PREDICCIÓN DE BROTES DE DENGUE, EN CALI, MEDELLÍN Y BUCARAMANGA UTILIZANDO MODELOS DE MACHINE LEARNING

Julian Mauricio Rayo Grajales Cod. 9013813

Julian Andres Pinto Montes Cod. 9013855

Christophe Eklouh Molinier Cod. 9014662

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

Directora
Delia Ortega Lenis

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, NOVIEMBRE DE 2025

FICHA RESUMEN
TRABAJO DE GRADO DE MAESTRÍA

TITULO: “Predicción de brotes de dengue en Cali, Medellín y Bucaramanga utilizando modelos de Machine Learning”

1. ÉNFASIS: Sistemas y Computación
2. TIPO DE PROYECTO: Aplicado
3. ÁREA DE TRABAJO: Ciencia de Datos, Epidemiología, Salud Pública
4. ESTUDIANTE (S): Julian Mauricio Rayo Grajales, Julian Andrés Pinto Montes, Christophe Eklouh Molinier
5. CORREO ELECTRÓNICO: julianrayo@javerianacali.edu.co, julianpinto@javerianacali.edu.co, eklouh23@javerianacali.edu.co
6. DIRECCIÓN Y TELÉFONO:
   Julian Rayo: Carrera 30 N0 4A 41 Palmira – Telefono 315 7970256
   Julian Pinto: Calle 10A #6-23 Chipre, Manizales – Telefono 310 6279205
   Christophe Eklouh Molinier: Carrera 84F #18-45, Medellín – Telefono 3244294846
7. DIRECTOR: Delia Ortega Lenis
8. VINCULACIÓN DEL DIRECTOR (en la universidad): Planta
9. CORREO ELECTRÓNICO DEL DIRECTOR: delia.ortega@javerianacali.edu.co
10. CO-DIRECTOR(ES) (Si aplica):
11. GRUPO O EMPRESA QUE LO AVALA (Si aplica):
12. OTROS GRUPOS O EMPRESAS:
13. PALABRAS CLAVE (al menos 5): Dengue, Machine Learning, Ciencia de Datos, Epidemiología, Variables Climáticas, Salud Pública
14. ODS QUE APLICA EL PROYECTO (Agenda 2030):
    ODS 3 – Salud y Bienestar
    ODS 9 – Industria, Innovación e Infraestructura
    ODS 11 – Ciudades y Comunidades Sostenibles
15. FECHA DE INICIO (Desarrollo del proyecto): 21/04/2025

16. RESUMEN (máximo 400 palabras):

El presente proyecto de Maestría en Ciencia de Datos se orienta a la predicción de brotes de dengue en las ciudades de Cali, Medellín y Bucaramanga utilizando modelos de aprendizaje automático. El dengue constituye un problema persistente de salud pública en Colombia, cuya dinámica está influenciada por factores climáticos, ambientales y socioeconómicos, lo que dificulta su control mediante métodos tradicionales de vigilancia epidemiológica.

Para el desarrollo del proyecto se construyó una base de datos integrada a partir de fuentes oficiales como SIVIGILA, IDEAM y DANE, que incluye registros semanales de casos de dengue desde 2007–2019, así como variables climáticas como temperatura, humedad y precipitación. Se realizó un proceso riguroso de limpieza, consolidación y análisis exploratorio de los datos para identificar patrones estacionales, correlaciones y rezagos temporales entre las variables.

El modelado predictivo se desarrolló bajo dos escenarios: uno basado únicamente en variables climáticas y otro que incorpora además la inercia epidemiológica mediante promedios móviles de contagios. Se implementaron diversos algoritmos de Machine Learning, incluyendo Random Forest, XGBoost, redes neuronales densas y modelos recurrentes tipo GRU. Los modelos fueron entrenados con partición temporal y evaluados con métricas de desempeño como el RMSE.

Los resultados evidencian que la incorporación de la inercia epidemiológica mejora significativamente la capacidad predictiva de los modelos, permitiendo anticipar semanas de alta incidencia con mayor precisión. Este proyecto aporta una herramienta analítica útil para fortalecer la vigilancia epidemiológica y apoyar la toma de decisiones en salud pública, contribuyendo a la gestión oportuna del riesgo y a la optimización de recursos en contextos urbanos vulnerables.

