEDCO según se señaló en la Sección 7.2, el modelo puede terminar reproduciendo los patrones de vigilancia ya existentes en lugar de medir el fenómeno de forma independiente de ellos (Lum e Isaac, 2016), y esa reproducción puede reforzarse a sí misma una vez el pronóstico empieza a orientar el propio despliegue de recursos que después alimenta al modelo con nuevos datos (Ensign et al., 2018). A esto se suma una precaución de escala que la evidencia criminológica reciente hace difícil de omitir, ya que el delito no se distribuye de forma homogénea ni siquiera dentro de un mismo municipio, sino que se concentra en una fracción mínima de microlugares (Weisburd, 2015), de modo que cualquier desagregación territorial que se detenga en el nivel municipal seguiría enfrentando, a menor escala, la misma limitación que hoy enfrenta la serie nacional agregada frente a la heterogeneidad documentada en la Figura 11. Conviene subrayar que, si esa desagregación llegara a mostrar una asociación estadística entre una condición territorial y un mayor nivel de homicidio, dicha asociación seguiría siendo, en los mismos términos ya señalados para la violencia intrafamiliar, una hipótesis por corroborar mediante herramientas de investigación social, y no un fundamento suficiente para justificar una narrativa de estigmatización o una asignación desproporcionada de vigilancia sobre un territorio.

Una tercera consideración, de naturaleza distinta a las dos anteriores, tiene que ver con la estabilidad estructural que el modelo asume de manera implícita. Todo el esquema

de validación descrito en la Sección 7.7 supone que las condiciones que generan el fenómeno se mantienen razonablemente estables entre el periodo de ajuste y el periodo de pronóstico, supuesto que un cambio regulatorio de magnitud suficiente puede alterar sin que el modelo tenga forma de anticiparlo. El Decreto 1368 de 2026, expedido por el Ministerio de Defensa Nacional, que levanta la suspensión general de los permisos para el porte de armas de fuego vigente en el país desde 2018, ofrece un ejemplo concreto y actual de esta situación, en la medida en que incide directamente sobre el arma de fuego, identificada en el análisis exploratorio como el medio ampliamente predominante en el homicidio colombiano según la Figura 10, y entra en vigencia apenas meses después de la ventana de prueba empleada en este estudio. Se trata, en los términos de la metodología Box-Jenkins presentada en la Sección 4.3, de una intervención externa (Box y Tiao, 1975) para la cual el modelo entrenado hasta diciembre de 2025 no cuenta con ningún mecanismo de ajuste incorporado, de forma análoga a como la serie ya registró, sin haberlo anticipado, el choque de la pandemia por COVID-19 en 2020. Por esta razón, cualquier uso prospectivo del pronóstico, y en particular el esquema de monitoreo mensual propuesto como trabajo futuro, debería incorporar de manera explícita la evaluación del efecto de este tipo de cambios regulatorios, en lugar de asumir que el desempeño documentado en este trabajo se mantiene de forma automática después de una modificación de esta naturaleza.

Finalmente, conviene precisar quién estaría en condiciones de emplear un pronóstico de este tipo y bajo qué resguardos. Los consumidores naturales de un ejercicio como este son entidades con responsabilidad directa en seguridad pública, entre ellas la Policía Nacional, el Ministerio de Defensa o del Interior, y la Fiscalía General de la Nación, y en todos los casos su uso operativo debería subordinarse a la supervisión humana explícita, sin convertirse en un disparador automático de despliegue de recursos o de

vigilancia. El Marco ético para la inteligencia artificial en Colombia (Guío Español et al., 2021) recoge precisamente este principio al establecer el control humano de las decisiones y la transparencia como condiciones para la implementación de proyectos de inteligencia artificial en el sector público, y su aplicación a este caso resulta directa: el modelo de mejor desempeño de este estudio, LightGBM, pertenece también al grupo de modelos cuya lógica interna no resulta directamente inspeccionable, de modo que su eventual incorporación operativa exigiría mantener y actualizar el análisis SHAP con la misma periodicidad que el pronóstico, no como un ejercicio técnico adicional sino como la condición que hace auditable, y por tanto legítima, su presencia en la toma de decisiones.

