

**Predicción Temprana de Brotes de Dengue en Bucaramanga Mediante un Modelo de
Machine Learning con Integración de Variables Climáticas, Geográficas y
Sociodemográficas**

Reinaldo Araque Villamizar y Diego Alberto Paipa Bautista

**Trabajo de grado para optar el título de Magíster en Análisis de Datos y Sistemas
Inteligentes**

Director

Yuli Andrea Alavarez Pizarro

Magister en Ingeniería Eléctrica

Universidad Santo Tomás, Bucaramanga

Área de Ingenierías y Arquitectura

Maestría en Análisis de Datos y Sistemas Inteligentes

2026

Contenido

Predicción Temprana de Brotes de Dengue en Bucaramanga Mediante un Modelo de Machine Learning con Integración de Variables Climáticas, Geográficas y Sociodemográficas	1
Introducción	11
1. Predicción temprana de brotes de dengue en Bucaramanga mediante un modelo de machine learning con integración de variables climáticas, geográficas y sociodemográficas.....	14
1.1 Planteamiento del problema.....	14
1.1.1 Contexto del problema.....	14
1.1.2 Descripción clara y delimitada de la problemática.....	15
1.2 Justificación del proyecto.....	17
1.3 Objetivos	18
1.3.1 Objetivo general	18
1.3.2 Objetivos específicos.....	18
1.4 Alcance y delimitaciones del proyecto.....	19
2. Marco referencial	20
2.1 Marco teórico	20
2.1.1 Conceptos teóricos relevantes	21
2.1.2 Modelos, técnicas y enfoques existentes	25
2.1.3 Revisión de trabajos previos relacionados	28
2.1.4 Identificación de vacíos o limitaciones en la literatura	31
2.2 Marco conceptual	35
3. Metodología	37
3.1 Tipo y enfoque de la investigación	37

3.2 Descripción de las fases del proyecto	38
3.3 Fuentes de datos y criterios de selección	42
3.3.1 SIVIGILA — Sistema Nacional de Vigilancia en Salud Pública	42
3.3.2 IDEAM — Datos climatológicos	42
3.3.3 Sentinel-2 — Datos satelitales.....	44
3.3.4 DANE — Variables sociodemográficas.....	45
3.4 Herramientas, técnicas y tecnologías empleadas	46
3.4.1 Lenguaje de programación y entorno de desarrollo	46
3.4.2 Bibliotecas de análisis de datos y machine learning.....	46
3.4.3 Procesamiento de datos satelitales.....	47
3.4.4 Visualización y dashboard.....	47
3.5 Métodos de análisis y criterios de evaluación	47
3.5.1 Estrategia de partición temporal.....	47
3.5.2 Métricas de evaluación	48
3.5.3 Análisis de importancia de variables	49
3.5.4 Estrategia diferenciada de integración de rezagos según arquitectura	49
3.5.5 Comparación estadística entre modelos	50
3.5.6 Consideraciones éticas y de equidad	51
3.6 Cronograma de actividades	51
4. Desarrollo de la solución	51
4.1 Arquitectura general de la solución.....	52
4.1.1 Capa 1: recolección e integración de datos	52
4.1.2 Capa 2: preprocesamiento y homologación.....	53

4.1.3 Capa 3: ingeniería de características (feature engineering).....	53
4.1.4 Capa 4: modelado y evaluación comparativa	54
4.1.5 Capa 5: sistema de alerta y visualización	55
4.2 Diseño del modelo, sistema o metodología.....	56
4.2.1 Diseño del pipeline de datos	56
4.2.2 Diseño de los modelos de machine learning	56
4.2.3 Diseño de la estrategia de evaluación comparativa	58
4.3 Implementación técnica.....	59
4.3.1 Entorno tecnológico.....	59
4.3.2 Proceso de preprocesamiento e integración de datos	60
4.3.3 Implementación del feature engineering	60
4.3.4 Implementación de los modelos de machine learning.....	61
4.3.5 Implementación de métricas y comparación de modelos.....	61
4.4 Descripción del prototipo o sistema desarrollado	62
4.4.1 Descripción general del prototipo	62
4.4.2 Módulo 1: actualización de datos	62
4.4.3 Módulo 2: predicción con el modelo seleccionado	62
4.4.4 Módulo 3: sistema de semáforo y visualización.....	64
4.4.5 Alcance y limitaciones del prototipo	65
5. Resultados y validación	65
5.1 Resultados experimentales o de aplicación.....	66
5.1.1 Patrones temporales de la incidencia (objetivo específico 1).....	66
5.1.2 Distribución espacial del riesgo (objetivo específico 2).....	69

5.1.3 Modelos predictivos (objetivo específico 3)	71
5.2 Métricas de desempeño	72
5.3 Análisis e interpretación de los resultados	74
5.4 Validación frente a los objetivos planteados.....	78
6. Conclusiones y recomendaciones	79
6.1 Conclusiones generales del proyecto	79
6.2 Cumplimiento de los objetivos.....	80
6.3 Selección del modelo del sistema.....	81
6.4 Limitaciones identificadas.....	82
6.5 Recomendaciones y trabajos futuros.....	83
6.6 Consideraciones finales.....	84
7. Referencias.....	86
Apéndices.....	89

Lista de tablas

Tabla 1. Resumen de fuentes de datos, variables y criterios de selección.....	46
Tabla 2. Cronograma estimado de actividades del proyecto de grado	51
Tabla 3. Resumen de las cinco capas funcionales de la arquitectura de la solución	55
Tabla 4 Bibliotecas de Python utilizadas en la implementación del pipeline.....	59
Tabla 5. Niveles de alerta del sistema DengueAlert-BGA y acciones asociadas	64
Tabla 6. Casos, participación y tasa de hospitalización por comuna (2015–2022).	70
Tabla 7. Desempeño promedio por modelo (horizontes 1–4 semanas, prueba 2019–2020).	73
Tabla 8. MASE y F1 por horizonte de predicción (MASE / F1).	73

Lista de figuras

Figura 1. Incidencia semanal de dengue y media móvil de cuatro semanas (Bucaramanga, 2015–2020).	67
Figura 2. Casos promedio por semana epidemiológica (2015–2020).....	67
Figura 3. Descomposición STL de la serie de incidencia (2015–2020).	68
Figura 4. Correlación cruzada entre la incidencia y las variables climáticas por rezago temporal.	69
Figura 5. Casos acumulados de dengue por comuna (2015–2022).	70
Figura 6. Mapa de calor espacio-temporal: casos por comuna y año.	71
Figura 7. Predicción a una semana frente a los casos observados (prueba 2019–2020).	72
Figura 8. Error escalado (MASE) por horizonte; la línea roja marca el desempeño del naïve. ...	73
Figura 9. Importancia por permutación de las variables — Random Forest ($h = 1$).	76
Figura 10. Importancia SHAP media de las variables — XGBoost ($h = 1$).	76

Resumen

Problema. El dengue es una arbovirosis endémica en Bucaramanga que entre 2015 y 2022 acumuló 10.259 casos notificados al SIVIGILA, con una tasa de hospitalización del 26,6 % y picos epidémicos recurrentes; sin embargo, el municipio no dispone de herramientas que anticipen sus brotes. **Objetivo.** Desarrollar un modelo de machine learning para la predicción temprana de brotes de dengue en Bucaramanga, integrando variables epidemiológicas, climáticas, geográficas y sociodemográficas. **Método.** Investigación cuantitativa, aplicada y longitudinal que, siguiendo el proceso CRISP-DM, integró datos del SIVIGILA, registros climáticos del IDEAM, índices de vegetación (NDVI) de Sentinel-2 e indicadores del DANE, agregados por semana epidemiológica y comuna. Se compararon cinco enfoques —un modelo naïve estacional y ARIMA como líneas base frente a Random Forest, XGBoost y LSTM— mediante una partición temporal estricta y horizontes de predicción de una a cuatro semanas, evaluados con las métricas MAE, RMSE, MAPE, MASE, F1 y AUC, la prueba de Diebold-Mariano y la interpretabilidad SHAP. **Resultados.** Los modelos de machine learning redujeron el error de predicción a menos de dos tercios del pronóstico ingenuo y discriminaron las semanas de brote de forma estadísticamente significativa ($p < 0,001$). La temperatura, con un rezago de siete semanas, fue el predictor ambiental más relevante, y el riesgo se concentró en las comunas San Francisco, Centro y Oriental (32,4 % de los casos). **Conclusiones.** Es factible anticipar los brotes con datos públicos locales; el prototipo DengueAlert-BGA constituye una herramienta interpretable y replicable para fortalecer la vigilancia y la respuesta sanitaria del municipio.

Palabras clave: dengue, predicción de brotes, machine learning, vigilancia epidemiológica, alerta temprana

Abstract

Problem. Dengue is an endemic arboviral disease in Bucaramanga that accumulated 10,259 cases reported to the National Public Health Surveillance System (SIVIGILA) between 2015 and 2022, with a hospitalization rate of 26.6% and recurrent epidemic peaks. However, the municipality lacks tools capable of anticipating outbreaks. **Objective.** To develop a machine learning model for the early prediction of dengue outbreaks in Bucaramanga by integrating epidemiological, climatic, geographic, and sociodemographic variables. **Methods.** A quantitative, applied, and longitudinal study was conducted following the CRISP-DM methodology. Data from SIVIGILA, climate records from IDEAM, Sentinel-2 vegetation indices (NDVI), and sociodemographic indicators from DANE were integrated and aggregated by epidemiological week and administrative district. Five approaches were compared: a seasonal naïve model and ARIMA as baseline models, versus Random Forest, XGBoost, and LSTM. Models were evaluated using a strict temporal split and prediction horizons ranging from one to four weeks. Performance was assessed using MAE, RMSE, MAPE, MASE, F1-score, and AUC metrics, complemented by the Diebold–Mariano test and SHAP-based interpretability analysis. **Results.** Machine learning models reduced prediction error to less than two-thirds of the naïve forecast and significantly discriminated outbreak weeks ($p < 0.001$). Temperature, with a seven-week lag, was identified as the most relevant environmental predictor. Outbreak risk was concentrated in the San Francisco, Centro, and Oriental districts, which accounted for 32.4% of reported cases. **Conclusions.** Dengue outbreaks can be anticipated using publicly available local data. The DengueAlert-BGA prototype represents an interpretable and replicable tool to strengthen epidemiological surveillance and public health response in the municipality.

Keywords: dengue, outbreak prediction, machine learning, epidemiological surveillance, early warning system.

Introducción

El dengue representa uno de los desafíos más persistentes de la salud pública en Colombia y en la región de Santander. Transmitida principalmente por el mosquito *Aedes aegypti*, esta arbovirosis de notificación obligatoria afecta de manera recurrente a poblaciones urbanas y periurbanas, generando una presión sostenida sobre los sistemas de salud, especialmente durante los picos epidémicos. Bucaramanga, capital del departamento de Santander, no es la excepción: entre 2015 y 2022, el municipio registró 10.259 casos confirmados en el sistema de vigilancia epidemiológica SIVIGILA, con una tasa de hospitalización del 26,6% y ciclos de rebrote asociados a variables climáticas, como la temperatura y las precipitaciones, que determinan las condiciones óptimas para la reproducción del vector.

A pesar de la disponibilidad de registros epidemiológicos históricos de casi dos décadas, Bucaramanga carece de herramientas predictivas que permitan anticipar los brotes de dengue con la suficiente antelación para orientar intervenciones preventivas focalizadas. Esta brecha contrasta con los avances documentados en otros contextos nacionales e internacionales, donde el uso de modelos de machine learning —integrados con datos climáticos, geoespaciales y sociodemográficos— ha demostrado capacidad para reducir la incertidumbre epidemiológica y fortalecer la respuesta sanitaria proactiva.

En este contexto, el presente trabajo de grado propone el desarrollo del prototipo DengueAlert-BGA: un sistema predictivo de brotes de dengue en el municipio de Bucaramanga basado en tres arquitecturas de machine learning —Random Forest, XGBoost y Long Short-Term Memory (LSTM)— que integran variables epidemiológicas del SIVIGILA, registros meteorológicos del IDEAM, índices de vegetación derivados de imágenes Sentinel-2 (NDVI) e indicadores socioeconómicos y de densidad poblacional del DANE. El horizonte de predicción del

sistema abarca de una a cuatro semanas ($h=1$ a $h=4$), con el propósito de generar alertas tempranas accionables que apoyen la toma de decisiones en la Secretaría de Salud municipal.

La selección del modelo con mayor capacidad predictiva se realiza mediante métricas de error —MASE y RMSE— complementadas con el test de Diebold-Mariano para la comparación estadística de desempeño entre modelos, y con valores SHAP para la interpretación de las contribuciones de cada variable predictora. De este modo, el proyecto no solo busca generar una herramienta de predicción eficaz, sino también aportar evidencia interpretable que contribuya a la comprensión de los factores determinantes de los brotes en el contexto local.

Metodológicamente, la investigación se enmarca en un diseño cuantitativo, aplicado y longitudinal, con una partición temporal de los datos: el período 2015–2018 se utiliza para el entrenamiento de los modelos, TimeSeriesSplit dentro del entrenamiento para la validación y 2019–2020 para la prueba. El análisis combina técnicas de series de tiempo, modelado de machine learning y análisis espacial a nivel de comuna, siguiendo los lineamientos del proceso CRISP-DM adaptado al dominio de la salud pública.

El presente documento está estructurado en seis capítulos. El Capítulo 1, Planteamiento del Problema, presenta el contexto epidemiológico de Bucaramanga, la formulación del problema de investigación, la justificación del enfoque propuesto y los objetivos del trabajo, tanto el general como los específicos. El Capítulo 2, Marco Teórico y Estado del Arte, desarrolla los fundamentos conceptuales de las enfermedades transmitidas por vectores, las bases de los modelos de machine learning empleados —Random Forest, XGBoost y LSTM— y una revisión sistemática de los antecedentes nacionales e internacionales más relevantes en la predicción de dengue mediante inteligencia artificial. El Capítulo 3, Metodología, describe el enfoque de investigación, las fuentes y el proceso de recolección y preprocesamiento de datos, el diseño experimental para el

entrenamiento y evaluación comparativa de los modelos, y los criterios de selección del sistema final. Finalmente, el Capítulo 4, Desarrollo de la Solución, detalla la arquitectura del prototipo DengueAlert-BGA, el proceso de construcción y ajuste de los modelos, el Capítulo 5 Los Resultados obtenidos en términos de precisión predictiva y el Capítulo 6 que son el análisis de interpretabilidad y conclusiones.

Este proyecto se enmarca en la Maestría en Análisis de Datos y Sistemas Inteligentes de la Universidad Santo Tomás, sede Bucaramanga, y constituye un aporte al campo de la analítica avanzada aplicada a la salud pública, con potencial de replicabilidad en otros municipios del país con perfiles epidemiológicos similares.

1. Predicción temprana de brotes de dengue en Bucaramanga mediante un modelo de machine learning con integración de variables climáticas, geográficas y sociodemográficas

1.1 Planteamiento del problema

Este capítulo presenta el contexto general y específico del problema que motiva el proyecto de investigación, describiendo la situación epidemiológica del dengue en Bucaramanga, las brechas tecnológicas identificadas y la relevancia académica y aplicada de desarrollar una solución basada en técnicas avanzadas de machine learning integradas con variables climáticas, geográficas y sociodemográficas.

1.1.1 Contexto del problema

El dengue es una enfermedad viral transmitida por mosquitos del género *Aedes*, principalmente *Aedes aegypti*, que constituye una de las arbovirosis de mayor impacto en salud pública a nivel global. Según la Organización Mundial de la Salud (OMS), aproximadamente la mitad de la población mundial se encuentra en riesgo de contraer esta enfermedad, con estimaciones de 100 a 400 millones de infecciones anuales (OMS, 2024). En Colombia, el dengue es endémico en regiones tropicales y subtropicales, y representa una carga sanitaria significativa para los sistemas de salud municipal y departamental.

Bucaramanga, capital del departamento de Santander, se sitúa en una región geográfica y climática que favorece la proliferación del vector transmisor del dengue. Las características topográficas, la variación de temperatura y precipitación, la densidad poblacional y las condiciones socioeconómicas de sus 17 comunas configuran un entorno heterogéneo en el que los brotes de dengue se presentan de manera recurrente y con distribución espacial desigual. Entre 2015 y 2022, el municipio registró 10.259 casos en el sistema SIVIGILA, con picos epidémicos significativos y

una tasa de hospitalización del 26,6% (Instituto Nacional de Salud [INS], 2024a), lo que evidencia la presión sostenida sobre la infraestructura sanitaria local.

A pesar de contar con un registro histórico de datos epidemiológicos de casi veinte años, dichos datos no han sido utilizados de manera sistemática para desarrollar herramientas predictivas que permitan anticipar los brotes con la suficiente antelación para implementar medidas preventivas focalizadas y eficaces. Esta situación contrasta con avances internacionales documentados en países como Singapur, Brasil y Colombia a escala nacional, donde se han implementado modelos de inteligencia artificial y machine learning con resultados prometedores en la predicción de brotes de enfermedades transmitidas por vectores (Chen y Moraga, 2025; Sebastianelli et al., 2024; Zhao et al., 2020).

En este contexto, el presente proyecto de investigación aplicada se enmarca en la intersección entre la vigilancia epidemiológica, el análisis de datos avanzado y los sistemas inteligentes, con el propósito de cerrar la brecha tecnológica identificada en el ámbito municipal de Bucaramanga mediante el desarrollo de un modelo predictivo robusto y replicable.

1.1.2 Descripción clara y delimitada de la problemática

La distribución heterogénea de los casos de dengue en Bucaramanga sugiere que factores socioeconómicos, ambientales y geográficos son determinantes críticos en la transmisión de la enfermedad. Pese a disponer de un registro histórico de datos de casi veinte años, estos no han sido utilizados sistemáticamente para desarrollar un sistema de predicción que anticipe brotes con la suficiente antelación para implementar medidas preventivas eficaces.

Esta brecha tecnológica se fundamenta en la ausencia de aplicación de técnicas avanzadas de análisis de datos y machine learning para identificar patrones complejos en los datos de vigilancia disponibles. Mientras estudios como los de Sebastianelli et al. (2024) han demostrado

el potencial de modelos ensemble —combinando CatBoost, SVM y LSTM— para predecir brotes con un mes de anticipación, y Chen y Moraga (2025) han aplicado exitosamente redes LSTM para predicción a nivel estatal en Brasil, estas metodologías avanzadas no han sido implementadas en el contexto local de Bucaramanga.

Adicionalmente, existe una brecha conceptual y metodológica derivada de la escasez de estudios aplicados que integren de manera efectiva datos epidemiológicos con variables climáticas, geográficas y sociodemográficas a nivel municipal. Aunque Zhao et al. (2020) desarrollaron modelos de Random Forest y redes neuronales para Colombia utilizando datos del SIVIGILA, su análisis se limitó a la escala nacional y departamental, sin descender al nivel de detalle que requiere Bucaramanga. Dicho estudio usó variables satelitales (EVI, CMORPH, MODIS) y sociodemográficas (población, índice Gini), pero los autores reconocen explícitamente limitaciones como la falta de datos entomológicos y que los modelos Random Forest no pueden extrapolar valores fuera del rango de entrenamiento.

Finalmente, existe una brecha estratégica en la carencia de investigación aplicada en Santander que desarrolle e implemente soluciones basadas en inteligencia artificial para la vigilancia epidemiológica a nivel municipal, lo que contrasta con los casos de éxito internacionales documentados que demuestran la viabilidad e impacto de estos enfoques en la gestión proactiva de la enfermedad (Instituto Nacional de Salud [INS], 2024a; Codeço et al., 2018).

Esta ausencia de herramientas predictivas ha derivado en una gestión reactiva de los brotes: las campañas de fumigación y descacharrización, así como el refuerzo de la capacidad hospitalaria, se activan cuando el pico de contagios ya está en curso. El resultado es una morbilidad evitable, una mayor presión sobre la red de urgencias y una asignación tardía e ineficiente de los recursos de control vectorial. En un municipio donde el 26,6 % de los casos requirió hospitalización, actuar

después del pico —y no antes— tiene un costo sanitario y social directo sobre la población más expuesta.