Tabla de contenido

LISTA DE FIGURAS

LISTA DE TABLAS

LISTA DE ANEXOS

1. Proyecto de Grado – Cali [Cuaderno de trabajo en Google Colab]. Google Colab. https://colab.research.google.com/drive/1SZJ9TGbEUqCMChMYhGgXibeXjmh_kDnf?usp=sharing

2. Proyecto de Grado – Medellín [Cuaderno de trabajo en Google Colab]. Google Colab. https://colab.research.google.com/drive/1ajwvON7qy1XSSOuV0dz5tVbv4OSAJF0v?usp=sharing

3. Proyecto de Grado – Bucaramanga [Cuaderno de trabajo en Google Colab]. Google Colab. https://colab.research.google.com/drive/1gddI-ibLzv8tBCTNZ6N3ArBpnZ-jY5XY?usp=sharing

INTRODUCCIÓN

Las enfermedades transmitidas por vectores, como el dengue, constituyen una grave amenaza para la salud pública en Colombia. Esta situación es especialmente preocupante en ciudades densamente pobladas y con climas tropicales como Cali, Medellín y Bucaramanga.

La naturaleza impredecible de los brotes de dengue dificulta su control y su prevención efectiva. Estos brotes están influenciados por factores climáticos, socioeconómicos y ambientales.

Este proyecto propone desarrollar un modelo de predicción de brotes de dengue utilizando técnicas de aprendizaje automático. A través del análisis de grandes volúmenes de datos históricos, se busca identificar patrones y correlaciones que permitan anticipar la ocurrencia de futuros brotes.

Los datos analizados incluyen registros de casos confirmados, variables climáticas como temperatura, precipitación y humedad, índices socioeconómicos como densidad poblacional y nivel de urbanización, así como datos geográficos.

El modelo de aprendizaje automático, una vez entrenado, será capaz de generar pronósticos a corto y mediano plazo sobre la incidencia de casos de dengue en las ciudades seleccionadas. Estos pronósticos proporcionarán a las autoridades de salud pública una herramienta valiosa para:

- **Detección temprana de brotes**: identificar áreas de alto riesgo y tomar medidas preventivas de manera oportuna.
- **Optimización de recursos**: asignar de forma eficiente los recursos humanos y materiales para el control de vectores y la atención a pacientes.
- **Evaluación de estrategias de intervención**: evaluar la eficacia de las diferentes intervenciones implementadas para combatir el dengue.

Al mejorar la capacidad de predicción de los brotes de dengue, este proyecto contribuirá a reducir el impacto de esta enfermedad en la población colombiana. Esto permitirá disminuir el número de casos, hospitalizaciones y muertes relacionadas.

1. CONTEXTUALIZACIÓN DEL PROYECTO

1.1 Definición del problema

El dengue es una infección viral transmitida por mosquitos del género *Aedes*. En cuestión de pocas décadas, ha alcanzado proporciones claramente epidémicas en América Latina.

En Colombia, la evolución reciente ha sido especialmente alarmante. Los Boletines Epidemiológicos Semanales del Instituto Nacional de Salud (INS) señalan que en 2023 se notificaron más de 230 000 casos. En 2024, la cifra rebasó los 300 000, lo que representa uno de los picos más altos de los últimos diez años [1-3].

Este aumento constante se hace aún más visible en las tres ciudades analizadas. En Cali, los reportes muestran decenas de miles de casos en los últimos años. Solo en 2023 se superaron los 13 000 casos registrados, y diferentes análisis locales estiman más de 20 000 casos acumulados en ciertos periodos. Estos valores sitúan a la ciudad entre las más afectadas del país [4, 5].

Medellín, por su parte, también ha experimentado un repunte notable. En 2024, la Secretaría de Salud declaró alerta por dengue tras detectar un crecimiento superior al 1000 % respecto al año anterior. En los primeros meses del año se registraron cerca de 800 casos y varios fallecimientos. Estas cifras superan ampliamente las del mismo periodo de 2023 [7, 8].

En Bucaramanga ocurre un comportamiento similar. Según los boletines municipales, para la semana epidemiológica 47 de 2024 ya se habían reportado 5 781 casos. De estos, 5 202 correspondieron a casos sin signos de alarma, 449 con signos de alarma y 4 de dengue grave. Además, se registraron 7 muertes en lo que va del año. Este comportamiento confirma un patrón endémico con momentos de transmisión particularmente intensa [9].

En conjunto, los datos indican que Cali, Medellín y Bucaramanga presentan una circulación sostenida del virus. Se observan brotes repetidos y una tendencia creciente en el número de casos. Este fenómeno ocurre en un entorno marcado por el cambio climático, la urbanización acelerada y la inequidad en el acceso a servicios básicos.

En la práctica, las autoridades sanitarias enfrentan un escenario complejo. La detección temprana y la planificación de intervenciones resultan difíciles debido a la limitada capacidad para anticipar la dinámica semanal de la enfermedad.