10. Conclusiones y Trabajo Futuro

Las conclusiones se refieren a los resultados alcanzados, los objetivos correspondientes y las hipótesis contrastadas. El trabajo futuro plantea las líneas de investigación que podrían derivarse de esta.

Conclusiones

La investigación se propuso comparar el desempeño de modelos estadísticos clásicos y de aprendizaje automático en el pronóstico mensual del homicidio intencional a escala nacional, evaluados frente a datos reales no utilizados durante el entrenamiento. A partir de los resultados, se derivan las siguientes conclusiones, alineadas con cada objetivo específico y contrastadas frente a las hipótesis planteadas en la Sección 5.

En respuesta al primer objetivo, se construyó una serie temporal mensual nacional de víctimas de homicidio intencional para el periodo 2015–2025, mediante la agregación de 110.381 registros correspondientes a 141.774 víctimas. Este subconjunto se obtuvo de una base consolidada de 3.785.395 registros de distintas temáticas delictivas del SIEDCO. La serie se complementó con variables exógenas socioeconómicas y delictivas, mediante un proceso de limpieza, agregación y construcción de variables con control explícito de fuga de información. El aporte central de esta fase fue establecer un pipeline temporalmente consistente, con un conjunto de prueba compuesto por datos reales de 2026 nunca utilizados en el ajuste de los modelos, que constituye una línea base metodológica replicable para el pronóstico del homicidio en el país.

En relación con el segundo objetivo y la Hipótesis 1, la comparación de nueve modelos, tres estadísticos (SARIMA, SARIMAX y VAR) y seis de aprendizaje automático (XGBoost, LightGBM, Random Forest, LSTM, GRU y Prophet), frente al homicidio efectivamente observado entre enero y mayo de 2026 mostró que LightGBM obtuvo el mejor desempeño con un margen amplio (RMSE de 16,61 y R^2 de 0,84), seguido de SARIMAX (R^2 = 0,37), Random Forest (R^2 = 0,26) y VAR (R^2 = 0,20), los únicos otros modelos con ajuste positivo; XGBoost, LSTM y GRU tuvieron un desempeño considerablemente peor que los modelos estadísticos. Estos resultados confirman la Hipótesis 1 solo de forma parcial: el aprendizaje automático puede superar de manera sustancial a los modelos estadísticos clásicos, pero este resultado depende del algoritmo

específico y no es una propiedad uniforme de la familia, como se discute en la Sección 9. Con cinco observaciones de prueba, además, la ventaja de LightGBM sobre SARIMAX es de magnitud amplia pero indicativa en términos inferenciales (prueba de Diebold-Mariano corregida, p = 0,078) (Tabla 45).

En cumplimiento del tercer objetivo y de la Hipótesis 2, el análisis de interpretabilidad mediante valores SHAP identificó a la violencia intrafamiliar del mes anterior como la variable individual más influyente en los tres modelos basados en árboles, incluido LightGBM, por encima del rezago anual de la propia serie de homicidios. Este hallazgo, junto con la ganancia predictiva directa de SARIMAX y VAR frente al modelo univariado, confirma parcialmente que las variables exógenas de naturaleza socioeconómica y delictiva aportan capacidad predictiva adicional más allá de la historia de la serie: la confirmación aplica en el sentido predictivo a los modelos estadísticos multivariados y en el sentido de importancia a los modelos de árboles, sin constituir evidencia de causalidad ni generalizarse a todas las familias evaluadas. El resultado sugiere, a la luz de la teoría de la desorganización social (Sampson y Groves, 1989), que la violencia intrafamiliar actúa como un indicador adelantado del deterioro del tejido social asociado a la violencia letal, lectura que debe mantenerse en el plano interpretativo dado el signo negativo de la relación estimada.

En relación con la Hipótesis 3, reformulada para el diseño de ventana única (Sección 5), el análisis del error por horizonte (Tabla 44) no la confirma: en el horizonte de un mes el menor error corresponde a un modelo estadístico (SARIMAX) y no a uno de aprendizaje automático, y el error de los modelos estadísticos no crece con el horizonte, sino que se concentra en los meses intermedios del periodo de prueba. Esta evaluación es descriptiva, pues dispone de una sola observación por horizonte; la formulación original

de la hipótesis, con horizontes de uno y doce meses, queda pendiente de verificación bajo un esquema de validación de origen móvil.