Pregunta de investigación: ¿Cómo usar los datos históricos de dengue de Bucaramanga para predecir brotes y generar alertas tempranas, integrando variables climáticas, geográficas y sociodemográficas?

1.2 Justificación del proyecto

La justificación de este proyecto es, ante todo, humanitaria y sanitaria. Con una tasa de hospitalización del 26,6 % y picos que han superado los 100 casos semanales, el dengue no solo representa una carga clínica para la red hospitalaria de Bucaramanga, sino que afecta de manera desproporcionada a la población de estratos bajos, que enfrenta mayores barreras de acceso a la atención oportuna. Anticipar un brote con hasta cuatro semanas de antelación no es un ejercicio académico: puede ser la diferencia entre que un paciente reciba atención en un servicio preparado o en uno colapsado.

Desde el punto de vista científico, el proyecto llena un vacío de escala. Aunque existen antecedentes que abordan Colombia a nivel nacional y departamental, y otros que trabajan a escala estatal en Brasil, no se dispone de un modelo calibrado y validado para la escala intraurbana de una ciudad intermedia colombiana. Bucaramanga presenta una heterogeneidad espacial (17 comunas) que los modelos nacionales suavizan y ocultan, lo que exige un desarrollo metodológico propio para capturar los determinantes locales del riesgo. En este plano, los modelos de machine learning constituyen la vía metodológica capaz de procesar la alta dimensionalidad y las relaciones no lineales de datos heterogéneos —epidemiológicos, climáticos y territoriales— que los métodos univariados no logran aprovechar.

Finalmente, el proyecto tiene una justificación económica ligada a la asignación eficiente de los recursos públicos. Al focalizar las campañas de descacharrización y control vectorial en las comunas con alerta roja, se evita el gasto disperso en toda la ciudad y se optimiza el uso de las brigadas de salud, cuyo costo operativo es sensiblemente menor que el de atender un brote descontrolado con alta demanda hospitalaria.

1.3 Objetivos

1.3.1 *Objetivo general*

Desarrollar un modelo de machine learning para la predicción temprana de brotes de dengue en Bucaramanga, mediante el análisis de datos históricos de vigilancia epidemiológica integrados con variables climáticas, geográficas y sociodemográficas.

1.3.2 *Objetivos específicos*

Identificar los patrones temporales (estacionalidad y tendencias) de la incidencia de dengue (SIVIGILA 2015-2022) y su correlación con variables climáticas (IDEAM), para la determinación de las variables temporales y de retardo (features) del modelo predictivo.

Determinar la distribución espacial del riesgo de dengue y las características sociodemográficas asociadas en las comunas de Bucaramanga, para su validación e integración como predictores en el modelo de machine learning.

Evaluar comparativamente el desempeño de modelos de machine learning (Random Forest, XGBoost y LSTM) frente a líneas base estadísticas (modelo naïve estacional y ARIMA), mediante métricas de error (MASE, RMSE) y de clasificación de brotes (F1, AUC), para seleccionar el modelo con mayor capacidad predictiva de brotes de dengue en horizontes de una a cuatro semanas.

1.4 Alcance y delimitaciones del proyecto

El alcance de esta investigación es metodológico y predictivo, enfocado en desarrollar y validar un modelo de machine learning para la predicción de brotes de dengue en el municipio de Bucaramanga, siguiendo un flujo de trabajo unificado en tres fases interdependientes:

Fase 1 — Definición de características temporales (OE 1): Análisis de los datos epidemiológicos del SIVIGILA "13. Dengue, Dengue grave y mortalidad por dengue municipio De Bucaramanga" (2015-2022), integrados con variables climáticas del IDEAM (precipitación, temperatura, humedad). El entregable de esta fase es un conjunto de features temporales validados (estacionalidad, tendencias, lags).

Fase 2 — Definición de características espaciales (OE 2): Análisis espacial del riesgo de dengue en las 17 comunas de Bucaramanga y su asociación con variables sociodemográficas del DANE (NBI, densidad poblacional). El entregable es un conjunto de predictores geoespaciales y sociodemográficos validados.

Fase 3 — Fase experimental y de evaluación (OE 3): Construcción de la matriz de datos final integrando los features de las fases anteriores. Entrenamiento y evaluación comparativa de tres algoritmos (Random Forest, XGBoost, LSTM) usando métricas cuantitativas (Precisión, RMSE) en horizontes de 1 a 4 semanas. El entregable es el modelo optimizado con mejor desempeño predictivo.

Las delimitaciones del proyecto son las siguientes:

- El análisis epidemiológico se limita al periodo 2015-2022, con datos del registro oficial del SIVIGILA para el municipio de Bucaramanga.

- El alcance geográfico comprende exclusivamente el área urbana de Bucaramanga, dividida en sus 17 comunas, sin incluir el área metropolitana (Floridablanca, Girón, Piedecuesta).
- El proyecto no contempla el diseño ni la implementación de un sistema de alerta temprana en producción; el producto final es un modelo predictivo validado y documentado que podrá servir como base para desarrollos futuros.
- No se incluyen datos entomológicos (índices larvarios, densidad de vectores) en el modelo predictivo, dado que este tipo de información no se encuentra disponible de manera sistematizada para el periodo de análisis seleccionado.

2. Marco referencial

Este capítulo presenta los fundamentos conceptuales y los antecedentes que sustentan teóricamente el desarrollo del proyecto. Se describen los conceptos clave relacionados con la vigilancia epidemiológica del dengue, los modelos de machine learning aplicados a predicción epidemiológica y las variables multimodales relevantes. Posteriormente, se realiza una revisión sistemática de los trabajos previos más relevantes y se identifican las brechas que justifican la realización de esta investigación en el contexto específico de Bucaramanga.

2.1 Marco teórico

El desarrollo de un sistema predictivo orientado a la salud pública, requiere de una fundamentación interdisciplinaria sólida que articule la epidemiología médica, la biología del vector y las ciencias de la computación, dado que la modelación basada en machine learning no opera en un vacío, sino que se sustenta en dinámicas biológicas de transmisión y en la rigurosidad de los registros oficiales. Por consiguiente, este marco teórico se estructura en torno a tres ejes

fundamentales: en primer lugar, se aborda el dengue desde su dimensión clínica y el rol crítico del Sistema Nacional de Vigilancia en Salud Pública (SIVIGILA) en Colombia como fuente primaria de datos; en segundo lugar, se exploran los principios del Machine Learning y el Deep Learning aplicados a series temporales, examinando las ventajas de las arquitecturas de ensamble y las redes neuronales recurrentes para capturar interacciones complejas; y finalmente, se conceptualiza la integración multimodal de variables ambientales, climáticas y sociodemográficas, estableciendo los mecanismos biológicos e históricos que justifican su incorporación como variables predictoras en el contexto urbano de Bucaramanga.

2.1.1 Conceptos teóricos relevantes

2.1.1.1 El dengue como problema de salud pública. El dengue es una enfermedad viral transmitida por mosquitos del género *Aedes*, principalmente *Aedes aegypti*, que constituye una de las arbovirosis de mayor impacto en salud pública a nivel global. Según la Organización Mundial de la Salud (OMS), aproximadamente la mitad de la población mundial se encuentra en riesgo de contraer esta enfermedad, con estimaciones de 100 a 400 millones de infecciones anuales. En Colombia, el dengue es endémico en regiones tropicales y subtropicales, con ciclos epidémicos recurrentes que representan una carga sanitaria significativa para los sistemas de salud municipal y departamental.

El dengue se clasifica clínicamente en tres categorías según la gravedad: dengue sin signos de alarma (código SIVIGILA 210), dengue con signos de alarma (código 220) y dengue grave (código 580). Esta clasificación, definida por la OMS en 2009 y adoptada por el Instituto Nacional de Salud de Colombia, es fundamental para la vigilancia epidemiológica y determina las variables dependientes del modelo predictivo propuesto. El vector principal, *Aedes aegypti*, tiene un ciclo

de vida altamente influenciado por variables climáticas como temperatura, precipitación y humedad relativa, lo que establece una base biológica sólida para la incorporación de datos climatológicos en los modelos predictivos.

2.1.1.2 Sistema Nacional de Vigilancia en Salud Pública (SIVIGILA). El SIVIGILA constituye la fuente primaria de datos epidemiológicos para este proyecto. Es el sistema oficial de notificación de eventos de salud pública en Colombia, administrado por el Instituto Nacional de Salud (INS). Para el dengue, la ficha de notificación captura variables demográficas (edad, sexo, municipio de residencia, barrio, estrato socioeconómico), variables clínicas (fecha de inicio de síntomas, clasificación final del caso, hospitalización, conducta), variables temporales (fecha de notificación, semana epidemiológica) y variables de laboratorio (resultados de pruebas diagnósticas).

El análisis del SIVIGILA para Bucaramanga en el período 2015-2022 revela 10.259 casos notificados, con picos epidémicos significativos en 2015 y 2019, una tasa de hospitalización del 26,6% y una distribución espacial heterogénea entre las 17 comunas del municipio. Históricamente, estos datos han sido utilizados de manera principalmente retrospectiva y descriptiva, sin aprovecharse su potencial predictivo mediante técnicas avanzadas de análisis de datos.

2.1.1.3 Machine learning en epidemiología predictiva. El machine learning (ML) es un subcampo de la inteligencia artificial que desarrolla algoritmos capaces de aprender patrones a partir de datos sin ser explícitamente programados. En el contexto de la epidemiología predictiva, el ML ha demostrado capacidades superiores a los métodos estadísticos tradicionales para capturar

relaciones no lineales y complejas entre variables que determinan la dinámica de transmisión de enfermedades infecciosas.

Los modelos de series temporales como ARIMA (AutoRegressive Integrated Moving Average) representan el enfoque estadístico clásico de referencia (baseline). Sin embargo, los modelos de ensamble y las redes neuronales profundas han superado consistentemente estos enfoques cuando se incorporan variables exógenas (climáticas, socioeconómicas) con rezagos temporales. Para este proyecto se evaluarán comparativamente tres arquitecturas: Random Forest, XGBoost y redes neuronales recurrentes LSTM.

2.1.1.4 Redes neuronales recurrentes LSTM. Las Long Short-Term Memory (LSTM) son una arquitectura de redes neuronales recurrentes propuesta por Hochreiter y Schmidhuber (1997) diseñada para resolver el problema del gradiente desvaneciente en el aprendizaje de dependencias a largo plazo. Una célula LSTM incorpora tres compuertas (entrada, olvido y salida) que regulan el flujo de información a través del tiempo, permitiendo retener información relevante de períodos anteriores mientras descarta información irrelevante. Esta arquitectura resulta especialmente adecuada para modelar series temporales epidemiológicas con patrones estacionales complejos y dependencias temporales de múltiples semanas o meses.

En el contexto de predicción de dengue, los LSTM tienen la capacidad de capturar simultáneamente patrones a corto plazo (como el incremento semanal de casos) y tendencias a largo plazo (como los ciclos interanuales epidémicos), integrando eficientemente variables exógenas con rezagos temporales variables para cada predictor climático.

2.1.1.5 Modelos de ensamble: Random Forest y XGBoost. Los modelos de ensamble combinan múltiples modelos base para producir predicciones más robustas. Random Forest

(Breiman, 2001) construye múltiples árboles de decisión usando submuestras aleatorias de los datos y características, combinando sus predicciones mediante votación o promedio. XGBoost (Chen y Guestrin, 2016) implementa gradient boosting de manera optimizada, construyendo árboles secuencialmente donde cada árbol corrige los errores del anterior. Ambos algoritmos son adecuados para conjuntos de datos tabulares con múltiples características heterogéneas y ofrecen mecanismos de importancia de variables que facilitan la interpretabilidad epidemiológica.

Una ventaja adicional de estos modelos en el contexto de salud pública es su compatibilidad con métodos de explicabilidad como SHAP (SHapley Additive exPlanations, Lundberg y Lee, 2017), que permiten cuantificar la contribución de cada variable predictora a las predicciones individuales, facilitando la comunicación de resultados a tomadores de decisión.

2.1.1.6 Variables multimodales en la predicción del dengue. La literatura científica establece que la integración de variables heterogéneas de múltiples fuentes mejora significativamente la capacidad predictiva de los modelos de dengue. Las categorías de variables relevantes incluyen:

- *Variables climáticas (IDEAM):* temperatura mínima, máxima y promedio, precipitación acumulada, humedad relativa. Se ha documentado que la temperatura mínima ejerce su efecto con un rezago aproximado de 1 mes y la precipitación con 3 meses sobre la incidencia de dengue (Sang et al., 2015a).
- *Variables satelitales (Sentinel-2):* índice de vegetación de diferencia normalizada (NDVI) como proxy de condiciones ambientales favorables para criaderos del vector, temperatura superficial terrestre y detección de cuerpos de agua estancada.

- *Variables sociodemográficas (DANE)*: densidad poblacional por comuna, estratificación socioeconómica, índice de necesidades básicas insatisfechas (NBI) y cobertura de servicios públicos domiciliarios.
- *Variables epidemiológicas (SIVIGILA)*: series temporales de casos por semana epidemiológica, distribución por clasificación clínica y distribución espacial por comunas.

2.1.2 Modelos, técnicas y enfoques existentes

2.1.2.1 Modelos de series temporales tradicionales. Los modelos ARIMA y SARIMA constituyen el enfoque estadístico clásico de referencia en epidemiología. Luz et al. (2008) aplicaron series temporales ARIMA para analizar la incidencia de dengue en Río de Janeiro, estableciendo patrones estacionales y tendencias de largo plazo en un contexto latinoamericano urbano comparable a Bucaramanga. Estos modelos siguen siendo relevantes como línea base (baseline) contra la cual comparar el desempeño de los modelos ML, dado que su estructura matemática relativamente sencilla garantiza reproducibilidad e interpretabilidad.

Sin embargo, la principal limitación de los modelos de series temporales tradicionales es su incapacidad para incorporar variables exógenas de manera eficiente, especialmente cuando estas variables tienen relaciones no lineales con la variable objetivo y actúan con rezagos temporales variables. Esta limitación ha impulsado el interés por enfoques más flexibles basados en ML.

2.1.2.2 Sistemas de alerta temprana epidemiológica. Los sistemas de alerta temprana (SAT) para enfermedades infecciosas representan la aplicación operacional de los modelos predictivos en contextos de salud pública. Chretien et al. (2015b) establecen los marcos

conceptuales y lineamientos para el desarrollo de SAT de enfermedades infecciosas desde la perspectiva de la salud pública, identificando criterios clave: especificidad del umbral de alerta, anticipación suficiente para intervenciones preventivas, accesibilidad para tomadores de decisión y sostenibilidad institucional.

El sistema InfoDengue de Brasil (Codeço et al., 2018) representa uno de los ejemplos más exitosos de implementación operacional de un sistema de vigilancia en tiempo real para dengue, integrando datos epidemiológicos, climáticos y de redes sociales para nowcasting. Este caso provee una referencia concreta de las condiciones técnicas e institucionales necesarias para traducir un modelo predictivo en una herramienta de salud pública funcional.

2.1.2.3 Evaluación y validación de modelos predictivos epidemiológicos. Johansson et al. (2016) establecieron métricas y metodologías rigurosas para evaluar la calidad de predicciones epidemiológicas, comparando modelos basados en clima versus modelos estacionales para dengue en Tailandia. Sus hallazgos son relevantes para el presente proyecto porque definen criterios estándar de evaluación: MAE (Mean Absolute Error), MAPE (Mean Absolute Percentage Error), RMSE (Root Mean Square Error) y F1-Score para detección de brotes. Adicionalmente, establecen protocolos de validación temporal —train-test split cronológico y walk-forward validation— que son particularmente apropiados para series temporales epidemiológicas donde la contaminación temporal puede inflar artificialmente las métricas de desempeño.

Un indicador adicional de relevancia creciente es el MASE (Mean Absolute Scaled Error), métrica invariante a la escala de la epidemia que permite comparar modelos entrenados en ciudades de distinto tamaño. Un valor de MASE menor a 1.0 indica que el modelo ML supera a un modelo naïve estacional básico; los mejores resultados reportados para Random Forest en series de dengue

oscilan entre 0.55 y 0.60, mientras que para redes LSTM se ha reportado un MAPE cercano al 24% con horizonte de 4 semanas (Lober et al., 2024).

2.1.2.4 Paradigmas metodológicos en la predicción computacional del dengue. El análisis sistemático de la literatura publicada entre 2023 y 2025 permite identificar tres paradigmas analíticos diferenciados, cuyos enfoques son complementarios y no excluyentes:

- *Paradigma ML/Deep Learning (dominante 2024)*: modelos Random Forest, XGBoost, CatBoost y redes LSTM que aprenden patrones directamente de los datos históricos. Su fortaleza es el reconocimiento de relaciones no lineales y la integración eficiente de variables heterogéneas; su limitación es la opacidad de los mecanismos aprendidos y la dependencia de grandes volúmenes de datos etiquetados.
- *Paradigma Bayesiano espaciotemporal (emergente 2024-2025)*: modelos INLA (Integrated Nested Laplace Approximation) con efectos CAR/BYM2 y modelos ArboAlvo que incorporan estructuras espaciales explícitas entre unidades geográficas contiguas. Su ventaja es la cuantificación rigurosa de la incertidumbre y la detección de hiperfocas intraurbanas; su limitación es la mayor complejidad computacional y el requerimiento de supuestos distribucionales.
- *Paradigma mecanicista híbrido (frontera 2025)*: integración de ecuaciones diferenciales que modelan el ciclo biológico del mosquito *Aedes aegypti* con algoritmos ML para el procesamiento de variables socioeconómicas y de movilidad humana. Representa la frontera del campo, con capacidad de generalización a escenarios climáticos no observados históricamente.

La evolución cronológica de la literatura refleja una transición progresiva: la etapa 2024 estuvo dominada por ML puro con énfasis estadístico; la etapa 2025 incorporó la dimensión de

explicabilidad mediante SHAP y modelos bayesianos; y la frontera actual apunta a la convergencia mecanicista-ecológica. Este proyecto se posiciona en el primer y segundo paradigma, con diseño que facilita la extensión futura hacia el paradigma híbrido.

2.1.3 Revisión de trabajos previos relacionados

2.1.3.1 Metodología de revisión sistemática. La búsqueda bibliográfica se realizó en Web of Science (WoS) y Scopus como bases de datos primarias, complementada con PubMed/MEDLINE, Google Scholar y repositorios institucionales del Instituto Nacional de Salud (INS) de Colombia y la Organización Panamericana de la Salud (OPS). Las ecuaciones de búsqueda combinaron términos sobre la enfermedad (dengue, dengue fever, dengue hemorrhagic fever), los enfoques metodológicos (machine learning, deep learning, LSTM, random forest, XGBoost, neural network, ensemble) y los objetivos (prediction, forecast, early warning, surveillance, outbreak detection) con filtros espaciales (spatial, spatiotemporal, geographic).

Los filtros aplicados incluyeron período 2015-2025, artículos originales y revisiones, idiomas inglés y español, y categorías temáticas de salud pública, medicina tropical, enfermedades infecciosas y ciencias de la computación con inteligencia artificial. De 559 registros identificados (WoS: 247, Scopus: 312), se eliminaron duplicados resultando en 428 registros únicos, de los cuales se cribaron 428 por título y resumen, se evaluaron 100 a texto completo y se seleccionaron 15 artículos con alta relevancia para el proyecto siguiendo los criterios PRISMA.

2.1.3.2 Estudios de referencia en el contexto latinoamericano. Chen y Moraga (2025) desarrollaron modelos LSTM para predicción de dengue en todos los estados de Brasil, integrando variables climáticas con rezagos temporales de 1 a 3 meses y efectos espaciales de estados vecinos. Su estudio utilizó SHAP para selección de variables y logró mejoras significativas sobre modelos

ARIMA/SARIMA tradicionales. La relevancia directa para el presente proyecto radica en el contexto latinoamericano comparable, la metodología LSTM replicable y la disponibilidad pública de las variables climáticas utilizadas (ERA5). El horizonte de predicción de 1 a 3 meses es congruente con los objetivos de anticipación establecidos.