Aunque existen sistemas de vigilancia como SIVIGILA (Sistema de Vigilancia en Salud Pública) y múltiples fuentes de datos climáticos y socioeconómicos, su integración en herramientas predictivas sigue siendo limitada. Esto conduce a respuestas mayormente reactivas, como campañas intensivas cuando el brote ya está instalado. También genera un uso ineficiente de los recursos y dificulta la focalización de acciones en las zonas de mayor riesgo.

Ante este panorama, resulta fundamental desarrollar modelos predictivos de dengue basados en técnicas de aprendizaje automático. Estos modelos deben combinar series históricas de casos, información climática y otros determinantes locales. El objetivo es anticipar el comportamiento semanal del dengue en Cali, Medellín y Bucaramanga.

La expectativa es que estas herramientas fortalezcan la vigilancia epidemiológica, orienten mejor las intervenciones durante los periodos críticos y permitan avanzar hacia sistemas de alerta temprana más sólidos para el control del dengue en contextos urbanos complejos.

1.1.1 Planteamiento del problema

El dengue ha representado un desafío creciente para la salud pública en las zonas urbanas de Colombia. Esto se debe no solo al aumento sostenido de casos, sino también a las dificultades para anticipar su aparición con suficiente antelación.

Como se explicó anteriormente, ciudades como Cali, Medellín y Bucaramanga presentan vulnerabilidades particulares frente a esta enfermedad. Estas vulnerabilidades están influenciadas por una combinación de factores climáticos, ambientales y socioeconómicos.

A pesar de la existencia de reportes y registros epidemiológicos detallados, la capacidad de prever brotes con precisión sigue siendo limitada. Esto dificulta una respuesta oportuna y eficaz por parte de las autoridades sanitarias [6].

En la práctica, muchas decisiones relacionadas con el control del dengue se basaban en datos históricos y análisis tradicionales. Si bien estos enfoques aportan información valiosa, no logran captar toda la complejidad del fenómeno.

La propagación del virus no depende de un solo elemento. Depende de una interacción dinámica entre múltiples variables, como temperatura, precipitaciones, densidad poblacional, calidad del entorno urbano y acceso a servicios básicos, entre otras.

Esta naturaleza multifactorial exige el desarrollo de enfoques más avanzados. En particular, se requieren métodos que integren distintas fuentes de información y permitan descubrir patrones no evidentes mediante el análisis convencional.

En ese contexto, el uso de técnicas de aprendizaje automático (*machine learning*) surge como una herramienta con alto potencial. Estos modelos pueden procesar grandes volúmenes de datos, identificar relaciones complejas y generar predicciones a partir de variables diversas.

Aplicar estas metodologías al análisis del dengue ofrece la posibilidad de construir sistemas de alerta temprana más precisos. Además, permite anticipar tanto las zonas como los periodos de mayor riesgo, incluso en contextos urbanos distintos como los de Cali, Medellín y Bucaramanga.

Por tal razón, este proyecto desarrolló modelos predictivos basados en aprendizaje automático. Para ello, integró variables climáticas, ambientales y socioeconómicas con el fin de anticipar brotes de dengue en las tres ciudades mencionadas.

Se espera que estos resultados sirvan como insumo para fortalecer la capacidad de respuesta de los sistemas de salud pública. También podrían mejorar la asignación de recursos y permitir una focalización más precisa de las intervenciones en las zonas vulnerables.

Asimismo, el enfoque permite realizar un seguimiento más dinámico de las estrategias implementadas. Esto facilita evaluar su efectividad y ajustar las políticas de prevención según el nivel de riesgo identificado.

El aporte central del trabajo es la comparación interurbana. Evaluamos el mismo enfoque en tres ciudades con dinámicas distintas, como carga alta persistente, repuntes abruptos y patrón endémico. Esto permite discutir tanto la generalización del modelo como su adaptación local. Este enfoque es valioso porque parte de la literatura trabaja a escala nacional. Además, muchos estudios no contrastan contextos urbanos diferentes.

En definitiva, este trabajo contribuye a una gestión más proactiva basada en datos y evidencia científica. El objetivo es reducir el impacto del dengue mediante decisiones más informadas, focalizadas y oportunas.

1.1.2 Formulación del problema

El dengue sigue siendo una enfermedad endémica en muchas regiones de Colombia. Afecta de manera particular a zonas urbanas con alta densidad poblacional.