En atención al cuarto objetivo, el análisis por horizonte no permitió establecer que el desempeño relativo de cada familia de modelos dependa sistemáticamente de la distancia al último dato observado, tal como planteaba la Hipótesis 3. La evidencia disponible sugiere, en cambio, que la competitividad relativa de cada familia depende más de las características del algoritmo específico y de su interacción con el tamaño reducido de la muestra disponible, más que de una superioridad intrínseca de los modelos estadísticos o de aprendizaje automático como categorías. Los modelos de bagging (Random Forest) y de boosting con una configuración contenida (LightGBM) generalizaron adecuadamente a datos reales no vistos, mientras que la configuración más flexible de boosting seleccionada para XGBoost no lo hizo en la ventana de prueba, y las redes neuronales recurrentes (LSTM, GRU), en la configuración evaluada y con entradas sin escalar, no lograron anclar su pronóstico al nivel real de la serie, una por sobreestimación sistemática y otra por subestimación.

En conjunto, estas conclusiones responden a la pregunta de investigación: el desempeño comparativo entre modelos estadísticos y de aprendizaje automático no favorece de manera uniforme a ninguna de las dos familias; depende del algoritmo específico, y el mejor resultado global lo obtuvo LightGBM. Las variables más influyentes en el pronóstico combinan información exógena, en particular la violencia intrafamiliar, con la memoria de la propia serie. El estudio aporta una línea base reproducible y metodológicamente rigurosa, evaluada frente a datos reales prospectivos y no solo frente a un holdout histórico, antes inexistente en la literatura colombiana sobre predicción del homicidio con control explícito de fuga de información temporal.

Las conclusiones anteriores deben leerse junto con las condiciones de uso planteadas: el pronóstico obtenido es nacional y agregado, está orientado al monitoreo estratégico mensual y no constituye una herramienta para asignar recursos o vigilancia entre territorios ni para focalizar poblaciones. Su eventual despliegue exige supervisión humana, reporte de intervalos de incertidumbre, monitoreo continuo del error y transparencia sobre las variables que impulsan el pronóstico y sobre las limitaciones de los datos de registro administrativo que lo alimentan.

Trabajo futuro

A partir de los resultados y de las limitaciones identificadas, se plantean las siguientes líneas de investigación futura.

Profundización en variables exógenas. Si bien este trabajo incorporó variables socioeconómicas y delictivas con rezago de un mes, la exploración de otras fuentes (presencia de actores armados, economías ilegales, indicadores institucionales) podría enriquecer aún más la capacidad explicativa de los modelos, en particular de aquellos cuyo desempeño fue más débil**.**

Mayor granularidad territorial. Descender del nivel nacional al departamental y municipal permitiría aprovechar la autocorrelación espacial documentada en la literatura y capturar la heterogeneidad del fenómeno, abriendo la puerta a modelos espacio-temporales que integren la dimensión geográfica con la temporal.

Validación con mayor volumen de datos y configuraciones alternativas de redes recurrentes. El tamaño reducido de la serie limitó la estabilidad de los modelos más complejos, en particular las redes neuronales recurrentes, cuyo bajo desempeño se observó en una configuración específica que no incluyó escalamiento de las entradas, parada temprana, regularización L2, tasas de aprendizaje inferiores ni una malla de

búsqueda comparable a la de los modelos de árboles. Ampliar el periodo histórico o trabajar con una frecuencia de reporte más alta, y a la vez evaluar esas variantes de configuración, permitiría establecer si el fallo obedece al volumen de datos, a la configuración o a ambos.

Protocolos de validación más robustos. Reemplazar el holdout único por esquemas de validación de origen móvil (rolling-origin) con múltiples ventanas de distinto horizonte, por ejemplo, contrastando un horizonte corto de un mes frente a uno largo de doce, permitiría poner a prueba directamente la Hipótesis 3 en su formulación original y obtener comparaciones más concluyentes entre familias de modelos. Un esquema de este tipo produciría además series de errores de longitud suficiente para aplicar pruebas de igualdad de precisión predictiva, como la de Diebold-Mariano, con la potencia estadística de la que carece una ventana de prueba de cinco meses, y permitiría reentrenar ambas familias de modelos sobre una ventana histórica común.