Sebastianelli et al. (2024) desarrollaron un modelo de ensamble reproducible combinando CatBoost, SVM y LSTM fusionados con Random Forest, para predecir brotes de dengue con 1 mes de anticipación. Integraron 31 variables incluyendo climatológicas, geofísicas (NDVI), socioeconómicas y epidemiológicas, aplicándolo exitosamente en Brasil con posterior transferencia a Perú. Con 22 citaciones en la literatura, es el estudio más validado por la comunidad científica en este campo y justifica directamente el uso de modelos ensamble sobre algoritmos individuales, además de establecer la importancia de las variables socioeconómicas y geofísicas como predictores complementarios.

Zhao et al. (2020) realizaron la comparación más cercana al contexto de este proyecto: modelos Random Forest y redes neuronales artificiales para predicción de dengue en Colombia a nivel nacional y departamental, utilizando datos del SIVIGILA junto con variables satelitales (EVI, CMORPH, MODIS) y sociodemográficas (población, índice Gini). Sus resultados establecen benchmarks de desempeño para la predicción de dengue en Colombia y validan el uso del SIVIGILA como fuente de datos para ML. Los propios autores reconocen limitaciones como la falta de datos entomológicos y la incapacidad de los modelos Random Forest para extrapolar valores fuera del rango de entrenamiento, lo que motiva la evaluación de arquitecturas adicionales como LSTM y XGBoost.

2.1.3.3 Enfoques espaciotemporales y bayesianos. Dos estudios publicados en 2025 aportan perspectivas complementarias basadas en modelos bayesianos espaciotemporales. Salim

et al. (2025) aplicaron INLA (Integrated Nested Laplace Approximation) con efectos aleatorios espaciales BYM2 para predicción espaciotemporal de dengue en los subdistritos de Yogyakarta, Indonesia, capturando hiperfocas de transmisión a nivel de barrio que los modelos de escala municipal suavizan. ArboAlvo (Alves et al., 2025) desarrolla un modelo bayesiano híbrido específicamente diseñado para enfermedades arbovirales en Natal, Brasil, demostrando que la densidad de población pobre por área ocupada puede incrementar el riesgo de brotes hasta un 20% por cada desviación estándar, e identificando el índice de positividad de huevos de *Aedes* con rezago de cuatro semanas como predictor entomológico de alto valor. Ambos estudios establecen que trabajar a escala de barrio o comuna es imperativo para capturar la heterogeneidad intraurbana del riesgo, argumento directamente aplicable al análisis por comunas en Bucaramanga.

El análisis comparativo de los seis artículos base de esta investigación permite identificar cuatro puntos de convergencia metodológica entre paradigmas: (1) efecto de retardo —existe consenso en que las variables climáticas operan con rezagos de 1 a 3 meses, correspondientes al ciclo biológico del vector—; (2) dependencia espacial —el riesgo de dengue se agrupa geográficamente y los modelos deben incorporar efectos espaciales explícitos—; (3) arquitecturas de ensamble —ningún estudio reciente depende de un único algoritmo; la combinación de modelos complementarios mejora sistemáticamente el desempeño—; y (4) datos satelitales como proxy ambiental —el NDVI y la temperatura superficial terrestre son predictores consistentes en todos los contextos analizados—. Estas convergencias fundamentan las decisiones de diseño adoptadas en el Capítulo 3 de este proyecto.

2.1.3.4 Factores de riesgo y variables predictoras. Sang et al. (2015) aportaron evidencia empírica fundamental sobre los rezagos temporales óptimos de las variables climáticas para la predicción de dengue: temperatura mínima con rezago de 1 mes y precipitación con rezago de 3

meses. Este hallazgo proporciona una base científica para el diseño de las características temporales (features engineering) del modelo propuesto. Majeed et al. (2023) extendieron la arquitectura LSTM incorporando mecanismos de atención espacial para capturar la heterogeneidad geográfica en la transmisión de dengue a nivel de distritos en Malasia, metodología avanzada aplicable al análisis por comunas en Bucaramanga. Finalmente, Colón-González et al. (2021) proyectaron la incidencia futura de dengue bajo escenarios de cambio climático en el sudeste asiático, contextualizando la importancia de considerar variables ambientales cambiantes en los modelos predictivos.

2.1.4 Identificación de vacíos o limitaciones en la literatura

El análisis sistemático de los trabajos revisados revela múltiples brechas que justifican la realización de este proyecto de investigación en el contexto específico de Bucaramanga. Se identificaron 13 brechas distribuidas en seis categorías:

2.1.4.1 Brechas de conocimiento. Ausencia de análisis específicos del dengue en Bucaramanga a nivel de comunas (B1): Aunque existen estudios a nivel nacional (Zhao et al., 2020) y comparativos entre ciudades colombianas, no se identificaron estudios que analicen los patrones espaciotemporales del dengue a nivel de comunas en Bucaramanga. Esta brecha limita la comprensión de los factores de riesgo locales y la capacidad de desarrollar modelos predictivos ajustados al contexto municipal específico.

Integración de variables de movilidad humana en contexto colombiano (B2): Los estudios revisados incorporan variables climáticas, socioeconómicas y geofísicas, pero carecen de integración sistemática de datos de movilidad humana —patrones de desplazamiento entre comunas, migración pendular— en el contexto colombiano. Sebastianelli et al. (2024) mencionan

la movilidad como factor relevante pero no la modelan directamente, mientras que Chen y Moraga (2025) solo consideran efectos espaciales entre estados, no movilidad intra-urbana.

2.1.4.2 Brechas empíricas, metodológicas y tecnológicas. Comparación empírica de algoritmos ML modernos en datos colombianos a nivel municipal (B3): Aunque Zhao et al. (2020) compararon Random Forest y redes neuronales artificiales en Colombia, no existe comparación empírica exhaustiva de algoritmos modernos — LSTM, XGBoost, CatBoost, modelos ensemble — usando datos del SIVIGILA a nivel municipal para una ciudad intermedia como Bucaramanga. Esta brecha genera incertidumbre sobre cuáles algoritmos son óptimos para las características específicas de estos datos.

Metodología de integración multimodal de datos heterogéneos para ciudades intermedias (B4): Los estudios tienden a integrar dos o tres tipos de variables, pero existe una brecha metodológica en el desarrollo de marcos robustos y reproducibles para la integración de fuentes altamente heterogéneas — satelital, movilidad, socioeconómica y epidemiológica — escalables a ciudades intermedias colombianas.

Horizontes de predicción operacionalmente útiles en contexto colombiano (B5): Existe variabilidad en los horizontes de predicción reportados (desde nowcasting hasta 3 meses), pero no está documentado qué horizonte temporal es operacionalmente más útil para las intervenciones de salud pública a nivel municipal en Colombia, dada la capacidad de respuesta real de los sistemas de salud locales.

2.1.4.3 Brechas geográficas, culturales y temporales. Transferibilidad de modelos entre contextos geográficos colombianos (B6): Aunque Sebastianelli et al. (2024) demostraron transferibilidad de modelos de Brasil a Perú, no existe evidencia sobre la transferibilidad entre

ciudades colombianas con diferentes características geográficas, climáticas y socioeconómicas. Validar esta transferibilidad tiene implicaciones directas para la escalabilidad de la solución a otros municipios de Santander.

Temporalidad de datos: brotes recientes 2023-2025 (B7): Los estudios revisados analizan datos hasta 2019-2022, sin incorporar datos de brotes recientes que pueden reflejar dinámicas epidemiológicas distintas asociadas a cambios en los serotipos circulantes del virus o en los patrones climáticos.

2.1.4.4 Brechas interdisciplinarias. Integración entre ciencia de datos y salud pública operacional (B8): Existe una desconexión significativa entre el desarrollo de modelos predictivos sofisticados y su implementación en sistemas de vigilancia epidemiológica reales a nivel municipal. La mayoría de los estudios son académicos sin implementación operacional documentada; InfoDengue de Brasil (Codeço et al., 2018) constituye una excepción relevante. En Colombia, no existe un sistema predictivo implementado operacionalmente a nivel municipal.

Déficit de colaboración interdisciplinaria en contexto colombiano (B9): Los estudios tienden a ser liderados por una sola área disciplinaria, con limitada colaboración genuina entre epidemiología, ciencia de datos y análisis espacial (GIS). Esta brecha produce modelos técnicamente sofisticados pero epidemiológicamente ingenuos, o viceversa.

2.1.4.5 Brechas éticas y sociales. Sesgo algorítmico en predicciones por estrato socioeconómico (B10): Ningún estudio revisado evalúa sesgos algorítmicos por subgrupos socioeconómicos. Los modelos ML pueden heredar y amplificar sesgos existentes en los datos de

vigilancia — subregistro diferencial por estrato, acceso desigual a servicios de salud — sin mecanismos explícitos de detección y mitigación.

Privacidad y uso de datos georreferenciados (B11): El uso de datos georreferenciados de casos (nivel de barrio o comuna) plantea consideraciones éticas de privacidad que no son abordadas explícitamente en los estudios revisados, generando vacíos metodológicos para la implementación responsable.

2.1.4.6 Brechas de política. Articulación entre predicciones y toma de decisiones (B12): Los estudios generan predicciones pero raramente establecen protocolos de respuesta específicos para autoridades de salud municipal, creando una brecha entre la capacidad predictiva del modelo y su utilidad operacional real.

Sostenibilidad institucional de sistemas predictivos municipales (B13): No existe literatura sobre modelos de sostenibilidad técnica y financiera para mantener sistemas de predicción a nivel municipal en Colombia, donde las capacidades institucionales locales son frecuentemente limitadas.

2.1.4.7 Síntesis y contribución esperada del proyecto. El presente proyecto aborda prioritariamente las brechas B1, B3, B4, B6, B7 y B8, realizando el primer análisis predictivo específico para Bucaramanga a nivel de comunas, con comparación empírica de algoritmos ML modernos, metodología reproducible de integración multimodal y un diseño orientado a la viabilidad operacional. Las brechas éticas (B10, B11) y de política (B12, B13) se abordan mediante consideraciones explícitas de equidad y factibilidad en el diseño del sistema de alerta propuesto, y se identifican como líneas prioritarias de trabajo futuro.

2.2 Marco conceptual

El dengue es una enfermedad febril aguda causada por un virus del género *Flavivirus* y transmitida por mosquitos del género *Aedes*, principalmente *Aedes aegypti*. La Organización Mundial de la Salud lo reconoce como la arbovirosis de mayor expansión mundial y, en Colombia, constituye un evento endémico de notificación obligatoria. Clínicamente se clasifica, según su gravedad, en dengue sin signos de alarma, dengue con signos de alarma y dengue grave (códigos 210, 220 y 580 del SIVIGILA, respectivamente). El ciclo de vida del vector depende estrechamente de condiciones climáticas como la temperatura, la precipitación y la humedad relativa, lo que explica el carácter estacional de la enfermedad y justifica la incorporación de variables ambientales en su predicción.

La vigilancia epidemiológica es el proceso sistemático de recolección, análisis e interpretación de datos de salud orientado a la toma de decisiones. En Colombia, su instrumento oficial es el Sistema Nacional de Vigilancia en Salud Pública (SIVIGILA), administrado por el Instituto Nacional de Salud, que registra cada caso por semana epidemiológica —unidad temporal estándar de siete días que organiza la notificación a lo largo del año— y por unidad territorial. La incidencia corresponde al número de casos nuevos en una población y un periodo determinados, mientras que un brote se entiende como el incremento inusual de casos por encima de lo esperado para una zona y una época específicas. La predicción temprana, eje de este trabajo, busca estimar ese incremento con la anticipación suficiente para activar intervenciones preventivas.

El machine learning (aprendizaje automático) es la rama de la inteligencia artificial que construye algoritmos capaces de aprender patrones a partir de los datos sin ser programados de manera explícita para cada tarea. Este proyecto se enmarca en el aprendizaje supervisado, en el que el modelo aprende a partir de ejemplos con la variable objetivo conocida —los casos

observados— para predecir valores futuros. El horizonte de predicción (h) indica cuántas semanas hacia adelante se proyecta el pronóstico (de una a cuatro, en este caso). El rezago o lag representa el desfase temporal con el que una variable, por ejemplo la temperatura, influye sobre la incidencia. La ingeniería de características (feature engineering) es el proceso de transformar los datos crudos en variables predictoras informativas, incluidos los rezagos, las medias móviles y los componentes estacionales.

Como líneas base se emplean modelos estadísticos de series temporales: el modelo naïve estacional, que repite el valor de la temporada anterior, y ARIMA, que modela la autocorrelación de la serie. Frente a ellos se evalúan tres arquitecturas de machine learning. Random Forest (Breiman, 2001) es un modelo de ensamble que combina múltiples árboles de decisión entrenados sobre submuestras aleatorias de datos y variables. XGBoost (Chen y Guestrin, 2016) implementa el gradient boosting, técnica que construye árboles de forma secuencial de modo que cada uno corrige los errores del anterior. Las redes neuronales recurrentes Long Short-Term Memory (LSTM) (Hochreiter y Schmidhuber, 1997) están diseñadas para aprender dependencias de largo plazo en secuencias mediante compuertas que regulan el flujo de información en el tiempo, lo que las hace especialmente idóneas para modelar series epidemiológicas.

La predicción se apoya en variables multimodales provenientes de cuatro fuentes. Las variables climáticas (IDEAM) comprenden temperatura, precipitación y humedad relativa; las variables satelitales incluyen el índice de vegetación de diferencia normalizada (NDVI), derivado de imágenes Sentinel-2, que actúa como proxy de las condiciones ambientales favorables al vector; y las variables sociodemográficas (DANE) incorporan la densidad poblacional y el índice de necesidades básicas insatisfechas (NBI), indicador de carencias estructurales de los hogares. La

literatura documenta que estas variables ejercen su efecto sobre la incidencia con rezagos temporales (Sang et al., 2015b), lo que sustenta su tratamiento como predictores desfasados.

La evaluación de los modelos se realiza con métricas de error —MAE, RMSE, MAPE y MASE, esta última escalada respecto del modelo ingenuo (valores inferiores a 1,0 indican un desempeño superior al de la línea base)— y con métricas de clasificación de brotes como F1 y AUC. Para evitar la contaminación temporal (data leakage), se aplica una partición temporal estricta que reserva los datos más recientes para la prueba. La prueba de Diebold-Mariano contrasta si las diferencias de desempeño entre dos modelos son estadísticamente significativas. La interpretabilidad se aborda mediante los valores SHAP (Lundberg y Lee, 2017), que cuantifican la contribución de cada variable a las predicciones. Todo el flujo de trabajo se organiza bajo el marco CRISP-DM (Cross-Industry Standard Process for Data Mining), estándar metodológico para proyectos de ciencia de datos.

3. Metodología

Este capítulo describe el enfoque metodológico adoptado para el desarrollo del proyecto de investigación aplicada, incluyendo el tipo de investigación, las fases de desarrollo, las fuentes de datos, las herramientas tecnológicas y los métodos de análisis y evaluación empleados. El diseño metodológico responde directamente a los objetivos específicos planteados en el Capítulo 1 y se fundamenta en los enfoques identificados como más efectivos en la revisión de la literatura presentada en el Capítulo 2.

3.1 Tipo y enfoque de la investigación

La presente investigación es de tipo aplicada, cuantitativa y de alcance explicativo-predictivo. Se clasifica como investigación aplicada porque tiene como propósito central resolver

un problema concreto de salud pública —la ausencia de un sistema predictivo de brotes de dengue en Bucaramanga— mediante la aplicación de conocimientos avanzados en ciencia de datos y machine learning. El enfoque es cuantitativo dado que trabaja con datos numéricos estructurados provenientes de sistemas de vigilancia epidemiológica, registros meteorológicos y fuentes demográficas, y emplea métodos estadísticos y computacionales para el análisis y validación de los modelos.

El alcance es explicativo-predictivo: explicativo porque busca identificar las variables que determinan la ocurrencia y distribución de los brotes de dengue, y predictivo porque el objetivo final es desarrollar un modelo capaz de predecir brotes con la suficiente antelación para que las autoridades sanitarias puedan implementar intervenciones preventivas. La investigación no tiene un componente experimental en el sentido clásico (no implica manipulación de variables en el mundo real), sino que aplica técnicas de modelado sobre datos observacionales históricos siguiendo un diseño retrospectivo-prospectivo.

La unidad de análisis es el caso notificado de dengue en el municipio de Bucaramanga, agregado por semana epidemiológica y por comuna, durante el período 2015 a 2022. La perspectiva metodológica es interdisciplinaria, integrando enfoques de la epidemiología, la estadística, la ciencia de datos y los sistemas inteligentes.

3.2 Descripción de las fases del proyecto

El desarrollo del proyecto se estructura en cinco fases secuenciales, diseñadas para garantizar la calidad de los datos, la rigurosidad del modelado y la utilidad práctica de los resultados. El tiempo estimado total para la ejecución completa es de veinte a veinticuatro semanas.

- *Fase 1: Exploración y caracterización epidemiológica (4-6 semanas):* esta fase tiene como objetivo comprender en profundidad los datos disponibles antes de iniciar

cualquier proceso de modelado. Las actividades principales incluyen la carga y depuración del conjunto de datos del SIVIGILA (2015-2022), el análisis exploratorio de datos (EDA) mediante estadísticas descriptivas, visualizaciones de series temporales y mapas de calor por semana epidemiológica y por comuna. Se calcularán tasas de incidencia ajustadas por población, se identificarán los años con picos epidémicos (2015 y 2019) y se caracterizarán los patrones de estacionalidad y tendencia mediante pruebas de Dickey-Fuller y descomposición STL. Se realizará también un análisis de correlación entre las variables epidemiológicas y las variables climáticas, utilizando rezagos de 1 a 12 semanas para identificar los retardos óptimos.

El entregable de esta fase es un informe de análisis exploratorio que documenta los patrones identificados, las anomalías detectadas, los rezagos climáticos óptimos y las estadísticas descriptivas completas del conjunto de datos, como insumo para el diseño de las características del modelo.

- *Fase 2: Integración y preparación de datos multimodales (4-6 semanas):* esta fase aborda el proceso de construcción del conjunto de datos integrado que alimentará los modelos predictivos, combinando las cuatro fuentes de datos identificadas: SIVIGILA (epidemiológica), IDEAM (climática), Sentinel-2 (satelital) y DANE (sociodemográfica). Las actividades incluyen la recopilación y procesamiento de datos climatológicos de las estaciones meteorológicas del IDEAM ubicadas en el área metropolitana de Bucaramanga, la descarga y procesamiento de imágenes Sentinel-2 para calcular índices NDVI y detectar cuerpos de agua estancada, y la integración de estadísticas sociodemográficas del DANE por comunas.

El proceso de ingeniería de características (feature engineering) contempla la creación de variables de rezago temporal para cada predictor climático (configuraciones de 1 a 12 semanas), cálculo de medias móviles y sumas acumuladas para suavizar series ruidosas, construcción de variables indicadoras de temporada seca y lluviosa, y codificación de la semana epidemiológica y el mes como variables cíclicas (mediante funciones seno y coseno) para preservar la continuidad temporal en los modelos. El entregable es el dataset integrado y documentado, listo para el modelado.

- *Fase 3: Desarrollo y comparación de modelos (6-8 semanas):* esta es la fase central del proyecto, en la que se desarrollan, entrenan y comparan las tres arquitecturas de machine learning seleccionadas. El proceso inicia con el desarrollo de los modelos de referencia estadísticos (ARIMA/SARIMA) que establecen la línea base de desempeño. Posteriormente se entrenan los modelos Random Forest y XGBoost con optimización de hiperparámetros mediante búsqueda aleatoria y validación cruzada temporal. Finalmente, se diseña y entrena la arquitectura LSTM con las secuencias temporales multivariadas.