Aunque se cuenta con datos sobre casos reportados y condiciones ambientales, aún existe una brecha importante en la capacidad para anticipar con precisión la aparición de nuevos brotes. Esta limitación ha dificultado una respuesta oportuna y efectiva por parte de los sistemas de salud pública. En muchos casos, las autoridades han debido actuar de forma reactiva en lugar de preventiva.

Frente a este desafío, se planteó la siguiente pregunta central de investigación:

¿Cómo utilizar técnicas de aprendizaje automático para desarrollar un modelo predictivo que permita anticipar la ocurrencia de brotes de dengue en ciudades como Cali, Medellín y Bucaramanga, considerando la influencia de factores climáticos, ambientales y socioeconómicos?

Esta pregunta orientadora guió el diseño de una propuesta basada en ciencia de datos. El objetivo fue integrar múltiples variables procedentes de fuentes diversas. Asimismo, se buscó ajustar el modelo a las particularidades de cada entorno urbano.

1.1.3 Sistematización

Para abordar esta problemática de manera estructurada, se formuló un conjunto de subpreguntas. Estas permiten descomponer el problema principal en componentes más específicos y manejables:

- **¿Qué fuentes de información epidemiológicas, climáticas y socioeconómicas resultan más relevantes y confiables para construir una base de datos integrada y coherente?**
- **¿Qué variables tienen mayor peso en la predicción de brotes, y cómo varía su influencia según la ciudad (Cali, Medellín y Bucaramanga)?**

- **¿Qué algoritmos de aprendizaje automático ofrecieron el mejor desempeño en la predicción de brotes en contextos urbanos diversos, y cómo se optimizaron sus parámetros?**
- **¿Qué métricas estadísticas permitieron evaluar con mayor precisión la eficacia, sensibilidad y robustez de los modelos desarrollados?**

Estas preguntas guiaron la planificación metodológica del proyecto. Asimismo, facilitaron la validación de los modelos predictivos. De este modo, se aseguró un enfoque riguroso y contextualizado frente al fenómeno del dengue en el país.

1.2. Objetivos

1.2.1. Objetivo General

Desarrollar un modelo predictivo que permita anticipar la posible ocurrencia de brotes de dengue en Cali, Bucaramanga y Medellín, considerando variables climáticas, ambientales y socioeconómicas.

1.2.2. Objetivos Específicos

- Construir una base de datos integrada y coherente, consolidando información de múltiples fuentes como sistemas de vigilancia epidemiológica, registros climáticos y datos socioeconómicos.

- Determinar cuáles son los factores climáticos, ambientales y socioeconómicos que tienen mayor influencia en la ocurrencia de brotes de dengue en las ciudades de Cali, Medellín y Bucaramanga.

- Construir modelos de aprendizaje automático utilizando datos históricos de casos de dengue, variables climáticas, ambientales y socioeconómicas.

- Evaluar la precisión y robustez de los modelos predictivos utilizando métricas estadísticas apropiadas.

1.3. Marco de Referencia

La predicción de brotes de dengue es un desafío complejo que requiere herramientas analíticas sofisticadas. En este proyecto, exploraremos el potencial de las técnicas de *machine learning* para abordar este problema. A través de un análisis exhaustivo de datos, evaluaremos la eficacia de diversos algoritmos para identificar patrones y generar predicciones precisas. Los resultados de esta investigación permitirán comprender mejor la dinámica de los brotes de dengue. Además, sentarán las bases para desarrollar sistemas de alerta temprana más robustos.

El dengue es una enfermedad viral transmitida principalmente por el mosquito *Aedes aegypti*. Su incidencia ha aumentado en regiones tropicales y subtropicales. Según la Organización Mundial de la Salud (OMS), el dengue afecta a millones de personas cada año, especialmente en América Latina [6]. En Colombia, ciudades como Cali, Medellín y Bucaramanga han registrado brotes recurrentes. Esto se asocia con factores como el cambio climático, la urbanización descontrolada y la falta de infraestructura sanitaria [7].

La aparición y magnitud de los brotes de dengue dependen de factores climáticos, ambientales y socioeconómicos. En particular, variables como temperatura, humedad relativa y precipitación desempeñan un papel central. Estas condiciones regulan el ciclo de vida del mosquito *Aedes aegypti*, su supervivencia y la velocidad de transmisión del virus.

Las temperaturas cálidas y la humedad elevada favorecen la eclosión de huevos y el desarrollo de larvas. También incrementan la actividad de los mosquitos adultos. Por su parte, los periodos lluviosos generan y mantienen criaderos, sobre todo en recipientes y ambientes urbanos vulnerables.