Cuantificación de la incertidumbre en los modelos de aprendizaje automático. Los intervalos reportados para los modelos de árboles y las redes recurrentes son aproximaciones basadas en los residuales de validación. Obtener intervalos con fundamento estadístico, mediante regresión cuantílica (disponible de forma nativa en LightGBM), predicción conforme o bootstrap de residuales, exige reentrenar los modelos y se plantea como extensión directa de este trabajo.

Despliegue y monitoreo. Finalmente, se propone avanzar hacia la fase de despliegue de CRISP-DM, no abordada en este trabajo, mediante un sistema de monitoreo mensual que combine el pronóstico del modelo seleccionado con el seguimiento de las variables exógenas más influyentes, incorpore un componente de interpretabilidad (SHAP) y preserve el control de fuga de información temporal aplicado en esta investigación, bajo las condiciones de uso y salvaguardas planteadas.

Referencias

Adhikari, A., DeNero, J., y Wagner, D. (2022). What is data science? En *Inferential thinking*. https://inferentialthinking.com/chapters/01/what-is-data-science.html

Alsubayhin, A., Ramzan, M. S., y Alzahrani, B. (2024). Crime prediction model using three classification techniques: Random Forest, Logistic Regression, and LightGBM. *International Journal of Advanced Computer Science and Applications, 15*(1). https://doi.org/10.14569/IJACSA.2024.0150123

Anselin, L. (1995). Local indicators of spatial association,LISA. *Geographical Analysis, 27*(2), 93–115. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x

Arbona, A., Giménez, V., López Estrada, S., y Prior, D. (2024). The relationship between homicides from armed conflict and efficiency of educational quality in Colombia. *International Journal of Educational Development, 110*, 103120. https://doi.org/10.1016/j.ijedudev.2024.103120

Becker, G. S. (1968). Crime and punishment: An economic approach. *Journal of Political Economy, 76*(2), 169–217. https://doi.org/10.1086/259394

Bentéjac, C., Csörgő, A., y Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review, 54*(3), 1937–1967. https://doi.org/10.1007/s10462-020-09896-5

Bergmeir, C., y Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences, 191*, 192–213. https://doi.org/10.1016/j.ins.2011.12.028

Blum, A., Hopcroft, J., y Kannan, R. (2018). *Foundations of data science*. Cornell University.

Bontempi, G., Ben Taieb, S., y Le Borgne, Y.-A. (2013). Machine learning strategies for time series forecasting. En *Business intelligence* (Lecture Notes in Business Information Processing, Vol. 138, pp. 62–77). Springer. https://doi.org/10.1007/978-3-642-36318-4_3

Box, G. E. P., y Jenkins, G. M. (1970). Time series analysis: Forecasting and control. Holden-Day.

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., y Ljung, G. M. (2015). *Time series analysis: Forecasting and control* (5.ª ed.). Wiley.

Box, G. E. P., y Tiao, G. C. (1975). Intervention analysis with applications to economic and environmental problems. *Journal of the American Statistical Association, 70*(349), 70–79. https://doi.org/10.1080/01621459.1975.10480264

Bruce, P., y Bruce, A. (2017). *Practical statistics for data scientists: 50 essential concepts*. O'Reilly Media.

Cambridge International Examinations. (2015). *Cambridge International AS & A Level Information Technology*. Cambridge International Examinations.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., y Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv. https://doi.org/10.48550/arXiv.1406.1078

Cohen, L. E., y Felson, M. (1979). Social change and crime rate trends: A routine activity approach. *American Sociological Review, 44*(4), 588–608. https://doi.org/10.2307/2094589

Dickey, D. A., y Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. Journal of the American Statistical Association, 74(366), 427–431. https://doi.org/10.2307/2286348

Diebold, F. X. (2015). Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of Diebold-Mariano tests. Journal of Business & Economic Statistics, 33(1), 1-9. https://doi.org/10.1080/07350015.2014.983236

Diebold, F. X., y Mariano, R. S. (1995). Comparing predictive accuracy. Journal of Business & Economic Statistics, 13(3), 253-263. https://doi.org/10.1080/07350015.1995.10524599

Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., y Venkatasubramanian, S. (2018). Runaway feedback loops in predictive policing. En *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (Vol. 81, pp. 160–171). PMLR.