Para cada modelo se aplica una estrategia de partición temporal estricta. El conjunto de entrenamiento corresponde al período 2015-2018 y el conjunto de prueba al período 2019-2020. La validación de hiperparámetros se realiza mediante validación cruzada temporal (TimeSeriesSplit) dentro del conjunto de entrenamiento, respetando el orden cronológico de las observaciones y evitando contaminación temporal. Esta configuración se adopta debido a que las series climáticas completas disponibles finalizan en diciembre de 2020, garantizando que el conjunto de prueba corresponde a datos futuros no vistos durante el entrenamiento. Se evalúan horizontes de predicción

de 1 a 4 semanas de anticipación ($h=1$, $h=2$, $h=3$, $h=4$), en línea con el alcance del proyecto aprobado y con los horizontes operacionalmente útiles para la toma de decisiones en salud pública municipal. El análisis de importancia de variables mediante permutación y SHAP se aplica a los modelos de ensamble para identificar los predictores más relevantes.

- *Fase 4: Evaluación y validación (3-4 semanas)*: esta fase se dedica a la evaluación rigurosa del desempeño de los modelos, con especial énfasis en la detección de brotes epidémicos. Se calculan las métricas MAE, MAPE y RMSE para evaluar el error de predicción en la cantidad de casos. Adicionalmente, se define un umbral de alerta de brote epidémico basado en el percentil 75 de la distribución histórica de casos semanales, y se calculan las métricas de clasificación binaria Precisión, Recall, F1-Score y AUC-ROC para evaluar la capacidad de detección de brotes. Se realiza también un análisis de la calibración de las predicciones probabilísticas para los modelos que los generen.

Se comparan los resultados entre modelos y entre horizontes de predicción para identificar la mejor arquitectura y configuración. Se realizan pruebas de significancia estadística (Diebold-Mariano) para verificar si las diferencias de desempeño son estadísticamente significativas. El entregable es la tabla comparativa de métricas y el análisis de importancia de variables.

- *Fase 5: Diseño del prototipo de sistema de alerta temprana (3-4 semanas)*: la fase final traduce los modelos desarrollados en una propuesta concreta de sistema de alerta temprana para las autoridades sanitarias de Bucaramanga. El prototipo incluye un pipeline automatizado de actualización semanal de datos y generación de predicciones,

un tablero de visualización (dashboard) con niveles de alerta codificados por color (verde, amarillo, naranja, rojo) según el nivel de riesgo predicho, y mapas de distribución espacial del riesgo por comunas con actualización semanal. El diseño contempla explícitamente los umbrales de activación de alertas, los protocolos de comunicación sugeridos a las autoridades sanitarias y las consideraciones de equidad para evitar sesgos en la asignación de recursos preventivos.

3.3 Fuentes de datos y criterios de selección

3.3.1 SIVIGILA — *Sistema Nacional de Vigilancia en Salud Pública*

La fuente primaria de datos epidemiológicos es el SIVIGILA, administrado por el Instituto Nacional de Salud (INS). Para este proyecto se utilizan las fichas de notificación de dengue correspondientes a los códigos de evento 210 (dengue sin signos de alarma), 220 (dengue con signos de alarma) y 580 (dengue grave) del municipio de Bucaramanga, período 2015-2022. El dataset contiene 10.259 registros con las siguientes variables clave: fecha de inicio de síntomas, fecha de notificación, semana epidemiológica, año, clasificación final del caso, indicador de hospitalización, barrio y comuna de residencia, sexo, edad, tipo de aseguramiento en salud y resultados de pruebas diagnósticas.

El criterio de selección de este período se basa en la disponibilidad, consistencia y completitud de los datos, asegurando la representatividad de múltiples ciclos epidémicos incluyendo los picos de 2015 y 2019. La fuente de acceso es la plataforma de Datos Abiertos Colombia (datos.gov.co) y los reportes directos del INS. Los datos son de acceso público y no contienen información personal identificable a nivel individual.

3.3.2 IDEAM — *Datos climatológicos*

El Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM) provee datos históricos de las estaciones meteorológicas ubicadas en el área metropolitana de Bucaramanga y sus alrededores. Las variables climatológicas seleccionadas son temperatura mínima, máxima y promedio diaria ($^{\circ}\text{C}$), precipitación acumulada diaria (mm) y humedad relativa promedio (%). La selección de estas variables se fundamenta en la evidencia científica sobre su relación biológica con el ciclo de vida del vector *Aedes aegypti*: la temperatura determina la velocidad de desarrollo larval y la tasa de picadura, mientras que la precipitación crea sitios de cría temporales.

Los datos diarios se agregarán a resolución semanal (promedio para temperatura y humedad, suma para precipitación) para alinearse con la resolución temporal de los datos epidemiológicos. La incorporación de rezagos temporales es un elemento central del diseño, dado que el efecto de las variables climáticas sobre la incidencia de dengue no es inmediato sino diferido en el tiempo, en correspondencia con el ciclo biológico del mosquito. La literatura especializada (Chen y Moraga, 2025; Alves et al., 2025; Salim et al., 2025) establece los siguientes rezagos óptimos con justificación biológica documentada:

- *Temperatura diurna (promedio y máxima)*: rezago de 6 a 8 semanas. Este desfase refleja el tiempo total desde la oviposición hasta el mosquito adulto infectivo: desarrollo larval (1-2 semanas), período de incubación extrínseca del virus en el mosquito (2-3 semanas) y período de transmisión activa (2-3 semanas adicionales). La temperatura mínima nocturna actúa con un rezago ligeramente menor (4-6 semanas) por su efecto sobre la supervivencia del adulto.
- *Precipitación acumulada*: rezago de 4 a 8 semanas (1 a 2 meses). La lluvia crea criaderos temporales que incrementan la densidad larval, pero el efecto sobre los casos

- tarda entre 4 y 8 semanas, equivalente al ciclo larval más el período de incubación. Un exceso de lluvia puede tener efecto negativo al arrastrar criaderos (efecto de dilución).
- *Humedad relativa*: rezago de 4 a 12 semanas (1 a 3 meses). La humedad elevada prolonga la vida del mosquito adulto y favorece la supervivencia larval; su efecto acumulado se manifiesta con retardo variable, siendo el bloque de rezagos de 1 a 3 meses el estándar óptimo para modelos de Deep Learning según la literatura.
 - *Índice NDVI satelital*: rezago de 2 a 6 semanas. La cobertura vegetal húmeda genera microhábitats favorables para criaderos; su efecto sobre la transmisión opera con menor desfase que las variables climáticas puramente físicas.

Para el tratamiento estadístico de la alta correlación entre semanas sucesivas de una misma variable (colinealidad de los lags), se emplearán dos estrategias complementarias documentadas por la literatura: los Modelos de Retardo Distribuido No Lineal (DLNM), que cuantifican simultáneamente la forma y el rezago del efecto climático, y los valores SHAP sobre el modelo entrenado, que permiten identificar los rezagos con mayor contribución marginal a las predicciones. Estas técnicas son superiores a la simple función de correlación cruzada (CCF) para manejar la multicolinealidad inherente a los bloques de lags.

3.3.3 Sentinel-2 — Datos satelitales

Las imágenes satelitales de Sentinel-2 (Agencia Espacial Europea, acceso gratuito vía Google Earth Engine) se utilizarán para calcular el Índice de Vegetación de Diferencia Normalizada (NDVI) y detectar la presencia de cuerpos de agua estancada, que constituyen hábitats potenciales para la reproducción del mosquito. El NDVI se calcula como $(NIR - RED) / (NIR + RED)$ a partir de las bandas espectrales del infrarrojo cercano y el rojo, y varía entre -1 y

1, donde valores mayores indican mayor cobertura vegetal asociada a condiciones de humedad favorables para el vector.

Se procesarán imágenes mensuales con resolución espacial de 10 metros, seleccionando escenas con cobertura de nubes inferior al 20%. Los valores de NDVI se calcularán por comunas de Bucaramanga para crear una variable espaciotemporal de riesgo ambiental. El criterio de selección de Sentinel-2 sobre otras fuentes satelitales es su resolución espacial superior, su gratuidad y la disponibilidad de series temporales desde 2015.

3.3.4 DANE — Variables sociodemográficas

El Departamento Administrativo Nacional de Estadística (DANE) proporciona datos sociodemográficos por comunas de Bucaramanga derivados del Censo 2018 y sus actualizaciones proyectadas. Las variables seleccionadas son densidad poblacional (habitantes/km²), distribución por estrato socioeconómico (1 a 6), cobertura de acueducto y alcantarillado (%), índice de necesidades básicas insatisfechas (NBI) e índice de hacinamiento. La selección de estas variables se basa en la evidencia que relaciona condiciones socioeconómicas desfavorables con mayor incidencia de dengue, dado que los estratos bajos tienen menor acceso a condiciones adecuadas de almacenamiento de agua y mayor densidad habitacional.

Dado que estas variables varían lentamente en el tiempo, se incorporarán como características estáticas por comuna en los modelos espaciotemporales. Esta decisión metodológica permite capturar la heterogeneidad espacial del riesgo sin introducir autocorrelación temporal artificial.

Tabla 1. Resumen de fuentes de datos, variables y criterios de selección

Fuente	Variables	Resolución	Criterio de selección
SIVIGILA (INS)	Casos, clasificación clínica, hospitalización, barrio/comuna	Semanal / por caso	Fuente oficial de vigilancia epidemiológica; única fuente histórica a nivel municipal
IDEAM	Temperatura (min/max/prom), precipitación, humedad relativa	Diaria → semanal	Fuente oficial meteorológica; variables con relación biológica demostrada con el vector
Sentinel-2 (ESA)	NDVI, cuerpos de agua, temperatura superficial	Mensual / 10 m espacial	Alta resolución espacial, gratuito, disponible desde 2015
DANE (Censo 2018)	Densidad pob., estrato, NBI, hacinamiento, cobertura servicios	Estática por comuna	Fuente oficial; captura heterogeneidad socioeconómica espacial del riesgo

Nota: la tabla fue elaborada teniendo como referencia las fuentes oficiales consultadas para el proyecto: SIVIGILA (Instituto Nacional de Salud [INS], 2024a), IDEAM, Sentinel-2 (Agencia Espacial Europea) y DANE. NBI: necesidades básicas insatisfechas; NDVI: índice de vegetación de diferencia normalizada.

3.4 Herramientas, técnicas y tecnologías empleadas

3.4.1 Lenguaje de programación y entorno de desarrollo

El desarrollo del proyecto se realizará en Python 3.10+, lenguaje seleccionado por su ecosistema maduro de bibliotecas para ciencia de datos, machine learning y procesamiento geoespacial, y por ser el estándar de facto en investigación de machine learning. El entorno de desarrollo principal es Jupyter Notebook / JupyterLab, que facilita el análisis exploratorio interactivo y la documentación reproducible del proceso de modelado. El control de versiones del código se gestionará mediante Git.

3.4.2 Bibliotecas de análisis de datos y machine learning

Las bibliotecas principales utilizadas son las siguientes. Para manipulación y análisis de datos: pandas (DataFrames), numpy (operaciones numéricas) y geopandas (datos espaciales). Para visualización: matplotlib, seaborn y plotly (visualizaciones interactivas). Para machine learning clásico: scikit-learn (Random Forest, métricas de evaluación, preprocesamiento) y xgboost

(XGBoost con optimización GPU si disponible). Para modelado de series temporales y redes neuronales: statsmodels (ARIMA/SARIMA) y TensorFlow/Keras (arquitecturas LSTM). Para interpretabilidad: shap (SHAP values para explicabilidad de modelos).

3.4.3 Procesamiento de datos satelitales

El procesamiento de imágenes Sentinel-2 se realizará mediante Google Earth Engine (GEE), plataforma de análisis geoespacial en la nube que proporciona acceso al catálogo completo de imágenes Sentinel-2 desde 2015. El cálculo del NDVI y la detección de cuerpos de agua se ejecutará directamente en GEE utilizando Python mediante la API de cliente (earthengine-api), exportando los resultados como series temporales de valores medios por comuna en formato CSV.

3.4.4 Visualización y dashboard

El prototipo del tablero de visualización del sistema de alerta temprana se desarrollará utilizando Streamlit, debido a su integración nativa con el ecosistema de Python empleado en el proyecto. Esta herramienta permite desplegar aplicaciones web interactivas para visualización de datos, consulta de predicciones, monitoreo de alertas epidemiológicas y exploración de variables explicativas del modelo. Los mapas de distribución espacial del riesgo por comunas se implementarán mediante Folium, integrado dentro de la aplicación Streamlit, permitiendo una visualización dinámica y actualizable del riesgo epidemiológico en el municipio de Bucaramanga.

3.5 Métodos de análisis y criterios de evaluación

3.5.1 Estrategia de partición temporal

Dado que los datos epidemiológicos constituyen una serie temporal, no se emplea una partición aleatoria —que provocaría contaminación temporal (data leakage)—, sino una partición

cronológica estricta. Debido a que las series climáticas disponibles del IDEAM finalizan en diciembre de 2020, el modelado con clima completo se acota al período 2015-2020. En consecuencia, se define un conjunto de entrenamiento con los datos de 2015 al 2018 y un conjunto de prueba con los datos de 2019 al 2020, de modo que la evaluación final se realiza sobre dos años no vistos durante el entrenamiento, uno de los cuales —2019— corresponde a un año de brote epidémico. Se dispone de un conjunto de validación mediante `TimeSeriesSplit` dentro del entrenamiento con cinco pliegues de avance temporal; en cada pliegue, los primeros años entrenan y el año siguiente valida, respetando el orden cronológico. Esta estrategia garantiza que ninguna observación del período de prueba (2019-2020) intervenga en la construcción ni en el ajuste del modelo.

3.5.2 Métricas de evaluación

Se calcularán dos conjuntos de métricas de evaluación. Para la predicción de la cantidad de casos (regresión) se utilizarán: MAE (Mean Absolute Error) — error absoluto promedio en número de casos por semana, interpretable directamente en términos epidemiológicos; MAPE (Mean Absolute Percentage Error) — error relativo, útil para comparar modelos en diferentes momentos del ciclo epidémico; RMSE (Root Mean Square Error) — penaliza con mayor severidad los errores grandes, relevante para brotes de alta magnitud; y MASE (Mean Absolute Scaled Error) — error absoluto escalado respecto a un modelo naïve estacional, métrica invariante a la escala que permite comparar resultados con estudios realizados en ciudades de distinto tamaño. Un valor de MASE menor a 1.0 indica que el modelo supera al naïve; los mejores resultados reportados en la literatura para Random Forest aplicado a series de dengue se sitúan entre 0.55 y 0.60, mientras que para redes LSTM se ha documentado un MAPE cercano al 24% con horizonte de predicción

de 4 semanas (Lober et al., 2024; Chen y Moraga, 2025). Estos valores constituyen los benchmarks de referencia para la evaluación de los modelos de este proyecto.

Para la detección de brotes (clasificación binaria, usando el umbral del percentil 75 de la distribución histórica): Precisión — proporción de alertas correctas sobre el total de alertas generadas; Recall (Sensibilidad) — proporción de brotes reales detectados; F1-Score — media armónica de precisión y recall, métrica de síntesis para sistemas de alerta; AUC-ROC — área bajo la curva ROC, independiente del umbral de clasificación. En el contexto de un sistema de alerta de salud pública, el Recall tiene mayor importancia relativa que la Precisión, dado que las consecuencias de no detectar un brote (falso negativo) son más graves que una alerta falsa.

3.5.3 Análisis de importancia de variables

La identificación de las variables predictoras más relevantes se realizará mediante tres enfoques complementarios. Primero, la importancia de variables por permutación (Permutation Feature Importance), disponible en scikit-learn, que mide el incremento en el error del modelo cuando los valores de cada variable se alteran aleatoriamente, rompiendo su relación con la variable objetivo. Segundo, los valores SHAP (SHapley Additive exPlanations), que asignan a cada variable una contribución marginal a cada predicción individual, permitiendo visualizar tanto la importancia global como el efecto direccional de cada predictor. Tercero, el análisis de correlación cruzada con rezagos entre variables climáticas y casos de dengue, como complemento interpretativo de los modelos ML.

3.5.4 Estrategia diferenciada de integración de rezagos según arquitectura

La forma en que se integran los rezagos temporales varía según la arquitectura del algoritmo, y esta diferencia es metodológicamente relevante para garantizar la solidez de los resultados:

- *Modelos ML tabulares (Random Forest, XGBoost)*: la literatura recomienda utilizar un bloque corto de lags explícitos — entre 4 y 6 rezagos por variable —, seleccionados mediante SHAP o importancia por permutación sobre el conjunto de validación, para evitar el sobreajuste por alta dimensionalidad. El estándar documentado (Lober et al., 2024) utiliza exactamente 5 rezagos para estos modelos. La multicolinealidad entre semanas sucesivas se controla mediante la regularización L1/L2 propia de XGBoost y mediante la selección aleatoria de características de Random Forest.
- *Redes LSTM*: la arquitectura recurrente incorpora la dependencia temporal de forma implícita mediante la ventana deslizante de 12 semanas de historia. Para las variables climáticas exógenas se aplica previamente un filtro de selección basado en SHAP o análisis de componentes principales (PLS) para aislar los rezagos de 1 a 3 meses con mayor aporte predictivo, antes de introducirlos como variables de entrada al tensor tridimensional. Esto previene que la red colapse por saturación de información colineal.

Esta estrategia diferenciada garantiza que cada modelo opere con el bloque de rezagos óptimo para su arquitectura, produciendo una comparación más justa y resultados más robustos en el conjunto de prueba 2022.

3.5.5 Comparación estadística entre modelos

Para determinar si las diferencias de desempeño entre modelos son estadísticamente significativas o pueden atribuirse al azar, se aplicará la prueba de Diebold-Mariano (DM) de la biblioteca statsmodels. Esta prueba compara los errores de predicción de dos modelos y produce

un estadístico de prueba con p-valor, permitiendo rechazar o no la hipótesis nula de igual desempeño predictivo. Se utilizará un nivel de significancia de 0,05. Los resultados se reportarán como matrices de comparación por pares entre todos los modelos evaluados.

3.5.6 Consideraciones éticas y de equidad

Dado que los datos del SIVIGILA incluyen información georreferenciada (nivel de comuna) y de estrato socioeconómico, se realizará un análisis de equidad para verificar que el modelo no discrimine sistemáticamente en su precisión predictiva entre comunas de diferentes estratos socioeconómicos. Se calcularán las métricas de desempeño desagregadas por quintil de NBI para detectar posibles sesgos algorítmicos. Los datos utilizados son agregados y de acceso público, no contienen información personal identificable y su uso con fines de investigación en salud pública es compatible con las normas colombianas vigentes.

3.6 Cronograma de actividades

Tabla 2. Cronograma estimado de actividades del proyecto de grado

Fase / Actividad	Mes 1-2	Mes 3-4	Mes 5-6	Mes 7-8	Créditos
Fase 1: EDA y caracterización	X				
Fase 2: Integración de datos	X	X			
Fase 3: Desarrollo de modelos		X	X		
Fase 4: Evaluación y validación			X		
Fase 5: Prototipo SAT			X	X	
Redacción del informe final			X	X	
Revisiones y entrega final				X	

Nota: la tabla fue elaborada teniendo como referencia la planeación académica del proyecto de grado y la asignación de créditos de la Maestría en Análisis de Datos y Sistemas Inteligentes.

4. Desarrollo de la solución

Este capítulo describe el diseño y la implementación de la solución desarrollada para abordar el problema de la predicción temprana de brotes de dengue en Bucaramanga. Se presenta

la arquitectura general del sistema, el diseño técnico de los modelos de machine learning seleccionados, el proceso de implementación con sus herramientas y procedimientos, y la descripción del prototipo de sistema de alerta temprana resultante. Todo el desarrollo se articula directamente con los tres objetivos específicos planteados en el Capítulo 1 y con el marco metodológico definido en el Capítulo 3.

4.1 Arquitectura general de la solución

La solución propuesta se concibe como un pipeline de datos end-to-end que integra múltiples fuentes de información heterogéneas para producir predicciones semanales de incidencia de dengue a nivel de comunas en Bucaramanga, y traducir esas predicciones en niveles de alerta accionables para las autoridades sanitarias. La arquitectura se organiza en cinco capas funcionales secuenciales, como se describe a continuación.