Como se observa en los análisis descriptivos y en los modelos presentados más adelante, pequeñas variaciones en estas condiciones pueden tener efectos relevantes. Este impacto es mayor cuando se combinan con la inercia epidemiológica de semanas

previas. En conjunto, estos elementos se asocian con incrementos significativos en los casos. Por ello, las variables climáticas se consideran un insumo fundamental para la predicción de brotes en Cali, Medellín y Bucaramanga.

De acuerdo con la Organización Mundial de la Salud [6], cerca de la mitad de la población mundial vive en zonas con riesgo de dengue. La carga anual estimada se ubica entre 100 y 400 millones de infecciones. Además, se observa una expansión sostenida hacia nuevas áreas geográficas.

La enfermedad es endémica en más de cien países de regiones tropicales y subtropicales. Se concentra principalmente en áreas urbanas y semiurbanas. Aunque la mayoría de infecciones son asintomáticas o leves, una proporción progresa a formas graves potencialmente mortales.

Dado que no existe un tratamiento antiviral específico, el manejo se basa en medidas de soporte. Entre ellas se incluyen el control del dolor, la hidratación y la detección oportuna de signos de alarma. Por esta razón, las estrategias de prevención se centran en el control del vector. También incluyen la eliminación de criaderos y la protección individual frente a picaduras [6].

En el contexto colombiano, la carga de enfermedad por dengue sigue siendo relevante, pese a la reducción reciente en la notificación. De acuerdo con el Boletín Epidemiológico Semanal del Instituto Nacional de Salud para la semana 34 de 2025, la incidencia nacional es de aproximadamente 289 casos por 100 000 habitantes en riesgo. Además, se reporta circulación simultánea de los cuatro serotipos virales. La mayor afectación se observa en departamentos como Vichada, Guaviare, Meta y Putumayo.

Aunque se ha registrado una disminución cercana al 60 % frente al mismo periodo de 2024, la transmisión persiste. Esta transmisión se mantiene de forma sostenida. Asimismo, se concentra principalmente en población infantil y adolescente. También se observa mayor frecuencia en hombres y en grupos socioeconómicos vulnerables. Esto

refuerza la importancia de integrar información epidemiológica nacional en estrategias de predicción y control.

En los últimos años, el análisis de datos y las técnicas de *machine learning* han ganado relevancia en la predicción de enfermedades infecciosas. El *machine learning* es una rama de la inteligencia artificial. Su objetivo es desarrollar métodos que permitan a las computadoras aprender a partir de datos. Estas herramientas permiten identificar patrones complejos en grandes conjuntos de información. Por ello, pueden apoyar la predicción de brotes con mayor precisión. Además, contribuyen a planificar intervenciones preventivas.

Entre las técnicas más utilizadas se encuentran los modelos basados en árboles de decisión y las redes neuronales artificiales. Estos enfoques permiten capturar relaciones no lineales. También pueden modelar interacciones entre predictores climáticos, ambientales y epidemiológicos.

- **Bosques aleatorios (*Random Forest*)**

El bosque aleatorio es un algoritmo de aprendizaje supervisado. Construye un conjunto de árboles de decisión entrenados con subconjuntos aleatorios de datos y variables. Cada árbol genera una predicción. El resultado final se obtiene mediante el promedio (en regresión) o el voto mayoritario (en clasificación).

Esta estrategia de *ensamble* reduce la varianza y mejora la robustez del modelo. Además, ayuda a evitar el sobreajuste y reduce el efecto del ruido en los datos. En el contexto del dengue, los bosques aleatorios son útiles por varias razones. En particular, manejan bien predictores heterogéneos, como temperatura, humedad, precipitación y semana epidemiológica. Asimismo, permiten estimar la importancia relativa de las variables. Esto aporta interpretabilidad y facilita identificar factores asociados con los brotes.

- **Modelos de gradiente reforzado (*XGBoost*)**

Los modelos de gradiente reforzado, como XGBoost, también se basan en árboles de decisión. Sin embargo, los construyen de forma secuencial. Cada nuevo árbol se ajusta sobre los errores residuales de los árboles anteriores. De este modo, el modelo corrige progresivamente los errores. Por ello, suele alcanzar un desempeño predictivo alto en datos tabulares.

XGBoost incorpora técnicas de regularización. También ofrece manejo eficiente de datos faltantes. Esto lo hace adecuado para bases reales donde la información climática y epidemiológica puede ser incompleta o ruidosa. En este proyecto, XGBoost se emplea como un modelo de referencia de alto desempeño. Su función es comparar su capacidad predictiva frente a otras familias de algoritmos.