Fajnzylber, P., Lederman, D., y Loayza, N. (2002). Inequality and violent crime. *Journal of Law and Economics, 45*(1), 1–39. https://doi.org/10.1086/338347

Ferguson, A. G. (2017). The rise of big data policing: Surveillance, race, and the future of law enforcement. New York University Press.

Fondo Monetario Internacional. (2024). *Violent crime and insecurity in Latin America and the Caribbean: A macroeconomic perspective* (Departmental Paper N.º 2024/009). https://doi.org/10.5089/9798400288470.087

Gelvez-Ferreira, J. D., Nieto-Rodríguez, M. P., y Rocha-Ruiz, C. A. (2022). Prediciendo el crimen en ciudades intermedias: Un modelo de machine learning en Bucaramanga, Colombia. *URVIO Revista Latinoamericana de Estudios de Seguridad*, (34), 82–98. https://doi.org/10.17141/urvio.34.2022.5395

Gil-Vera, V., y Seguro-Gallego, C. (2022). Machine learning aplicado al análisis del rendimiento de desarrollos de software. *Revista Politécnica, 18*(35), 128–139. https://doi.org/10.33571/rpolitec.v18n35a9

Guío Español, A., Tamayo Uribe, E., Gómez Ayerbe, P., y Mujica, M. P. (2021). *Marco ético para la inteligencia artificial en Colombia*. Departamento Administrativo de la Presidencia de la República.

Harvey, D., Leybourne, S., y Newbold, P. (1997). Testing the equality of prediction mean squared errors. International Journal of Forecasting, 13(2), 281-291. https://doi.org/10.1016/S0169-2070(96)00719-4

Herrán, V. R., Martínez, M. Á. P., y Sogorb, A. O. (2008). Análisis de la eficacia de las medias móviles en el mercado intradiario de renta variable español. En *Universidad, sociedad y mercados globales* (pp. 56–68). Asociación Española de Dirección y Economía de la Empresa.

Hochreiter, S., y Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hyndman, R. J., y Athanasopoulos, G. (2021). *Forecasting: Principles and practice* (3.ª ed.). OTexts. https://otexts.com/fpp3/

Instituto Nacional de Medicina Legal y Ciencias Forenses. (2017). Forensis: Datos para la vida. Grupo Centro de Referencia Nacional sobre Violencia.

Instituto Nacional de Medicina Legal y Ciencias Forenses. (2019). Forensis: Datos para la vida. Grupo Centro de Referencia Nacional sobre Violencia.

Instituto Nacional de Medicina Legal y Ciencias Forenses. (2021). Forensis: Datos para la vida. Grupo Centro de Referencia Nacional sobre Violencia.

Instituto Nacional de Medicina Legal y Ciencias Forenses. (2022). *Forensis: Datos para la vida*. Grupo Centro de Referencia Nacional sobre Violencia.

Instituto Nacional de Medicina Legal y Ciencias Forenses. (2023). Forensis: Datos para la vida. Grupo Centro de Referencia Nacional sobre Violencia.

Instituto Nacional de Medicina Legal y Ciencias Forenses. (2024). Forensis: Datos para la vida. Grupo Centro de Referencia Nacional sobre Violencia.

International Business Machines. (s.f.). *¿Qué son las redes neuronales?* https://www.ibm.com/es-es/think/topics/neural-networks

Kapoor, S., y Narayanan, A. (2023). Leakage and the reproducibility crisis in machine learning-based science. *Patterns, 4*(9), 100804. https://doi.org/10.1016/j.patter.2023.100804

Kavlakoglu, E., y Russi, E. (s.f.). *¿Qué es XGBoost?* International Business Machines. https://www.ibm.com/mx-es/think/topics/xgboost

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., y Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. En *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates. https://papers.nips.cc/paper_files/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html

Kelleher, J. D., y Tierney, B. (2018). *Data science*. MIT Press.

Kim, B., Khanna, R., y Koyejo, O. (2016). Examples are not enough, learn to criticize! Criticism for interpretability. En *Proceedings of the 30th International Conference on Neural Information Processing Systems* (pp. 2288–2296). Curran Associates. https://papers.nips.cc/paper_files/paper/2016/hash/5680522b8e2bb01943234bce7bf84534-Abstract.html

Kleinberg, J., Ludwig, J., Mullainathan, S., y Obermeyer, Z. (2015). Prediction policy problems. *American Economic Review, 105*(5), 491–495.