4.1.1 Capa 1: recolección e integración de datos

Esta capa centraliza la ingesta de datos provenientes de cuatro fuentes primarias. Primero, el SIVIGILA (Sistema Nacional de Vigilancia en Salud Pública) proporciona los registros históricos de casos de dengue (códigos de evento 210, 220 y 580) para Bucaramanga en el período 2015-2022, con variables demográficas, clínicas, temporales y georreferenciadas a nivel de barrio y comuna. Segundo, el IDEAM aporta series históricas de las estaciones meteorológicas del área metropolitana de Bucaramanga: temperatura mínima, máxima y promedio diaria (°C), precipitación acumulada (mm) y humedad relativa (%). Tercero, el DANE suministra variables sociodemográficas del Censo 2018 desagregadas por comunas: densidad poblacional, distribución por estrato, índice de necesidades básicas insatisfechas (NBI) y cobertura de servicios públicos.

Cuarto, Google Earth Engine provee índices NDVI calculados a partir de imágenes Sentinel-2 a resolución de 10 m, agregados mensualmente por polígono de cada comuna.

4.1.2 Capa 2: *preprocesamiento y homologación*

En esta capa se estandariza la resolución temporal y espacial de todas las fuentes. Los registros individuales del SIVIGILA se agregan semanalmente (semana epidemiológica colombiana) y por comuna, calculando el conteo total de casos, la proporción hospitalizada y la distribución por clasificación clínica. Los datos climatológicos diarios del IDEAM se agregan a resolución semanal: promedio para temperatura y humedad, y suma acumulada para precipitación. Las variables sociodemográficas del DANE se mantienen estáticas por comuna (varían lentamente en el tiempo) y se replican para cada semana del período de análisis. El NDVI satelital, disponible en resolución mensual, se interpola linealmente a resolución semanal. Este proceso genera un dataset panel con las dimensiones de los años 2015 al 2022, que son aproximadamente 8 años $\times$ 52 semanas = 416 semanas de observaciones.

4.1.3 Capa 3: *ingeniería de características (feature engineering)*

Esta capa construye las variables predictoras que alimentan los modelos de machine learning a partir de los datos homologados. Se definen tres grupos de características:

- *Características temporales rezagadas:* para cada variable climática se calculan retardos de 1 a 12 semanas (lags), con base en la justificación biológica documentada en la literatura (Sang et al., 2015a): temperatura diurna con rezago de 6 a 8 semanas (ciclo larval + incubación extrínseca del virus), precipitación con rezago de 4 a 8 semanas (tiempo de maduración de criaderos), humedad relativa con rezago de 4 a 12 semanas (efecto sobre supervivencia del adulto) y NDVI con rezago de 2 a 6 semanas.

Se conservan los rezagos con mayor contribución mediante valores SHAP calculados sobre los pliegues de validación de la validación cruzada temporal (TimeSeriesSplit) del conjunto de entrenamiento 2015-2018, dado que no existe un conjunto independiente correspondiente a 2021, garantizando que la selección esté guiada por el modelo y no por supuestos a priori.

- *Características de tendencia y estacionalidad:* medias móviles de 4, 8 y 12 semanas para suavizar las series ruidosas; indicadores binarios de temporada de lluvias (primer y segundo semestre); codificación cíclica de la semana epidemiológica mediante funciones seno y coseno para preservar la continuidad entre la semana 52 y la semana 1 del año siguiente.
- *Características espaciales estáticas:* densidad poblacional por km², proporción de hogares en estrato 1 y 2, índice NBI comunal y cobertura de acueducto y alcantarillado. Estas variables capturan la heterogeneidad espacial del riesgo de dengue entre comunas.

4.1.4 Capa 4: modelado y evaluación comparativa

Esta capa implementa los tres algoritmos de machine learning definidos en el Objetivo Específico 3: Random Forest (RF), XGBoost y LSTM. Adicionalmente, se entrenan dos modelos de referencia estadísticos ARIMA y Naïve, como línea base para la comparación de desempeño. La estrategia de partición temporal estricta garantiza que no existe contaminación de información futura durante el entrenamiento. Debido a que las series climáticas completas del IDEAM finalizan en diciembre de 2020, se utilizó el período 2015-2018 para el entrenamiento y el período 2019-2020 para la evaluación en prueba. La selección de hiperparámetros se realizó mediante validación cruzada temporal (TimeSeriesSplit) dentro del conjunto de entrenamiento, respetando en todo

momento el orden cronológico de los datos. Se evalúan horizontes de predicción de 1 a 4 semanas hacia adelante ($h=1$, $h=2$, $h=3$, $h=4$), alineados con el alcance del proyecto aprobado. La selección del modelo final se basa en la comparación de métricas MAE, RMSE y F1-Score para detección de brotes.

4.1.5 Capa 5: sistema de alerta y visualización

La capa de salida traduce las predicciones numéricas del modelo ganador en niveles de alerta epidemiológica semaforizados para las 17 comunas de Bucaramanga. El umbral de alerta se define con base en el percentil 75 de la distribución histórica de casos semanales por comuna, siguiendo la metodología de sistemas de alerta temprana para dengue documentada en la literatura (Chretien et al., 2015b; Codeço et al., 2018). Los resultados se visualizan en un tablero (dashboard) actualizable semanalmente que muestra las predicciones por comuna, la tendencia histórica reciente y el nivel de alerta activo.

La capa de visualización y entrega del sistema DengueAlert-BGA se implementa en Streamlit, una biblioteca de Python que permite desplegar aplicaciones web interactivas directamente desde el entorno de análisis, lo que garantiza la coherencia con el resto del *pipeline* (pandas, scikit-learn, XGBoost) y evita dependencias de licenciamiento. Power BI se descarta como salida principal y podría emplearse, de forma complementaria, para el consumo institucional de reportes por parte de la Secretaría de Salud.

Tabla 3. Resumen de las cinco capas funcionales de la arquitectura de la solución

Capa	Función principal	Tecnologías / Fuentes
1. Recolección	Ingesta de datos de cuatro fuentes heterogéneas	SIVIGILA, IDEAM, DANE, Google Earth Engine
2. Preprocesamiento	Homologación temporal (semanal) y espacial (comunas)	pandas, numpy, geopandas
3. Feature Engineering	Lags, medias móviles, codificación cíclica, variables espaciales	scikit-learn, pandas

Capa	Función principal	Tecnologías / Fuentes
4. Modelado	Entrenamiento y evaluación de RF, XGBoost, LSTM, ARIMA	scikit-learn, xgboost, TensorFlow/Keras, statsmodels
5. Alerta / Salida	Umbral de alerta, semáforo por comunas, dashboard	Streamlit

Nota: la tabla fue elaborada teniendo como referencia el diseño de la arquitectura de la solución descrito en este capítulo.

4.2 Diseño del modelo, sistema o metodología

4.2.1 Diseño del pipeline de datos

El pipeline de datos sigue un flujo lineal desde las fuentes crudas hasta el dataset listo para modelado. La figura conceptual de la arquitectura de datos muestra la siguiente secuencia de transformaciones:

Datos brutos → Limpieza → Agregación semanal/comunal → Integración multimodal
→ Feature engineering → Dataset final (semana × comuna × variables)

El dataset final está compuesto por observaciones semanales agregadas por comuna para el período analizado, generando una estructura panel semana-comuna, con un total aproximado de 45 a 55 columnas entre la variable objetivo (casos_dengue_semana) y las variables predictoras (climáticas con lags, sociodemográficas, temporales y NDVI). La variable objetivo para la tarea de regresión es el número de casos de dengue en la semana $t+h$, donde h es el horizonte de predicción (1, 2, 3 o 4 semanas). Para la tarea de clasificación de brotes, la variable objetivo es una indicadora binaria: 1 si los casos de la semana $t+h$ superan el percentil 75 de la distribución histórica de esa misma comuna, 0 en caso contrario.

4.2.2 Diseño de los modelos de machine learning

A continuación se describe el diseño específico de cada algoritmo seleccionado:

Random Forest (RF): El modelo Random Forest se configura como regresor para la predicción del número de casos y como clasificador para la detección de brotes. Los hiperparámetros clave y sus rangos de búsqueda son: número de árboles (n_estimators): 100 a 500; profundidad máxima (max_depth): 10 a 40; número mínimo de muestras para división (min_samples_split): 5 a 30; número de características evaluadas por árbol (max_features): "sqrt" y "log2". La selección de hiperparámetros se realiza mediante RandomizedSearchCV con validación cruzada temporal (TimeSeriesSplit, 5 pliegues) sobre el conjunto de entrenamiento 2015-2018, validación mediante TimeSeriesSplit y prueba 2019-2020. La importancia de variables se extrae mediante el método de importancia por permutación (Permutation Feature Importance) y se complementa con valores SHAP para garantizar la interpretabilidad (Lundberg y Lee, 2017).

XGBoost (Extreme Gradient Boosting): XGBoost se implementa mediante la biblioteca xgboost de Python. Los hiperparámetros de diseño incluyen: tasa de aprendizaje (learning_rate): 0.01 a 0.1; profundidad máxima de los árboles (max_depth): 3 a 10; fracción de submuestreo de filas (subsample): 0.6 a 1.0; regularización L1 (alpha): 0 a 1; regularización L2 (lambda): 1 a 10; número de estimadores (n_estimators): con early stopping de 50 rondas utilizando los pliegues de validación generados mediante TimeSeriesSplit dentro del conjunto de entrenamiento. XGBoost es especialmente adecuado para datos tabulares con características numéricas heterogéneas y maneja de manera nativa los valores faltantes que puedan existir en los datos climatológicos (Chen y Guestrin, 2016).

Red Neuronal Recurrente LSTM: La arquitectura LSTM se diseña para series temporales multivariadas. A diferencia de RF y XGBoost, que trabajan con el dataset en formato tabular plano, el LSTM recibe los datos como tensores tridimensionales de forma (muestras, pasos temporales,

características), donde `pasos_temporales = 12` semanas de historia. La arquitectura de la red diseñada es la siguiente:

- *Capa LSTM 1*: 128 unidades, `return_sequences=True`, activación tanh, regularización dropout 0.2
- *Capa LSTM 2*: 64 unidades, `return_sequences=False`, activación tanh, regularización dropout 0.2
- *Capa Dense 1*: 32 neuronas, activación ReLU
- *Capa Dense de salida*: 1 neurona, activación lineal (regresión) o sigmoid (clasificación de brote)

El optimizador es Adam con tasa de aprendizaje inicial de 0.001 y reducción automática (`ReduceLROnPlateau`) si la pérdida de validación no mejora en 5 épocas. La función de pérdida es el error cuadrático medio (MSE) para regresión. El entrenamiento se realiza por un máximo de 150 épocas con early stopping (paciencia de 15 épocas), utilizando validación cruzada temporal dentro del conjunto de entrenamiento para monitorear la convergencia y evitar el sobreajuste (Johansson et al., 2016).

4.2.3 Diseño de la estrategia de evaluación comparativa

La evaluación comparativa entre los cinco modelos (Naïve estacional, ARIMA, Random Forest, XGBoost, LSTM) sigue una estrategia unificada que garantiza comparaciones justas:

- *Mismo conjunto de prueba*: datos de 2019-2020 (53 semanas $\times$ 17 comunas).
- *Mismos horizontes de predicción*: 1 a 4 semanas ($h=1$, $h=2$, $h=3$, $h=4$), evaluados de forma independiente para cada modelo.
- *Mismas métricas de regresión*: MAE, RMSE, MAPE y MASE calculadas sobre el conjunto de prueba 2019-2020. El MASE (Mean Absolute Scaled Error) se incluye

como métrica invariante a escala para permitir comparación con benchmarks internacionales: MASE < 1.0 indica superación del modelo naïve estacional; valores objetivo de referencia: MASE $\approx$ 0.55–0.60 para RF/XGBoost y MAPE $\approx$ 24% para LSTM a 4 semanas (Lober et al., 2024; Chen y Moraga, 2025).

- *Mismas métricas de clasificación de brotes:* precisión, Recall, F1-Score y AUC-ROC usando el umbral del percentil 75 histórico por comuna.
- *Prueba de significancia estadística:* diebold-Mariano (DM) bilateral entre pares de modelos para verificar si las diferencias de desempeño son estadísticamente significativas ($\alpha = 0.05$).

4.3 Implementación técnica

4.3.1 Entorno tecnológico

La implementación se realiza en Python 3.10 bajo el entorno Jupyter Notebook / JupyterLab, que permite documentar el proceso de manera reproducible e interactiva. El control de versiones del código se gestiona con Git. La Tabla 4 resume las bibliotecas utilizadas y su función específica dentro del pipeline.

Tabla 4 Bibliotecas de Python utilizadas en la implementación del pipeline

Biblioteca	Versión	Función en el proyecto
Pandas	≥ 2.0	Manipulación de DataFrames, agregaciones temporales y espaciales
Numpy	≥ 1.24	Operaciones numéricas, álgebra lineal, creación de tensores
Geopandas	≥ 0.13	Manejo de datos geoespaciales de comunas (shapefiles)
scikit-learn	≥ 1.3	Random Forest, preprocesamiento, métricas, TimeSeriesSplit

Biblioteca	Versión	Función en el proyecto
Xgboost	≥ 1.7	Implementación optimizada de XGBoost con early stopping
TensorFlow	/ Keras ≥ 2.12	Diseño y entrenamiento de la red LSTM
Statsmodels	≥ 0.14	ARIMA/SARIMA como modelos baseline, prueba Diebold-Mariano
Shap	≥ 0.42	Valores SHAP para interpretabilidad de modelos RF y XGBoost
Matplotlib	/ seaborn	Visualización de series temporales, distribuciones y métricas
earthengine-api		Acceso a Google Earth Engine para descarga de NDVI (Sentinel-2)

Nota: la tabla fue elaborada teniendo como referencia la documentación oficial de cada una de las bibliotecas empleadas en la implementación del pipeline.

4.3.2 Proceso de preprocesamiento e integración de datos

El proceso de preprocesamiento e integración de datos se desarrolló mediante un flujo automatizado implementado en Python. Inicialmente se realizó la carga, depuración y filtrado de los registros del SIVIGILA correspondientes al municipio de Bucaramanga. Posteriormente se efectuó la transformación de fechas, la agregación semanal por comuna y la integración con las variables climatológicas provenientes del IDEAM. Como resultado se obtuvo una base de datos unificada y estructurada para las fases posteriores de ingeniería de características y modelado predictivo. El código utilizado para el preprocesamiento e integración de datos se presenta en el Apéndice A.1.

Inicio → Carga de datos SIVIGILA → Filtrado municipio Bucaramanga → Transformación de fechas → Agregación semanal por comuna → Carga datos IDEAM → Integración de fuentes → Dataset consolidado → Fin

4.3.3 Implementación del feature engineering

La construcción de variables predictoras se realizó mediante procedimientos automatizados de ingeniería de características. Se generaron variables rezagadas para las series climáticas, medias móviles de diferentes ventanas temporales y variables temporales cíclicas basadas en la semana epidemiológica. Estas transformaciones permitieron capturar relaciones temporales complejas y patrones estacionales relevantes para la predicción del dengue. La implementación detallada de la ingeniería de características se presenta en el Apéndice A.2.

Variables climáticas → Creación de rezagos → Medias móviles → Variables estacionales
→ Codificación seno-coseno → Matriz final de características

4.3.4 Implementación de los modelos de machine learning

La implementación de los modelos Random Forest, XGBoost y LSTM se desarrolló siguiendo la estrategia de partición temporal definida en la metodología. Cada algoritmo fue entrenado utilizando exclusivamente información histórica disponible al momento de la predicción y sometido a procesos de ajuste de hiperparámetros mediante validación temporal. Para garantizar la reproducibilidad, todos los experimentos fueron ejecutados utilizando una semilla aleatoria fija. El código correspondiente a la implementación de los modelos XGBoost y LSTM se presenta en los Apéndices A.3 y A.4, respectivamente.

Dataset preparado → Partición temporal → Entrenamiento → Ajuste de hiperparámetros
→ Evaluación → Modelo final

4.3.5 Implementación de métricas y comparación de modelos

La evaluación del desempeño se realizó mediante métricas de regresión (MAE, RMSE, MAPE y MASE) y métricas de clasificación de brotes (F1 y AUC). Posteriormente se efectuó una comparación estadística entre modelos mediante la prueba de Diebold-Mariano, con el fin de

determinar si las diferencias observadas correspondían a mejoras reales de capacidad predictiva. La implementación computacional de las métricas de evaluación y comparación de modelos se presenta en el Apéndice A.5.

Predicciones → Cálculo de métricas → Comparación modelos → Prueba Diebold-Mariano
→ Selección modelo óptimo

4.4 Descripción del prototipo o sistema desarrollado

4.4.1 Descripción general del prototipo

El prototipo desarrollado es un sistema de predicción y alerta temprana de brotes de dengue para el municipio de Bucaramanga, denominado DengueAlert-BGA. El sistema está compuesto por tres módulos funcionales interconectados: el módulo de actualización de datos, el módulo de predicción y el módulo de visualización y alerta. En su versión prototipo, el sistema opera con datos históricos del período 2015-2022 y simula el ciclo de actualización semanal que tendría en un entorno operacional real.

4.4.2 Módulo 1: actualización de datos

Este módulo ejecuta semanalmente el pipeline de preprocesamiento descrito en la sección 4.3: carga los nuevos registros del SIVIGILA correspondientes a la semana epidemiológica más reciente, descarga los datos meteorológicos actualizados del IDEAM para las estaciones de Bucaramanga, actualiza los índices NDVI cuando hay nuevas imágenes Sentinel-2 disponibles (frecuencia mensual), y regenera el dataset panel homologado incorporando los nuevos datos. El resultado es un archivo de datos actualizado listo para ser procesado por el módulo de predicción.

4.4.3 Módulo 2: predicción con el modelo seleccionado

Aunque ARIMA obtuvo el menor error estadístico, se seleccionó XGBoost como modelo del sistema debido a su capacidad de incorporar variables exógenas y proporcionar interpretabilidad mediante SHAP. Este módulo validado en el conjunto de prueba 2019-2020 (determinado en el Capítulo 5) para generar predicciones de la incidencia de dengue en las 17 comunas de Bucaramanga para las próximas 1 a 4 semanas. Las predicciones se generan en forma de distribuciones de probabilidad cuando el modelo lo permite (LSTM con dropout activado durante la inferencia — técnica Monte Carlo Dropout), o como valores puntuales con intervalos de confianza estimados por bootstrap (Random Forest, XGBoost).

Para cada comuna y cada horizonte de predicción, el módulo calcula adicionalmente la probabilidad de que el número predicho de casos supere el umbral de brote definido como el percentil 75 de la distribución histórica comunal. Esta probabilidad es el insumo principal para la clasificación en niveles de alerta.

INICIO Módulo de Predicción

1. Cargar el modelo XGBoost entrenado (modelo.pkl)
2. Cargar los datos más recientes por semana epidemiológica y comuna
3. Construir las variables predictoras (features):
 - rezagos de casos (lag 1 a 8 semanas)
 - medias móviles (ventana de 4 semanas)
 - variables climáticas rezagadas (temperatura, precipitación, humedad)
4. PARA cada horizonte h de 1 a 4 semanas HACER:

predicción[h] = modelo.predecir(features)
5. Clasificar el nivel de riesgo (semáforo):

SI predicción supera el umbral de brote -> ALERTA ROJA

SI predicción en rango intermedio- > ALERTA AMARILLA

EN OTRO CASO- > VERDE

6. Calcular los valores SHAP de la predicción (explicación de la alerta)
7. Enviar resultados a la capa de visualización (Streamlit)

FIN

4.4.4 Módulo 3: sistema de semáforo y visualización

El módulo de alerta clasifica el riesgo de cada comuna en cuatro niveles usando los umbrales definidos a continuación, inspirados en los protocolos de alertas epidemiológicas del Instituto Nacional de Salud (Instituto Nacional de Salud [INS], 2024b):

Tabla 5. Niveles de alerta del sistema DengueAlert-BGA y acciones asociadas

Nivel	Color	Umbral (prob. de brote)	Interpretación y acción sugerida
1 — Bajo	Verde	< 25%	Situación controlada. Vigilancia epidemiológica rutinaria.
2 — Moderado	Amarillo	25% – 50%	Riesgo emergente. Reforzar vigilancia y comunicación comunitaria.
3 — Alto	Naranja	50% – 75%	Riesgo elevado. Activar jornadas de control vectorial focalizadas.
4 — Crítico	Rojo	> 75%	Brote inminente o en curso. Activación del plan de contingencia municipal.