- **Redes neuronales densas**

Las redes neuronales densas (*feedforward*) están compuestas por capas de neuronas conectadas. Cada neurona aplica una transformación lineal y luego una función de activación no lineal. Gracias a esta estructura, pueden aproximar funciones complejas. También aprenden patrones que modelos lineales no capturan.

En la predicción de dengue, estas redes permiten combinar información climática, estacional y epidemiológica. Sin embargo, su poder de representación puede aumentar el riesgo de sobreajuste. Por ello, requieren técnicas de regularización, como *dropout*. También necesitan una validación cuidadosa de hiperparámetros.

- **Redes neuronales recurrentes tipo GRU**

Las redes neuronales recurrentes (RNN) están diseñadas para procesar secuencias temporales. Dentro de esta familia, las unidades recurrentes con compuertas (GRU) incorporan una memoria interna. Esto les permite retener información de semanas anteriores. Luego, pueden utilizarla para predecir contagios futuros.

Este enfoque es relevante para el dengue, debido a la inercia epidemiológica de la serie de casos. Es decir, el comportamiento reciente influye en la evolución de los brotes. Las GRU pueden modelar dependencias de corto y mediano plazo. No obstante, su entrenamiento suele ser más exigente computacionalmente. Además, su interpretación es menos directa que la de modelos basados en árboles.

En conjunto, estas familias de modelos permiten abordar la predicción de brotes desde perspectivas complementarias. Los algoritmos basados en árboles (Random Forest y XGBoost) ofrecen buena capacidad predictiva y mayor interpretabilidad. En cambio, las redes neuronales, tanto densas como recurrentes, aportan flexibilidad. En particular, facilitan capturar dinámicas temporales complejas.

En los capítulos 4 y 5 se describe en detalle la implementación de estas técnicas. También se presentan las configuraciones de hiperparámetros utilizadas. Finalmente, se comparan los desempeños mediante métricas como MSE, RMSE y MAE.

En este trabajo, las variables climáticas de temperatura del aire, humedad relativa y precipitación se construyeron a partir de registros horarios. Estos registros provienen de la red de estaciones hidrometeorológicas administradas por el IDEAM. Además, están disponibles en el portal Datos Abiertos Colombia.

En particular, la temperatura ambiente del aire se obtuvo del conjunto de datos “Temperatura Ambiente del Aire”. Este conjunto recopila observaciones horarias en múltiples estaciones del territorio nacional [7].

De manera análoga, la humedad relativa del aire se obtuvo del conjunto de datos “Humedad del Aire”. Este conjunto recopila mediciones horarias a 2 metros de altura en estaciones hidrometeorológicas del IDEAM. También se publica a través de la plataforma Datos Abiertos Colombia [8].

Finalmente, la precipitación se obtuvo del conjunto de datos “Precipitación”. Este conjunto registra la cantidad de lluvia observada en estaciones automáticas y convencionales.

Dichas estaciones se distribuyen en el territorio nacional. La información está disponible en la plataforma Datos Abiertos Colombia (Instituto de Hidrología, Meteorología y Estudios Ambientales [9].

Evaluación de las medidas y modelos

La evaluación de los modelos predictivos será una parte esencial de este proyecto. Para ello, se calcularán indicadores como el **MSE (Error Cuadrático Medio)**. Esta métrica mide la magnitud promedio de los errores al cuadrado. También se calculará el **MAE (Error Absoluto Medio)**, que estima el error promedio sin considerar el signo.

Estas métricas permitirán medir el desempeño en términos de anticipación de brotes. Además, ayudarán a reducir errores relevantes en la práctica, como falsos positivos y falsos negativos. Asimismo, se realizarán validaciones cruzadas para evitar el sobreajuste. Esto permitirá evaluar la robustez del modelo en diferentes contextos.

La comparación entre modelos permitirá identificar el mejor balance entre complejidad y rendimiento. El objetivo es que los enfoques sean aplicables en escenarios reales. Finalmente, los resultados serán interpretados según las condiciones locales de cada ciudad. Esto facilitará su implementación práctica.

1.3.1. Marco Teórico

El dengue es una enfermedad viral transmitida por el mosquito *Aedes aegypti*. Su propagación se ha intensificado en las últimas décadas debido a factores como el cambio climático, la urbanización acelerada y las deficiencias en infraestructura sanitaria.

Este escenario ha generado la necesidad de desarrollar herramientas analíticas más robustas. El objetivo es anticipar la aparición de brotes y fortalecer la capacidad de respuesta de los sistemas de salud pública [10].

En este contexto, la ciencia de datos y, en particular, las técnicas de aprendizaje automático (*machine learning*), se han posicionado como alternativas prometedoras para

abordar este desafío. El aprendizaje automático permite procesar grandes volúmenes de datos e identificar patrones complejos entre variables. Además, no requiere reglas programadas explícitamente.