Kontopoulou, V. I., Panagopoulos, A. D., Kakkos, I., y Matsopoulos, G. K. (2023). A review of ARIMA vs. machine learning approaches for time series forecasting in

data driven networks. *Future Internet, 15*(8), 255. https://doi.org/10.3390/fi15080255

Kroese, D. P., Botev, Z. I., Taimre, T., y Vaisman, R. (2024). *Data science and machine learning: Mathematical and statistical methods*. Chapman and Hall/CRC.

Kurama, V. (2025, 26 de mayo). *Making sense of gradient boosting in classification: A clear guide*. DigitalOcean. https://www.digitalocean.com/community/tutorials/gradient-boosting-for-classification

Kwiatkowski, D., Phillips, P. C. B., Schmidt, P., y Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. Journal of Econometrics, 54(1-3), 159–178. https://doi.org/10.1016/0304-4076(92)90104-Y

Ley 1581 de 2012. *Por la cual se dictan disposiciones generales para la protección de datos personales*. Congreso de la República de Colombia.

Lum, K., e Isaac, W. (2016). To predict and serve? *Significance, 13*(5), 14–19.

Lundberg, S. M., y Lee, S.-I. (2017). A unified approach to interpreting model predictions. En *Advances in Neural Information Processing Systems* (Vol. 30, pp. 4765–4774). Curran Associates. https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

Martinez, I., Viles, E., y Olaizola, I. G. (2021). Data science methodologies: Current challenges and future approaches. *Big Data Research, 24*, 100183. https://doi.org/10.1016/j.bdr.2020.100183

Messner, S. F., Anselin, L., y Baller, R. D. (1999). The spatial patterning of county homicide rates: An application of exploratory spatial data analysis. *Journal of*

Quantitative Criminology, 15(4), 423–450. https://doi.org/10.1023/A:1007544208712

Ministerio de Asuntos Económicos y Transformación Digital. (2021). *Guía práctica de introducción al análisis exploratorio de datos*. https://datos.gob.es/sites/default/files/doc/file/analisis_exploratorio_de_datos_2021.pdf

Núñez Rivera, L. C., Tolentino Pulido, F., y Rodríguez Barrios, H. (2023). Factores sociodemográficos en la dinámica del comportamiento delictivo: Análisis descriptivo de criminalidad en Colombia, año 2022. *Revista Criminalidad, 65*(3), 161–280. https://doi.org/10.47741/17943108.525

Observatorio de la Unidad para las Víctimas. (2023). *Índice de riesgo de victimización 2023*. https://datospaz.unidadvictimas.gov.co/archivos/datosPaz/infografias/Documento-IRV-2023.pdf

OpenStax. (2025). *Principles of data science*. Rice University. https://openstax.org/details/books/principles-data-science

Ordóñez-Eraso, H. A., Pardo-Calvache, C. J., y Cobos-Lozada, C. A. (2020). Detection of homicide trends in Colombia using machine learning. *Revista Facultad de Ingeniería, 29*(54), e11740. https://doi.org/10.19053/01211129.v29.n54.2020.11740

Pérez, A., Madera, W., Cortés, J., Ortiz, S., y Taborda, A. (2024). Correlación de las modalidades de homicidio con los delitos de impacto en cuadrantes atípicos de la Policía Nacional de Colombia. En J. Lozano-Medina (Ed.), *Analítica de datos para la seguridad ciudadana* (pp. 113–140). Editorial de la Dirección de Educación Policial. https://doi.org/10.22335/y1jrrh33

Policía Nacional de Colombia. (s.f.). *Homicidios* [Base de datos]. En Estadística delictiva. https://www.policia.gov.co/estadistica-delictiva

Polyzos, E., y Siriopoulos, C. (2023). Autoregressive random forests: Machine learning and lag selection for financial research. *Computational Economics*. https://doi.org/10.1007/s10614-023-10429-9

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Said, S. E., y Dickey, D. A. (1984). Testing for unit roots in autoregressive-moving average models of unknown order. Biometrika, 71(3), 599–607. https://doi.org/10.1093/biomet/71.3.599