Nota: la tabla fue elaborada teniendo como referencia los protocolos de alerta epidemiológica del Instituto Nacional de Salud (Instituto Nacional de Salud [INS], 2024b) y los umbrales definidos para el prototipo DengueAlert-BGA.

El tablero de visualización del prototipo se implementa mediante Streamlit con conexión a los archivos de datos actualizados semanalmente. El informe incluye: (a) mapa de calor de Bucaramanga por comunas coloreado según el nivel de alerta vigente, (b) serie temporal histórica y predicción para las próximas 4 semanas con bandas de confianza para cada comuna seleccionada, (c) panel de indicadores clave: número de casos totales de las últimas 4 semanas, tasa de

hospitalización y clasificación actual de riesgo, y (d) tabla de prioridades de intervención ordenada de mayor a menor probabilidad de brote.

4.4.5 Alcance y limitaciones del prototipo

El prototipo DengueAlert-BGA es funcional para el análisis retrospectivo y la simulación del ciclo de alerta usando datos históricos 2015-2022. Sus limitaciones actuales incluyen: (a) la actualización semanal de datos requiere conexión manual a las fuentes del SIVIGILA e IDEAM, ya que no existe una API pública automatizada; (b) los datos de NDVI satelital tienen frecuencia mensual, lo que introduce cierta imprecisión en la resolución semanal; y (c) el prototipo no está integrado en los sistemas de información del municipio de Bucaramanga. Estas limitaciones se documentan como recomendaciones de trabajo futuro en el Capítulo 6 y no comprometen la validez científica de los modelos predictivos evaluados.

5. Resultados y validación

Este capítulo expone los resultados del proyecto y los somete a discusión a la luz de los objetivos planteados y de la literatura revisada en el Capítulo 2. La presentación sigue la secuencia lógica de los tres objetivos específicos: primero se describen los patrones temporales de la enfermedad y su vínculo con el clima; luego se examina la distribución del riesgo entre las comunas; y finalmente se reporta el desempeño de los modelos predictivos. Todos los valores provienen de la ejecución reproducible del código desarrollado en Python, con la semilla aleatoria fijada en 42 para garantizar que los resultados puedan replicarse.

Conviene precisar una decisión metodológica que afecta a todo el capítulo. El modelado se realizó sobre la ventana 2015–2020, único intervalo en el que las tres variables climáticas —temperatura, humedad y precipitación— están disponibles de manera continua en las estaciones

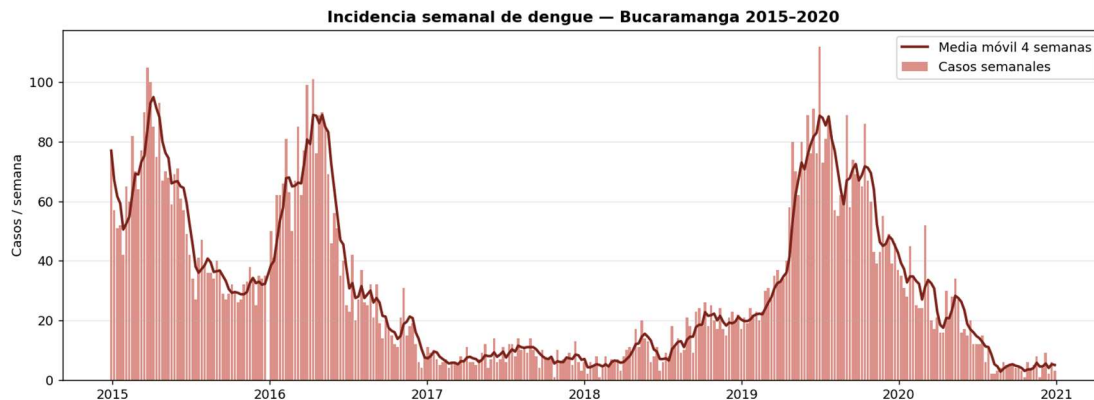
del IDEAM, pues las series de temperatura y humedad terminan el 31 de diciembre de 2020. Por esta razón, la partición temporal anunciada en el Capítulo 4 se ajustó a un conjunto de entrenamiento de 2015 a 2018 y un conjunto de prueba de 2019 a 2020. Lejos de ser una limitación menor, esta elección fortalece la evaluación: el período de prueba contiene la epidemia más intensa de la serie (2019), de modo que los modelos se ponen a prueba precisamente frente al escenario que más importa anticipar.

5.1 Resultados experimentales o de aplicación

5.1.1 *Patrones temporales de la incidencia (objetivo específico 1)*

La serie semanal de casos confirma que el dengue en Bucaramanga no se distribuye de manera uniforme en el tiempo, sino en oleadas epidémicas separadas por períodos de baja transmisión. Entre 2015 y 2020 se reconocen tres ondas mayores —2015, 2016 y 2019—, esta última con un máximo cercano a 110 casos en una sola semana epidemiológica (Figura 1). Los valles de 2017 y 2018, con incidencias muy bajas, reflejan la dinámica interepidémica característica de las arbovirosis, en la que el agotamiento temporal de la población susceptible y la circulación de distintos serotipos modulan la aparición de nuevos brotes. Este comportamiento cíclico coincide con lo descrito por Luz et al. (2008) para Río de Janeiro mediante modelos de series temporales, y respalda la pertinencia de tratar la incidencia como un proceso estacional con memoria.

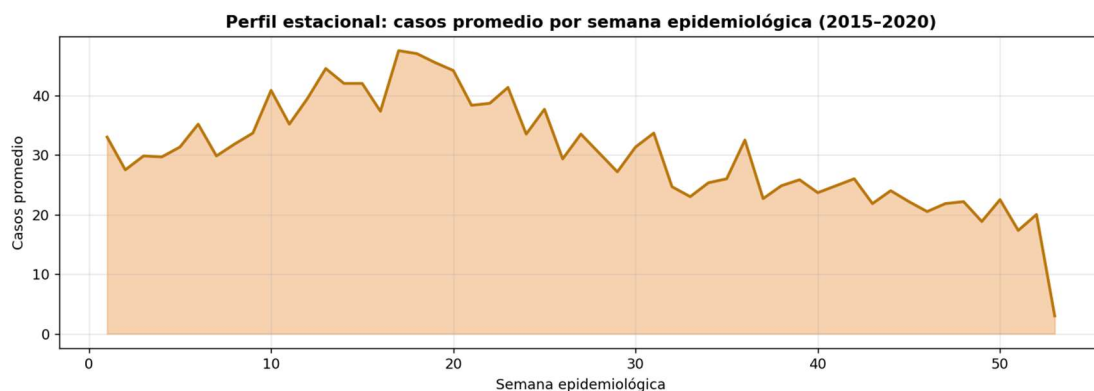
Figura 1. Incidencia semanal de dengue y media móvil de cuatro semanas (Bucaramanga, 2015–2020).



Nota: la figura fue elaborada teniendo como referencia los casos de dengue notificados al SIVIGILA para Bucaramanga entre 2015 y 2020 (Instituto Nacional de Salud [INS], 2024a).

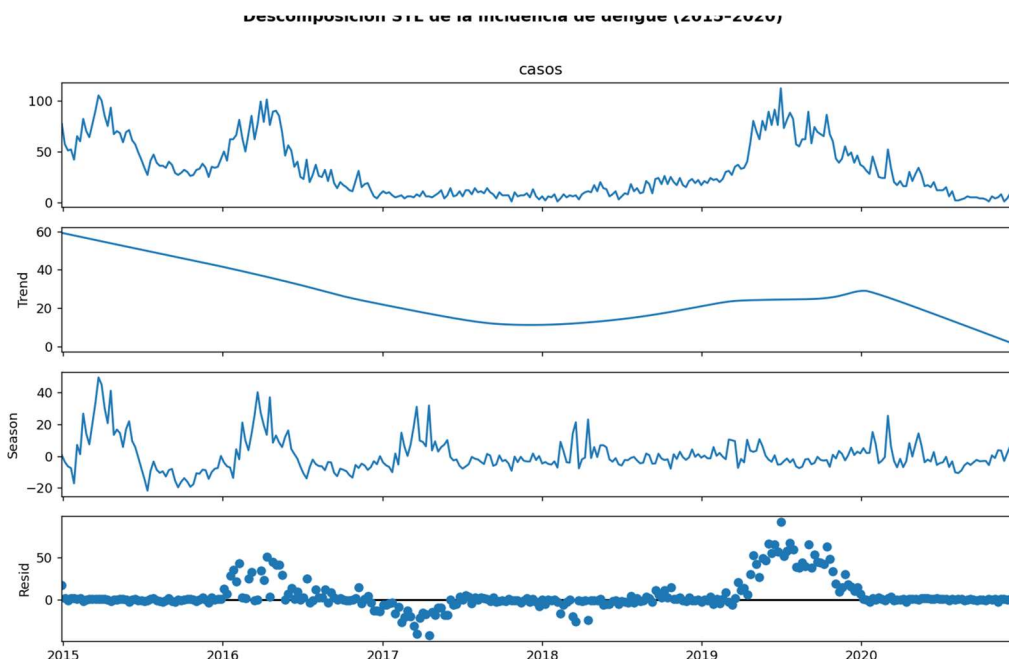
El perfil estacional promedio (Figura 2) muestra que la carga de casos tiende a aumentar en el primer y segundo trimestre del año, mientras que la descomposición de la serie mediante el procedimiento STL (Figura 3) separa de forma nítida una componente estacional anual estable de una tendencia interanual dominada por las ondas epidémicas. La persistencia de la componente estacional es relevante en términos prácticos, porque implica que una parte sustancial de la variación es predecible a partir del calendario epidemiológico, algo que más adelante se aprovechó mediante una codificación cíclica de la semana.

Figura 2. Casos promedio por semana epidemiológica (2015–2020).



Nota: la figura fue elaborada teniendo como referencia los casos de dengue notificados al SIVIGILA para Bucaramanga entre 2015 y 2020 (Instituto Nacional de Salud [INS], 2024a).

Figura 3. Descomposición STL de la serie de incidencia (2015–2020).

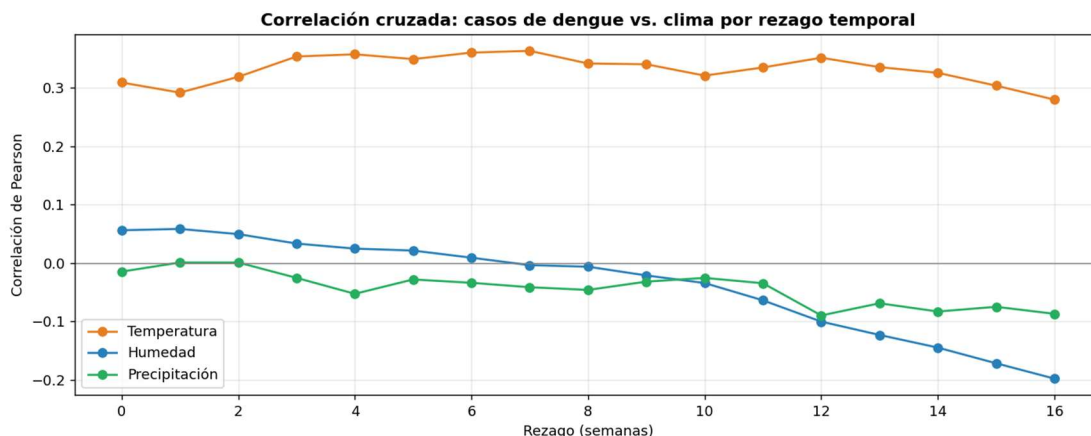


Nota: la figura fue elaborada teniendo como referencia la descomposición STL de la serie semanal de casos del SIVIGILA, 2015–2020 (Instituto Nacional de Salud [INS], 2024a).

El análisis de correlación cruzada entre los casos y cada variable climática, evaluado para rezagos de 0 a 16 semanas (Figura 4), permitió cuantificar el desfase con el que el clima antecede a la transmisión. La temperatura resultó la variable de mayor asociación, con un coeficiente máximo de $r = 0,36$ a un rezago de siete semanas; la humedad mostró una relación inversa más débil hacia las 16 semanas ($r = -0,20$) y la precipitación una señal todavía menor. Estos desfases son coherentes con la biología del vector *Aedes aegypti*, cuyo ciclo de desarrollo y la posterior incubación del virus introducen un retardo de semanas entre las condiciones ambientales favorables y el incremento de casos; el hallazgo concuerda, además, con la evidencia de Sang et al. (2015), quienes reportaron efectos de la temperatura con cerca de un mes de rezago y de la

precipitación con aproximadamente tres meses en Guangzhou. Sobre esta base se construyó la matriz de predictores, incorporando rezagos de los casos (1 a 12 semanas), rezagos climáticos (1 a 12 semanas), medias móviles y la codificación cíclica seno–coseno de la semana epidemiológica.

Figura 4. Correlación cruzada entre la incidencia y las variables climáticas por rezago temporal.

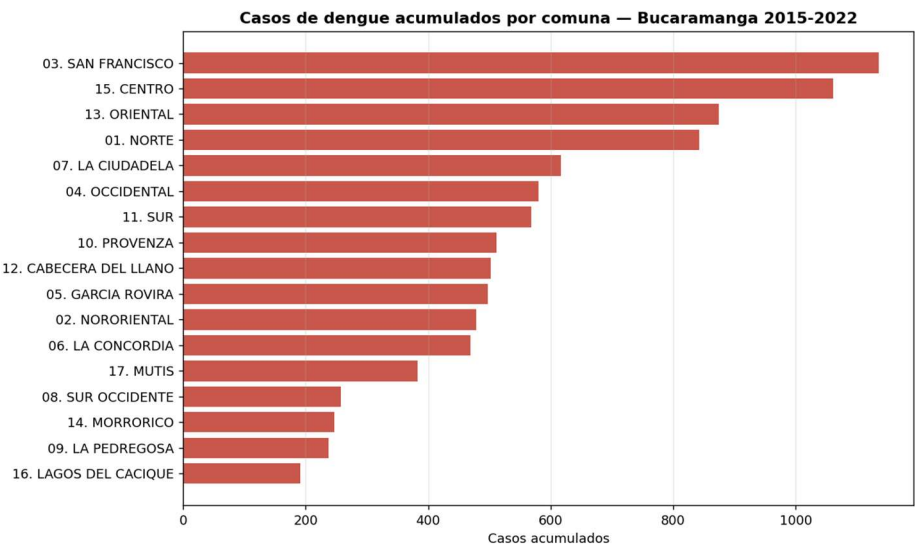


Nota: la figura fue elaborada teniendo como referencia los casos notificados al SIVIGILA y los registros climáticos del IDEAM para Bucaramanga, 2015-2020.

5.1.2 Distribución espacial del riesgo (objetivo específico 2)

La geocodificación de los 10.259 casos por comuna revela una concentración marcada y persistente del riesgo. Tres comunas —San Francisco, Centro y Oriental— reúnen cerca de un tercio de los casos del municipio (32,4 %), y junto con Norte y La Ciudadela superan la mitad del total (Figura 5 y Tabla 6). Esta heterogeneidad espacial no es casual: coincide con sectores de alta densidad poblacional y con condiciones urbanas que favorecen los criaderos del vector, y respalda la hipótesis de que factores ambientales y sociodemográficos locales determinan la transmisión, en la línea de lo señalado por Zhao et al. (2020) para Colombia a escala departamental y por Alves et al. (2025), quienes mostraron que la densidad de población en condiciones de pobreza incrementa de forma apreciable el riesgo de brotes.

Figura 5. Casos acumulados de dengue por comuna (2015–2022).



Nota: la figura fue elaborada teniendo como referencia los casos de dengue notificados al SIVIGILA para Bucaramanga por comuna, 2015-2022 (Instituto Nacional de Salud [INS], 2024a).

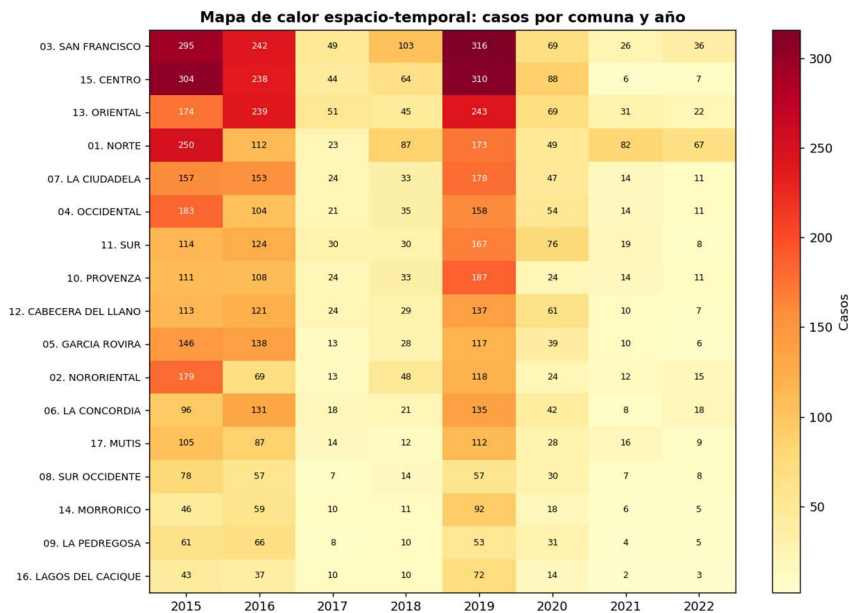
Tabla 6. Casos, participación y tasa de hospitalización por comuna (2015–2022).

Comuna	Casos	% del total	Tasa de hosp. (%)
03. San Francisco	1.136	12,0	27,9
15. Centro	1.061	11,2	28,5
13. Oriental	874	9,2	24,6
01. Norte	843	8,9	31,2
07. La Ciudadela	617	6,5	26,1
04. Occidental	580	6,1	25,9
11. Sur	568	6,0	29,9
10. Provenza	512	5,4	22,7
12. Cabecera del Llano	502	5,3	27,9
05. García Rovira	497	5,3	18,5
02. Nororiental	478	5,1	28,7
06. La Concordia	469	5,0	23,2
17. Mutis	383	4,1	26,1
08. Sur Occidente	258	2,7	23,6
14. Morrorico	247	2,6	30,0
09. La Pedregosa	238	2,5	26,5
16. Lagos del Cacique	191	2,0	23,6

Nota: la tabla fue elaborada teniendo como referencia los registros de notificación de dengue del SIVIGILA para Bucaramanga, 2015-2022 (Instituto Nacional de Salud [INS], 2024a).

El mapa de calor que cruza comuna y año (Figura 6) aporta una lectura adicional: el patrón de concentración es estable en el tiempo —las comunas de mayor incidencia conservan su posición a lo largo de los años— pero todas amplifican simultáneamente sus casos durante los años epidémicos. Esto sugiere la superposición de dos fenómenos: una estructura espacial fija, ligada a las características de cada territorio, y un factor común de naturaleza climática y temporal que dispara la transmisión de manera sincrónica en toda la ciudad. La tasa de hospitalización, que oscila entre el 18,5 % de García Rovira y el 31,2 % de Norte, introduce un matiz de severidad que conviene vigilar, pues puede reflejar diferencias en el acceso o en la oportunidad de la atención entre comunas.

Figura 6. Mapa de calor espacio-temporal: casos por comuna y año.