Su aplicación en salud pública ha permitido avanzar en la predicción de enfermedades transmisibles. En muchos casos, esto supera limitaciones de modelos tradicionales. Diversos algoritmos, como Random Forest y redes neuronales artificiales, han mostrado un rendimiento sobresaliente en la predicción de series temporales. Entre estas series se incluyen los casos de dengue.

Estas metodologías permiten integrar múltiples fuentes de información. Esto favorece una aproximación multifactorial al fenómeno epidemiológico.

En ciudades como Cali, Medellín y Bucaramanga confluyen climas distintos, estructuras urbanas disímiles y niveles heterogéneos de vulnerabilidad social. Por ello, resulta indispensable contar con modelos adaptables que consideren estas diferencias territoriales.

En este marco, el uso de modelos predictivos no solo aporta al conocimiento científico. También ofrece aplicaciones prácticas en la planificación de intervenciones, la distribución de recursos y el diseño de campañas de prevención. Asimismo, puede fortalecer la vigilancia epidemiológica.

La literatura revisada respalda la hipótesis de que el comportamiento del dengue está estrechamente vinculado a variables climáticas. Entre ellas se destacan la **temperatura**, la **humedad relativa** y la **precipitación**. Estas variables condicionan la supervivencia y reproducción del mosquito *Aedes aegypti* [11, 12].

Asimismo, factores **socioeconómicos** y **ambientales** inciden de forma indirecta en la exposición al vector y en la propagación de la enfermedad. Entre ellos se incluyen la densidad poblacional, la calidad del saneamiento básico y el nivel de urbanización [13].

El número de casos de dengue observado semana a semana constituye una **serie temporal**. En este tipo de datos, las observaciones no son independientes, ya que suelen existir **autocorrelación**, **tendencias** y **estacionalidad**. Por ello, el pronóstico requiere considerar explícitamente la estructura temporal.

Desde un enfoque clásico, el análisis de series temporales propone conceptos como la **descomposición** (tendencia–estacionalidad–residuo), la identificación de **rezagos (lags)** y el estudio de la autocorrelación para justificar el uso de variables retardadas. Asimismo, modelos estadísticos como ARIMA/SARIMA o enfoques de suavizamiento exponencial se utilizan como líneas base para pronóstico y comparación metodológica [14].

En el contexto del dengue, el uso de **promedios móviles** y **variables climáticas con rezago** puede interpretarse como una forma de capturar la "inercia" del proceso epidemiológico y la influencia retardada del clima sobre la dinámica del vector.

Además de los enfoques basados en datos, el dengue también puede estudiarse mediante modelos epidemiológicos mecanicistas. Los modelos compartimentales clásicos (por ejemplo, SIR/SEIR) describen la evolución de la enfermedad dividiendo la población en estados como susceptibles, expuestos, infectados y recuperados. Estos modelos permiten interpretar parámetros epidemiológicos clave y analizar escenarios de transmisión [15].

En enfermedades transmitidas por vectores, existen extensiones que representan explícitamente la dinámica humano–mosquito. En estos marcos, variables climáticas como temperatura y precipitación influyen en parámetros biológicos del vector (supervivencia, reproducción y tiempos de incubación), lo cual conecta de manera natural con la evidencia que relaciona clima y transmisión del dengue.

En la práctica, estos modelos requieren parámetros que no siempre están disponibles en tiempo real o con suficiente resolución. Por ello, los enfoques de aprendizaje automático

pueden considerarse **complementarios**: buscan maximizar capacidad predictiva integrando múltiples fuentes (clima, demografía, variables epidemiológicas), sin sustituir la interpretación epidemiológica, sino apoyando la vigilancia y la toma de decisiones.

En este trabajo se adopta un enfoque predictivo basado en aprendizaje automático, complementario a los modelos epidemiológicos clásicos, con énfasis en la capacidad de pronóstico operativo semanal.

1.3.2. Antecedentes

A continuación, se presentan investigaciones y trabajos relacionados con el problema de estudio:

1. Zhao et al. (2020) [16] realizaron un estudio comparativo del rendimiento de dos modelos de predicción de casos de dengue en Colombia: Random Forest (RF) y Redes Neuronales Artificiales (ANN). Los autores encontraron que los modelos RF eran más precisos, especialmente cuando se utilizaban datos agregados a nivel nacional.