Sampson, R. J., y Groves, W. B. (1989). Community structure and crime: Testing social-disorganization theory. *American Journal of Sociology, 94*(4), 774–802. https://doi.org/10.1086/229068

Sánchez Jabba, A., Díaz, A. M., Peláez, A., Castelblanco, O., Tautiva, J., González, C., y Ángel, L. (2012). *Evolución geográfica del homicidio en Colombia* (Documentos de Trabajo sobre Economía Regional, n.º 169). Banco de la República. https://doi.org/10.32468/dtseru.169

Sánchez, J., y Núñez, J. (2001). Determinantes del crimen violento en un país altamente violento: El caso de Colombia. *Coyuntura Económica, 31*(1), 91–116. http://hdl.handle.net/11445/2098

Sandoval, L. (2014). Transición demográfica, condición económica y tasa de homicidio: El caso de Colombia, 1993–2013. *Ecos de Economía, 18*(39), 97–111. https://doi.org/10.17230/ecos.2014.39.5

Santos-Marquez, F. (2022). Spatial beta-convergence forecasting models: Evidence from municipal homicide rates in Colombia. *Journal of Forecasting, 41*(2), 294–302. https://doi.org/10.1002/for.2816

Santos-Marquez, F., y Ribeiro-Silveira, R. (2021). Regional convergence, spatial scale, and spatial dependence: Evidence from homicides and personal injuries in Colombia 2010–2018. *Regional Science Policy & Practice, 15*(4), 893–910. https://doi.org/10.1111/rsp3.12479

Shapley, L. S. (1953). A value for n-person games. En H. W. Kuhn y A. W. Tucker (Eds.), *Contributions to the theory of games* (Vol. 2, pp. 307–317). Princeton University Press.

Sims, C. A. (1980). Macroeconomics and reality. Econometrica, 48(1), 1–48. https://doi.org/10.2307/1912017

Small, M. L. (2011). How to conduct a mixed methods study: Recent trends in a rapidly growing literature. *Annual Review of Sociology, 37*, 57–86.

Soares Filho, A. M., Vasconcelos, C. H., Cunningham, M., Ribeiro, A., Naghavi, M., y Carvalho Malta, D. (2024). Spatial association of homicide rate with violence, sociodemographic, and public security factors: Global Burden of Disease Study 2018 for municipalities in Brazil. *Public Health, 227*, 16–23. https://doi.org/10.1016/j.puhe.2023.10.052

Soares, R. R., y Naritomi, J. (2010). Understanding high crime rates in Latin America: The role of social and policy factors. En R. Di Tella, S. Edwards, y E. Schargrodsky (Eds.), *The economics of crime: Lessons for and from Latin America* (pp. 19–55). University of Chicago Press.

Sosa, J., Briz-Redón, A., Flores, M., Abril, M., y Mateu, J. (2023). A spatio-temporal multinomial model of firearm death in Ecuador. *Spatial Statistics, 54*, 100738. https://doi.org/10.1016/j.spasta.2023.100738

Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: An analysis and review. *International Journal of Forecasting, 16*(4), 437–450. https://doi.org/10.1016/S0169-2070(00)00065-0

Taylor, S. J., y Letham, B. (2018). Forecasting at scale. The American Statistician, 72(1), 37–45. https://doi.org/10.1080/00031305.2017.1380080

United Nations Office on Drugs and Crime. (2023). *Global study on homicide 2023*. United Nations. https://www.unodc.org/documents/data-and-analysis/gsh/2023/Global_study_on_homicide_2023_web.pdf

Verrey, J., Ariel, B., Harinam, V., y Dillon, L. (2023). Using machine learning to forecast domestic homicide via police data and super learning. *Scientific Reports, 13*, 22932. https://doi.org/10.1038/s41598-023-50274-2

Weisburd, D. (2015). The law of crime concentration and the criminology of place. *Criminology, 53*(2), 133–157.

Wirth, R., y Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. En *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining* (pp. 29–39). Practical Application Company.

Zhang, X., Liu, L., Xiao, L., y Ji, J. (2020). Comparison of machine learning algorithms for predicting crime hotspots. *IEEE Access, 8*, 181302–181310. https://doi.org/10.1109/ACCESS.2020.3028420