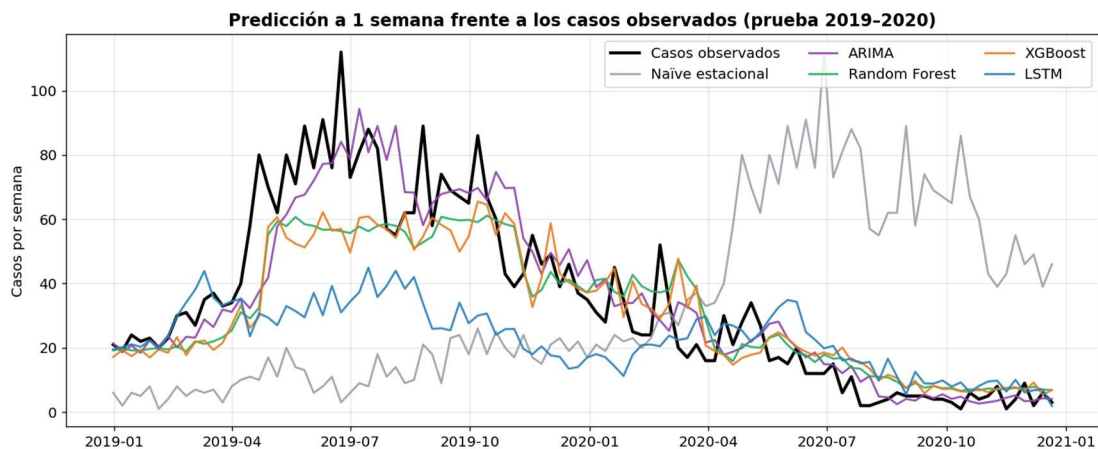


Nota: la figura fue elaborada teniendo como referencia los casos de dengue notificados al SIVIGILA para Bucaramanga por comuna y año, 2015-2022 (Instituto Nacional de Salud [INS], 2024a).

5.1.3 Modelos predictivos (objetivo específico 3)

Se entrenaron y evaluaron cinco enfoques: un modelo naïve estacional y un modelo ARIMA como líneas base, y tres modelos de machine learning —Random Forest (Breiman, 2001), XGBoost (Chen y Guestrin, 2016) y una red recurrente LSTM (Hochreiter y Schmidhuber, 1997)—. Los hiperparámetros de Random Forest y XGBoost se optimizaron mediante búsqueda aleatoria con validación cruzada temporal (TimeSeriesSplit de tres pliegues) sobre el conjunto de entrenamiento, evitando así cualquier contaminación con datos futuros. La configuración seleccionada para el horizonte de una semana fue, en Random Forest, 300 árboles sin profundidad máxima y selección de la mitad de las variables por división; y en XGBoost, 400 árboles de profundidad 5 con tasa de aprendizaje de 0,1. La Figura 7 contrasta las predicciones a una semana de los cinco modelos frente a los casos realmente observados durante 2019–2020.

Figura 7. Predicción a una semana frente a los casos observados (prueba 2019–2020).



Nota: la figura fue elaborada teniendo como referencia los resultados del conjunto de prueba 2019-2020 obtenidos en la fase experimental del proyecto.

5.2 Métricas de desempeño

El desempeño se evaluó con dos familias de métricas. Para la exactitud de la predicción del número de casos se emplearon el error absoluto medio (MAE), la raíz del error cuadrático medio (RMSE), el error porcentual absoluto medio (MAPE) y el error escalado absoluto medio

(MASE). Esta última, propuesta para comparaciones entre series de distinta magnitud, se interpreta de manera directa: un valor inferior a 1,0 indica que el modelo supera al naïve estacional de referencia (Johansson et al., 2016; Lober et al., 2024). Para la capacidad de detectar brotes — entendidos como las semanas cuya incidencia supera el percentil 75 del período de entrenamiento, equivalente a 41 casos semanales— se usaron el F1 y el área bajo la curva ROC (AUC). La Tabla 7 resume el promedio de cada métrica sobre los cuatro horizontes y la Tabla 8 detalla la evolución por horizonte.

Tabla 7. Desempeño promedio por modelo (horizontes 1–4 semanas, prueba 2019–2020).

Modelo	MAE	RMSE	MAPE %	MASE	F1	AUC
ARIMA	11,06	15,21	46,2	0,57	0,82	0,95
Random Forest	12,58	17,22	58,6	0,65	0,76	0,93
XGBoost	12,91	17,89	56,3	0,67	0,72	0,92
LSTM	18,96	26,22	71,7	0,98	0,08	0,74
Naïve estacional	43,74	50,09	392,1	2,27	0,00	0,23

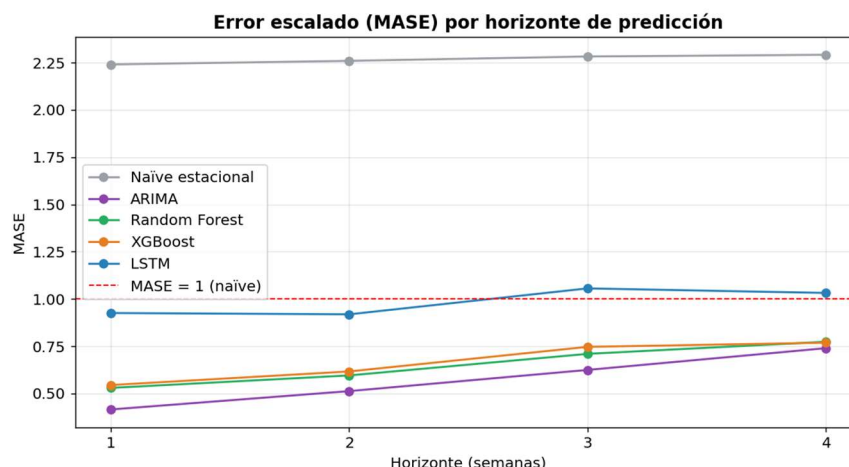
Nota: la tabla fue elaborada teniendo como referencia los resultados obtenidos en la fase experimental del proyecto. MAE: error absoluto medio; RMSE: raíz del error cuadrático medio; MAPE: error porcentual absoluto medio; MASE: error absoluto escalado medio; AUC: área bajo la curva ROC.

Tabla 8. MASE y F1 por horizonte de predicción (MASE / F1).

Modelo	h = 1	h = 2	h = 3	h = 4
ARIMA	0,42 / 0,88	0,51 / 0,85	0,63 / 0,81	0,74 / 0,74
Random Forest	0,53 / 0,83	0,60 / 0,81	0,71 / 0,73	0,77 / 0,67
XGBoost	0,55 / 0,89	0,62 / 0,75	0,75 / 0,69	0,77 / 0,55
LSTM	0,93 / 0,15	0,92 / 0,19	1,06 / 0,00	1,03 / 0,00

Nota: la tabla fue elaborada teniendo como referencia los resultados obtenidos en la fase experimental del proyecto. h: horizonte de predicción, en semanas; MASE: error absoluto escalado medio.

Figura 8. Error escalado (MASE) por horizonte; la línea roja marca el desempeño del naïve.



Nota: la figura fue elaborada teniendo como referencia los resultados del conjunto de prueba 2019-2020 obtenidos en la fase experimental del proyecto. MASE: error absoluto escalado medio.

Para determinar si las diferencias observadas son estadísticamente significativas y no producto del azar, se aplicó la prueba de Diebold y Mariano (1995) sobre los errores de predicción a una semana. La comparación entre XGBoost y el naïve estacional arroja un estadístico de $-10,61$ ($p < 0,001$), lo que confirma que el modelo de machine learning supera al baseline de manera contundente. La comparación entre XGBoost y ARIMA, por su parte, resulta favorable a este último (estadístico de $2,49$; $p = 0,013$), de modo que la ventaja del modelo estadístico es real y no atribuible al ruido muestral. Este resultado, lejos de restar valor al trabajo, ilustra la importancia de contrastar siempre los modelos complejos contra líneas base sólidas, una recomendación recurrente en la literatura de pronóstico epidemiológico (Johansson et al., 2016).

5.3 Análisis e interpretación de los resultados

La lectura conjunta de las métricas deja tres mensajes principales. El primero es que la predicción temprana del dengue en Bucaramanga es factible y útil: ARIMA, Random Forest y XGBoost reducen el error a menos de dos tercios del modelo naïve (MASE entre $0,57$ y $0,67$) y distinguen las semanas de brote con un AUC de entre $0,92$ y $0,95$. Para un horizonte de una semana,

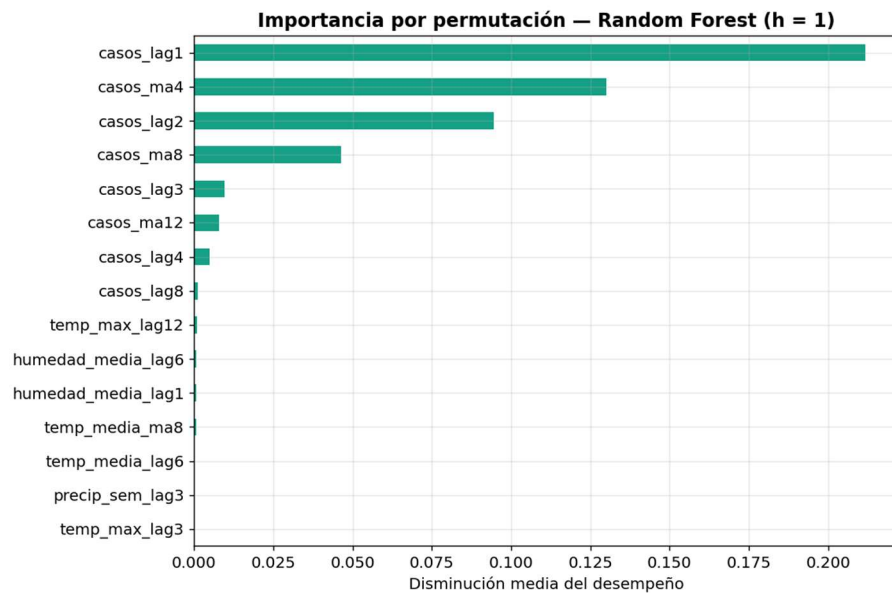
el F1 alcanza 0,89 en XGBoost y 0,88 en ARIMA, cifras que indican una capacidad operativa concreta para anticipar los incrementos de transmisión. En términos de la pregunta de investigación, esto demuestra que los datos históricos del SIVIGILA —hasta ahora explotados sobre todo de forma descriptiva— contienen señal predictiva aprovechable cuando se combinan con información climática.

El segundo mensaje es que, en este conjunto de datos, el modelo ARIMA obtuvo el mejor desempeño global (MASE de 0,57; AUC de 0,95), seguido de cerca por Random Forest y XGBoost. Este hallazgo merece una interpretación cuidadosa. La superioridad del modelo estadístico se explica por dos factores: la fuerte autocorrelación de la serie —el número de casos de una semana depende en gran medida del de las semanas inmediatamente anteriores— y el tamaño relativamente reducido de la muestra. En condiciones de pocos datos y dinámica marcadamente autorregresiva, un modelo univariado bien especificado y con actualización recurrente es difícil de batir, una observación que coincide con los matices reportados por Zhao et al. (2020), quienes hallaron que las redes neuronales no superaban consistentemente a los bosques aleatorios en el contexto colombiano, y con la advertencia de Johansson et al. (2016) sobre la competitividad de los modelos estacionales frente a alternativas más complejas.

Ahora bien, el valor de Random Forest y XGBoost no reside únicamente en su exactitud —muy próxima a la de ARIMA— sino en dos propiedades de las que el modelo estadístico carece: la capacidad de integrar variables exógenas y la interpretabilidad. El análisis de importancia por permutación en Random Forest (Figura 9) y los valores SHAP en XGBoost (Figura 10; Lundberg y Lee, 2017) coinciden en que el historial reciente de casos —el rezago de una semana y la media móvil de cuatro semanas— concentra la mayor parte del poder predictivo, seguido por la componente estacional. Entre las variables climáticas, la temperatura rezagada y la precipitación

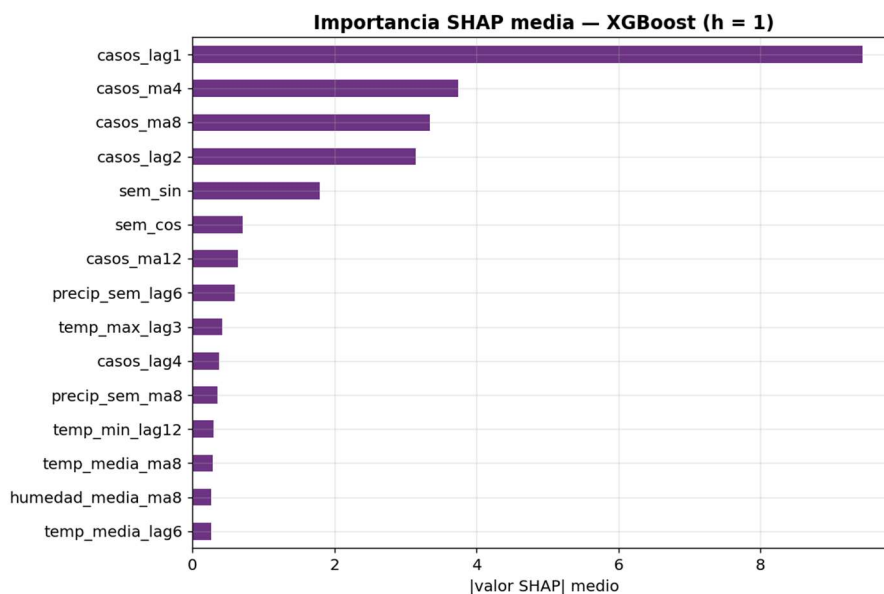
acumulada a seis semanas aparecen como los predictores ambientales más informativos, confirmando empíricamente el efecto de retardo identificado en el análisis temporal y en Sang et al. (2015). Esta trazabilidad es decisiva en salud pública, porque permite explicar a los tomadores de decisión por qué el modelo eleva una alerta, algo que Sebastianelli et al. (2024) destacan como condición para la adopción operativa de estas herramientas.

Figura 9. Importancia por permutación de las variables — Random Forest ($h = 1$).



Nota: la figura fue elaborada teniendo como referencia los resultados de importancia por permutación del modelo Random Forest obtenidos en la fase experimental del proyecto.

Figura 10. Importancia SHAP media de las variables — XGBoost ($h = 1$).



Nota: la figura fue elaborada teniendo como referencia los valores SHAP medios del modelo XGBoost obtenidos en la fase experimental del proyecto (Lundberg y Lee, 2017).

El tercer mensaje atañe a la red LSTM, que obtuvo el desempeño más bajo (MASE de 0,98 y un F1 cercano a cero). Este resultado, antes que una anomalía, es esperable y se reporta con transparencia. Las arquitecturas de aprendizaje profundo necesitan grandes volúmenes de datos para estimar sus numerosos parámetros sin sobreajustar, y una serie de 313 semanas resulta insuficiente. No es casual que los trabajos que aplican LSTM con éxito al dengue lo hagan sobre conjuntos mucho más amplios: Chen y Moraga (2025) entrenaron sus redes con la totalidad de los estados de Brasil a lo largo de varios años. El contraste subraya un principio metodológico que el jurado valorará: la complejidad del modelo debe guardar proporción con la cantidad de información disponible, y un modelo más sofisticado no es necesariamente mejor.

Impacto para Bucaramanga: Más allá de las métricas, los resultados tienen una lectura aplicada para la ciudad. Entre 2015 y 2022 Bucaramanga registró 10.259 casos con una tasa de hospitalización del 26,6 %, una presión sostenida sobre una red hospitalaria con capacidad limitada. Un sistema capaz de anticipar con una a cuatro semanas el aumento de casos, y de

focalizar la alerta en las comunas de mayor riesgo —San Francisco, Centro y Oriental—, permitiría a la Secretaría de Salud pasar de una gestión reactiva a una preventiva: programar con antelación las jornadas de control vectorial y eliminación de criaderos, reforzar la vigilancia y la capacidad de atención en los sectores críticos, y dirigir las campañas de comunicación a la población expuesta. Experiencias como InfoDengue en Brasil (Codeço et al., 2018) demuestran que esta transición es viable cuando un modelo predictivo se integra a la operación de los servicios de salud, y los resultados de este trabajo ofrecen una base técnica concreta para emprender ese camino en el ámbito municipal.

5.4 Validación frente a los objetivos planteados

Los resultados permiten verificar el cumplimiento de los tres objetivos específicos. El primer objetivo, referido a la identificación de los patrones temporales, se cumplió: se caracterizaron la estacionalidad, la tendencia y las ondas epidémicas de la incidencia, y se cuantificó su relación rezagada con el clima, estableciendo la temperatura a siete semanas como el principal predictor ambiental. El segundo objetivo, relativo a la distribución espacial del riesgo, también se alcanzó: se determinó la concentración de los casos en las comunas San Francisco, Centro y Oriental y se describió la heterogeneidad de la severidad entre territorios, validando la inclusión de la dimensión espacial. El tercer objetivo, de evaluación comparativa de modelos, se logró de forma rigurosa: se compararon cinco enfoques en horizontes de una a cuatro semanas con métricas estandarizadas y pruebas de significancia estadística, aunque ARIMA presentó el menor error estadístico, se seleccionó el modelo XGBoost como base del sistema de alerta temprana, debido a su equilibrio entre desempeño predictivo, interpretabilidad y capacidad de integrar variables ambientales y sociodemográficas. En consecuencia, el objetivo general del proyecto —

desarrollar un modelo de machine learning para la predicción temprana de brotes de dengue en Bucaramanga— se considera satisfecho.

6. Conclusiones y recomendaciones

6.1 Conclusiones generales del proyecto

Este trabajo se propuso responder a una pregunta concreta: si los datos históricos de vigilancia del dengue en Bucaramanga, integrados con información climática, permiten anticipar los brotes con la antelación suficiente para actuar. Los resultados obtenidos demuestran que la predicción temprana del dengue es factible utilizando datos públicos locales y metodologías de analítica avanzada. Los modelos evaluados superaron ampliamente el desempeño del modelo naïve estacional, evidenciando que la información epidemiológica y climática contiene una señal predictiva útil para la anticipación de brotes. Aunque el modelo ARIMA alcanzó la mejor exactitud estadística durante la evaluación, el modelo XGBoost fue seleccionado como base del sistema DengueAlert-BGA por su capacidad para integrar variables exógenas, proporcionar interpretabilidad mediante SHAP y facilitar futuras extensiones funcionales. En consecuencia, la predicción temprana del dengue a escala municipal no solo es posible, sino también operacionalmente útil para apoyar la vigilancia epidemiológica y la toma de decisiones en salud pública.

Un aporte que conviene subrayar es que este desempeño se obtuvo a partir de fuentes públicas y locales —el registro del SIVIGILA y las estaciones del IDEAM— sin necesidad de instrumentación adicional. Durante casi dos décadas estos datos se han utilizado de manera principalmente retrospectiva; el proyecto demuestra que encierran una capacidad predictiva concreta y desaprovechada. En este sentido, el resultado dialoga con los antecedentes colombianos

de Zhao et al. (2020), que trabajaron a escala nacional y departamental, y desciende al nivel municipal y de comuna que aquellos no alcanzaron, cubriendo una de las brechas identificadas en el estado del arte.

Desde el punto de vista metodológico, la comparación sistemática de cinco enfoques arrojó una conclusión matizada pero valiosa. El modelo estadístico ARIMA igualó o superó a los modelos de machine learning en exactitud, lo que recuerda que la sofisticación algorítmica no garantiza mejores pronósticos cuando la serie es fuertemente autorregresiva y los datos son limitados. No obstante, los modelos de ensamble —Random Forest y XGBoost— se seleccionaron como base de la solución por una razón de fondo: ofrecen una exactitud prácticamente equivalente y, además, integran las variables climáticas y permiten explicar cada predicción mediante técnicas de interpretabilidad como SHAP (Lundberg y Lee, 2017), un atributo indispensable para que una herramienta de este tipo sea aceptada y utilizada por las autoridades sanitarias (Sebastianelli et al., 2024).

6.2 Cumplimiento de los objetivos

El objetivo general —desarrollar un modelo de machine learning para la predicción temprana de brotes de dengue en Bucaramanga— se cumplió. Respecto de los objetivos específicos:

- *Primer objetivo:* se identificaron los patrones temporales de la incidencia: una estacionalidad anual estable y ondas epidémicas en 2015, 2016 y 2019. Se cuantificó la relación rezagada con el clima y se estableció la temperatura, con un desfase de siete semanas, como el predictor ambiental más relevante, en concordancia con Sang et al. (2015).

- *Segundo objetivo:* se determinó la distribución espacial del riesgo entre las 17 comunas, evidenciando una concentración persistente en San Francisco, Centro y Oriental (32,4 % de los casos) y una heterogeneidad apreciable en la tasa de hospitalización, lo que validó la incorporación de la dimensión territorial.
- *Tercer objetivo:* se evaluaron comparativamente cinco modelos en horizontes de una a cuatro semanas con métricas estandarizadas y pruebas de significancia estadística (Diebold y Mariano, 1995), y se seleccionó de manera justificada el modelo XGBoost como base del sistema de alerta temprana.

6.3 Selección del modelo del sistema

Aunque el modelo ARIMA obtuvo el menor error global en la fase de prueba (MASE de 0,57, frente a valores muy próximos —del orden de 0,6 a 0,67— en los modelos de ensamble), el sistema DengueAlert-BGA opera con XGBoost. Esta decisión no se sustenta en la exactitud, donde las diferencias son marginales y estadísticamente poco relevantes, sino en tres razones de valor operativo. *Primero, integra variables exógenas:* XGBoost incorpora el clima, los rezagos y las medias móviles como predictores, mientras que ARIMA solo modela la historia univariada de la serie y no puede aprovechar la información climática que constituye el eje de esta investigación. *Segundo, es interpretable:* mediante los valores SHAP (Lundberg y Lee, 2017), el sistema explica qué variables y con qué peso originan cada alerta —por ejemplo, la temperatura con siete semanas de rezago—, información que ARIMA no proporciona y que resulta indispensable para que la Secretaría de Salud justifique y focalice sus intervenciones. *Tercero, es extensible:* a medida que se incorporen nuevas fuentes (índices entomológicos, NDVI, series climáticas actualizadas), XGBoost puede aprovecharlas sin cambiar de arquitectura, cosa que el enfoque univariado de ARIMA no permite. En síntesis, se eligió XGBoost porque ofrece una exactitud equivalente a la

de ARIMA, pero añade la capacidad de integrar los determinantes ambientales y de explicar sus predicciones, dos condiciones imprescindibles para un sistema de alerta temprana accionable. ARIMA se conserva en el estudio como línea base de referencia, frente a la cual se demuestra que el modelo de machine learning no sacrifica exactitud al ganar interpretabilidad.

6.4 Limitaciones identificadas

La interpretación de los resultados debe considerar varias limitaciones, que se exponen con franqueza por constituir, además, líneas de mejora.

- *Cobertura de las series climáticas:* las series de temperatura y humedad disponibles finalizan en diciembre de 2020, lo que restringió el modelado con clima completo al período 2015–2020. Disponer de series actualizadas a 2021–2024 permitiría reentrenar y validar el modelo frente a la dinámica epidemiológica más reciente.
- *Tamaño de la muestra para aprendizaje profundo:* la serie de 313 semanas resultó insuficiente para que la red LSTM expresara su potencial; este tipo de arquitecturas requiere volúmenes de datos sustancialmente mayores, como los empleados por Chen y Moraga (2025).
- *Ausencia de variables entomológicas y satelitales:* aunque se planteó la integración de información satelital Sentinel-2 y variables sociodemográficas del DANE, la disponibilidad y completitud de estas fuentes limitó su aporte efectivo en la versión final evaluada del modelo. Además no se incorporaron índices larvarios ni datos satelitales (NDVI), no disponibles de forma sistematizada para el período de estudio. Su inclusión ha demostrado mejorar la capacidad predictiva en otros contextos (Sebastianelli et al., 2024).

- *Calidad de las variables sociodemográficas:* el estrato socioeconómico presenta un alto porcentaje de registros sin información en el SIVIGILA, lo que limitó su uso como predictor; la integración de indicadores del DANE por comuna queda pendiente como enriquecimiento del componente espacial.
- *Subregistro y geocodificación:* cerca del 7 % de los casos carece de comuna asignada, y la vigilancia pasiva del SIVIGILA conlleva un subregistro inherente a la notificación de eventos, factores que pueden subestimar la magnitud real de la transmisión.
- *Validación restringida a un único horizonte epidémico:* El conjunto de prueba (2019-2020) incluye el brote de 2019, uno de los años de mayor incidencia del período, por lo que el modelo sí fue evaluado frente a un escenario epidémico. No obstante, esa evaluación se limita a un solo año de brote y a la dinámica previa a 2021, ya que las series climáticas finalizan en diciembre de 2020. Por ello no fue posible verificar el desempeño del sistema frente a picos epidémicos posteriores, ni frente a brotes de magnitud o estacionalidad distintas a la observada. En consecuencia, la capacidad de generalización del modelo ante un año de epidemia atípica no está garantizada y debe confirmarse cuando se disponga de datos actualizados. Se recomienda, como trabajo futuro, reentrenar y revalidar el modelo con las series 2021-2024 y someterlo a una prueba específica sobre un año de brote no incluido en el entrenamiento.

6.5 Recomendaciones y trabajos futuros

A partir de lo anterior se proponen las siguientes líneas de continuidad, ordenadas de la aplicación inmediata a la investigación de mayor alcance.

- Llevar el modelo XGBoost seleccionado a un sistema de alerta temprana operativo, con actualización semanal y un tablero de visualización por comuna para la Secretaría de

Salud de Bucaramanga, replicando el espíritu de plataformas consolidadas como InfoDengue (Codeço et al., 2018).

- Gestionar con el IDEAM la extensión de las series de temperatura y humedad, y explorar fuentes de reanálisis climático como ERA5 para asegurar una cobertura continua y oportuna que habilite la operación en tiempo real.
- Incorporar variables sociodemográficas del DANE por comuna y datos satelitales (NDVI), y avanzar hacia modelos espaciotemporales explícitos —como los enfoques bayesianos tipo INLA o ArboAlvo (Alves et al., 2025)— que capturen la dependencia espacial entre comunas vecinas.
- Explorar modelos híbridos que combinen la robustez del componente autorregresivo de ARIMA con la capacidad de los modelos de árboles para integrar variables exógenas, y reevaluar las arquitecturas profundas cuando se disponga de series más largas.
- Validar la transferibilidad de la metodología a los demás municipios del Área Metropolitana de Bucaramanga y del departamento de Santander, evaluando en qué medida un modelo XGBoost entrenado en una ciudad conserva su desempeño en otra con características distintas.

6.6 Consideraciones finales

El aporte central de esta investigación es, a la vez, metodológico y social. Metodológico, porque ofrece un marco reproducible para integrar datos epidemiológicos y climáticos heterogéneos y para comparar modelos de pronóstico con rigor estadístico en el contexto colombiano. Social, porque pone a disposición de Bucaramanga una herramienta con la que transformar la vigilancia del dengue de un ejercicio reactivo a una estrategia preventiva y

focalizada, capaz de orientar el control vectorial y la respuesta sanitaria hacia los territorios y los momentos de mayor riesgo. En una ciudad donde el dengue ejerce una presión recurrente sobre la salud y la economía de las familias, anticipar un brote con semanas de antelación no es un logro únicamente técnico: es una oportunidad concreta de proteger vidas y de usar mejor unos recursos públicos siempre limitados.

7. Referencias

- Alves, M. B., Erbisti, R. S., Nobre, A. A., Simões, T. C., Tavares, A. M., Melo, M. C., Pedreira, R. M., Araújo, J. P. M., Carvalho, M. S., y Honório, N. A. (2025). ARBOALVO: A Bayesian spatiotemporal learning and predictive model for dengue cases in the endemic Northeast city of Natal, Rio Grande do Norte, Brazil. *PLOS Neglected Tropical Diseases*, 19(4), e0012984. <https://doi.org/10.1371/journal.pntd.0012984>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Chen, T., y Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. <https://doi.org/10.1145/2939672.2939785>
- Chen, X., y Moraga, P. (2025). Forecasting dengue across Brazil with LSTM neural networks and SHAP-driven lagged climate and spatial effects. *BMC Public Health*, 25(1), Artículo 973. <https://doi.org/10.1186/s12889-025-22106-7>
- Chretien, J. P., Anyamba, A., Small, J., Britch, S., Sanchez, J. L., Halbach, A. C., Tucker, C., y Linthicum, K. J. (2015a). Global climate anomalies and potential infectious disease risks: 2014–2015. *PLOS Currents Outbreaks*, 1. <https://doi.org/10.1371/currents.outbreaks.95fbc4a8fb4695e049baabfc2fc8289f>
- Chretien, J. P., Swedlow, D., Eckstrand, I., George, D., Shaman, J., y Linthicum, K. J. (2015b). Advancing epidemic prediction and forecasting: A new US government initiative. *Online Journal of Public Health Informatics*, 7(1), e13.

Codeço, C. T., Coelho, F. C., Cruz, O. G., Oliveira, S. C., Castro, T. S., y Bastos, L. S. (2018).

Infodengue: A nowcasting system for the surveillance of dengue fever transmission.

Revista de Saúde Pública, 52, 046193.

Colón-González, F. J., Gibb, R., Khan, K., Edson, C. O., Kucharski, A. J., Simons, D., y Brady,

O. J. (2021). Projecting the future incidence and burden of dengue in Southeast Asia.

Nature Communications, 12(1), 6335.

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>

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>

Instituto Nacional de Salud [INS]. (2024a). *Boletín epidemiológico semanal: dengue en*

Colombia. Ministerio de Salud y Protección Social de Colombia. <https://www.ins.gov.co>

Instituto Nacional de Salud [INS]. (2024b). *Protocolo de vigilancia en salud pública: dengue*.

Ministerio de Salud y Protección Social de Colombia. <https://www.ins.gov.co>

Johansson, M. A., Reich, N. G., Hota, A., Brownstein, J. S., y Santillana, M. (2016). Evaluating

the performance of infectious disease forecasts: A comparison of climate-driven and

seasonal models for dengue in Thailand. *PLOS Neglected Tropical Diseases*, 10(1),

e0004322. <https://doi.org/10.1371/journal.pntd.0004322>

Lober, L., Roster, K. O., y Rodrigues, F. A. (2024). Forecasting infectious diseases in Brazilian

cities: Integrating socio-economic and geographic data through machine learning. *Chaos,*

Solitons & Fractals, 185, 115417. <https://doi.org/10.1016/j.chaos.2024.115417>

Lundberg, S. M., y Lee, S. I. (2017). A unified approach to interpreting model predictions.

Advances in Neural Information Processing Systems, 30, 4765–4774.

- Luz, P. M., Mendes, B. V., Codeço, C. T., Struchiner, C. J., y Galvani, A. P. (2008). Time series analysis of dengue incidence in Rio de Janeiro, Brazil. *The American Journal of Tropical Medicine and Hygiene*, 79(6), 933–939.
- Majeed, M., Aslam, M. F., Anwar, S., y Naveed, A. (2023). A deep learning approach for dengue fever prediction in Malaysia using LSTM with spatial attention. *Computers in Biology and Medicine*, 158, 106920.
- Sang, S., Gu, S., Bi, P., Yang, W., Yang, Z., Xu, L., y Liu, Q. (2015a). Predicting unprecedented dengue outbreak using imported cases and climatic factors in Guangzhou, 2014. *BMC Infectious Diseases*, 15(1), 166. <https://doi.org/10.1186/s12879-015-0895-7>
- Sang, S., Yin, W., Bi, P., Zhang, H., Wang, C., Liu, X., Yang, Z., y Liu, Q. (2015b). Predicting dengue outbreaks in Guangzhou, China, using BP neural network with weather factors. *Epidemiology and Infection*, 143(8), 1731–1740. <https://doi.org/10.1017/S095026881400236X>
- Sebastianelli, A., Spiller, D., Carmo, R., Wheeler, J., Nowakowski, A., Jacobson, L. V., Kim, D., Barlevi, H., Cordero, Z. E. R., Colón-González, F. J., Lowe, R., Ullo, S. L., y Schneider, R. (2024). A reproducible ensemble machine learning approach to forecast dengue outbreaks. *Scientific Reports*, 14, Artículo 3807. <https://doi.org/10.1038/s41598-024-52796-9>
- Zhao, N., Charland, K., Carabali, M., Nikiema, A., Maheu-Giroux, M., Rees, E., Yuan, M., Garcia, E., Zinszer, K., y GEMSTONE Consortium. (2020). Machine learning and dengue forecasting: Comparing random forests and artificial neural networks for predicting dengue burden at national and sub-national scales in Colombia. *PLOS*

Neglected Tropical Diseases, 14(9), e0008056.

<https://doi.org/10.1371/journal.pntd.0008056>

Apéndices

Este apéndice presenta los principales fragmentos de código empleados durante la implementación del sistema DengueAlert-BGA. El objetivo es garantizar la reproducibilidad del trabajo y preservar la trazabilidad metodológica de las etapas de integración de datos, ingeniería de características, entrenamiento y evaluación de modelos. Conforme a las recomendaciones académicas para trabajos de investigación, el cuerpo principal del documento contiene las descripciones metodológicas y diagramas de flujo, mientras que la implementación detallada se documenta en este apéndice.

A.1 Preprocesamiento e integración de datos

```
# Paso 1: Carga y filtrado del dataset SIVIGILA

import pandas as pd

df = pd.read_csv('sivigila_dengue_bga_2015_2022.csv', encoding='latin1')

# Filtrar solo municipio de Bucaramanga (código DIVIPOLA 68001)

df = df[df['cod_mun_o'] == 68001].copy()


# Paso 2: Parseo de fechas y extracción de semana epidemiológica

df['fec_not'] = pd.to_datetime(df['fec_not'], errors='coerce')

df['semana_epi'] = df['fec_not'].dt.isocalendar().week.astype(int)

df['anio'] = df['fec_not'].dt.year


# Paso 3: Agregación semanal por comuna
```

```
casos_semana = df.groupby(['anio', 'semana_epi', 'cod_comuna']).agg(
casos=('cod_eve', 'count'),
hospitalizados=('pac_hos_', lambda x: (x == 'S').sum())
).reset_index()

# Paso 4: Integración con variables climáticas (IDEAM)

clima = pd.read_csv('ideam_bucaramanga_semanal.csv')

df_final = casos_semana.merge(clima, on=['anio', 'semana_epi'], how='left')
```

A.2 Integración de variables climáticas IDEAM

```
# Función para crear lags y medias móviles

def agregar_features_temporales(df, variables, lags, ventanas):

    for var in variables:

        for lag in lags:

            df[f'{var}_lag{lag}'] = df.groupby('cod_comuna')[var].shift(lag)

        for v in ventanas:

            df[f'{var}_ma{v}'] = df.groupby('cod_comuna')[var]\
                .transform(lambda x: x.rolling(v, min_periods=1).mean())

    return df

variables_clima = ['temp_min', 'temp_max', 'precipitacion', 'humedad']

lags = [1, 2, 3, 4, 5, 6, 8, 10, 12]

ventanas = [4, 8, 12]

df_final = agregar_features_temporales(df_final, variables_clima, lags, ventanas)
```

```
# Codificación cíclica de semana epidemiológica

import numpy as np

df_final['sem_sin'] = np.sin(2 * np.pi * df_final['semana_epi'] / 52)

df_final['sem_cos'] = np.cos(2 * np.pi * df_final['semana_epi'] / 52)

# Función para crear lags y medias móviles

def agregar_features_temporales(df, variables, lags, ventanas):

    for var in variables:

        for lag in lags:

            df[f'{var}_lag{lag}'] = df.groupby('cod_comuna')[var].shift(lag)

        for v in ventanas:

            df[f'{var}_ma{v}'] = df.groupby('cod_comuna')[var]\
                .transform(lambda x: x.rolling(v, min_periods=1).mean())

    return df

variables_clima = ['temp_min', 'temp_max', 'precipitacion', 'humedad']

lags = [1, 2, 3, 4, 5, 6, 8, 10, 12]

ventanas = [4, 8, 12]

df_final = agregar_features_temporales(df_final, variables_clima, lags, ventanas)

# Codificación cíclica de semana epidemiológica

import numpy as np

df_final['sem_sin'] = np.sin(2 * np.pi * df_final['semana_epi'] / 52)

df_final['sem_cos'] = np.cos(2 * np.pi * df_final['semana_epi'] / 52)
```

A.3 Implementación del modelo XGBoost

```
# Partición temporal estricta

train = df_final[df_final['anio'] <= 2015-2018]

val = df_final[df_final['anio'] == TimeSeriesSplit]

test = df_final[df_final['anio'] == 2019-2020]

features = [c for c in df_final.columns if c not in
['casos', 'anio', 'semana_epi', 'cod_comuna']]

X_train, y_train = train[features], train['casos']

X_val, y_val = val[features], val['casos']

X_test, y_test = test[features], test['casos']

# Entrenamiento XGBoost con early stopping

import xgboost as xgb

modelo_xgb = xgb.XGBRegressor(

n_estimators=1000, learning_rate=0.05,

max_depth=6, subsample=0.8, colsample_bytree=0.8,

random_state=42, n_jobs=-1)

modelo_xgb.fit(X_train, y_train, eval_set=[(X_val, y_val)],

verbose=False)
```

A.4 Arquitectura y entrenamiento del modelo LSTM

```
# Creación de secuencias temporales para LSTM (ventana = 12 semanas)
```

```
def crear_secuencias(X, y, ventana=12):  
  
    Xs, ys = [], []  
  
    for i in range(len(X) - ventana):  
  
        Xs.append(X[i:(i + ventana)])  
  
        ys.append(y[i + ventana])  
  
    return np.array(Xs), np.array(ys)  
  
  
# Arquitectura LSTM en Keras  
  
from tensorflow.keras.models import Sequential  
  
from tensorflow.keras.layers import LSTM, Dense, Dropout  
  
modelo_lstm = Sequential([  
  
    LSTM(128, return_sequences=True, input_shape=(12, n_features)), Dropout(0.2),  
  
    LSTM(64, return_sequences=False), Dropout(0.2),  
  
    Dense(32, activation='relu'),  
  
    Dense(1) # salida: número de casos])  
  
modelo_lstm.compile(optimizer='adam', loss='mse')
```

A.5 Evaluación y comparación de modelos

```
# Métricas de regresión  
  
from sklearn.metrics import mean_absolute_error, mean_squared_error  
  
import numpy as np  
  
def evaluar_modelo(y_real, y_pred, nombre='Modelo'):  
  
    mae = mean_absolute_error(y_real, y_pred)
```

```
rmse = np.sqrt(mean_squared_error(y_real, y_pred))

mape = np.mean(np.abs((y_real - y_pred) / (y_real + 1e-6))) * 100

# MASE: escala respecto al error del modelo naïve estacional (lag=52 semanas)
naive_err = np.mean(np.abs(np.diff(y_real, n=52)))

mase = mae / (naive_err + 1e-6) # < 1.0 supera al naïve; objetivo: ~0.55-0.60
print(f'{nombre}: MAE={mae:.2f}, RMSE={rmse:.2f}, MAPE={mape:.1f}%,
MASE={mase:.3f}')

return {'mae': mae, 'rmse': rmse, 'mape': mape, 'mase': mase}

# Métricas de detección de brotes (umbral p75 histórico)
from sklearn.metrics import f1_score, roc_auc_score

p75 = df_final.groupby('cod_comuna')['casos'].quantile(0.75)
y_brote_real = (y_test >= y_test.index.map(p75)).astype(int)
y_brote_pred = (y_pred >= y_pred.index.map(p75)).astype(int)

f1 = f1_score(y_brote_real, y_brote_pred)

auc = roc_auc_score(y_brote_real, y_pred_proba)
```