El análisis incluyó predictores ambientales y sociodemográficos. Un hallazgo relevante fue que la importancia de los predictores variaba según el horizonte de predicción. Los factores ambientales resultaron más útiles para pronósticos a corto plazo. En contraste, las variables sociodemográficas fueron más relevantes para predicciones a largo plazo.

- **Relación con este proyecto:** Este estudio aporta una base sólida para el desarrollo de modelos de aprendizaje automático aplicados al dengue. Además, resalta la importancia de combinar distintos tipos de predictores.

2. Torres et al. (2023) [17] desarrollaron un sistema de predicción subestacional de casos de dengue en Colombia. Para ello, utilizaron un modelo LSTM Seq2Seq.

El análisis incorporó variables epidemiológicas, meteorológicas, sociales y demográficas recopiladas entre 2007 y 2021.

La capacidad predictiva del modelo se evaluó a nivel nacional, municipal y en las principales ciudades. Los resultados mostraron que la inclusión de variables meteorológicas y demográficas mejoró de manera significativa la precisión de las predicciones.

- **Relación con este proyecto:** Esta investigación respalda el uso de redes neuronales recurrentes en la predicción de enfermedades infecciosas. Asimismo, subraya la relevancia de integrar variables ambientales y sociales.

3. Leung et al. (2023) [18] realizaron una revisión sistemática de los modelos utilizados para predecir brotes de dengue. El estudio analizó 99 modelos provenientes de 64 investigaciones. La mayoría de estos modelos se basaban principalmente en datos climáticos.

Los autores observaron que pocos modelos incluían factores no climáticos o ajustaban adecuadamente los retrasos en la notificación de datos. También señalaron que el uso de validación externa es poco frecuente. Además, destacaron la necesidad de mayor transparencia en la descripción del desarrollo y la evaluación de los modelos.

- **Relación con este proyecto:** Este trabajo proporciona un marco metodológico útil para la evaluación de modelos predictivos. Además, enfatiza la importancia de incorporar factores no climáticos y el uso de datos primarios en tiempo real.

2. CONSTRUCCIÓN DE LA BASE DE DATOS

2.1 Revisión de fuentes, planificación técnica y cronograma

Se llevó a cabo una revisión exhaustiva de la literatura científica y de fuentes institucionales. El objetivo fue identificar los factores determinantes de la incidencia del dengue y las metodologías más empleadas en su predicción.

Este proceso incluyó la consulta de artículos indexados, informes técnicos y bases de datos abiertas. Las fuentes estaban relacionadas con salud pública, cambio climático y variables socioeconómicas.

A partir de esta revisión se identificaron tres grandes categorías de variables relevantes. Estas categorías fueron climáticas, demográficas y sanitarias.

Las variables climáticas incluyeron temperatura, humedad, precipitación y radiación solar. Las variables demográficas consideraron densidad poblacional, distribución etaria y crecimiento urbano. Por su parte, las variables sanitarias incluyeron medidas de control vectorial.

Con base en estos insumos se elaboró una planificación técnica detallada. Esta planificación comprendió los siguientes elementos:

- La definición de objetivos específicos para la recolección, integración y modelamiento de los datos.

- La selección de herramientas tecnológicas para el procesamiento, como Python, Google Colab, *pandas*, *geopandas* y *scikit-learn*.

- El diseño de una arquitectura de datos estructurada por ciudad, con estandarización de formatos y codificación geográfica.

Asimismo, se estableció un cronograma de trabajo dividido en fases. Estas fases incluyeron recopilación, limpieza, consolidación, análisis exploratorio y modelamiento.

Este enfoque permitió un control continuo del avance del proyecto. Además, garantizó una trazabilidad completa de las decisiones metodológicas tomadas durante el proceso.

2.2. Recopilación de datos (Sivirep, IDEAM, DANE)

Se recopilaron datos provenientes de fuentes oficiales de carácter público y verificable:

- **SIVIGILA (Sistema Nacional de Vigilancia en Salud Pública):** registros semanales de casos confirmados de dengue por ciudad y semana epidemiológica.

- **IDEAM (Instituto de Hidrología, Meteorología y Estudios Ambientales):** series temporales de variables climáticas, incluyendo temperatura media, máxima y mínima, humedad relativa, precipitación acumulada y velocidad del viento.

- **DANE (Departamento Administrativo Nacional de Estadística):** indicadores demográficos y socioeconómicos a nivel municipal, como densidad poblacional, acceso a servicios, índice de pobreza y cobertura en salud.

Los datos se descargaron en formatos CSV, Excel y JSON, según la fuente de origen. Posteriormente, se almacenaron en Google Drive para facilitar su integración en entornos colaborativos. Este repositorio también permitió su posterior procesamiento en Python.

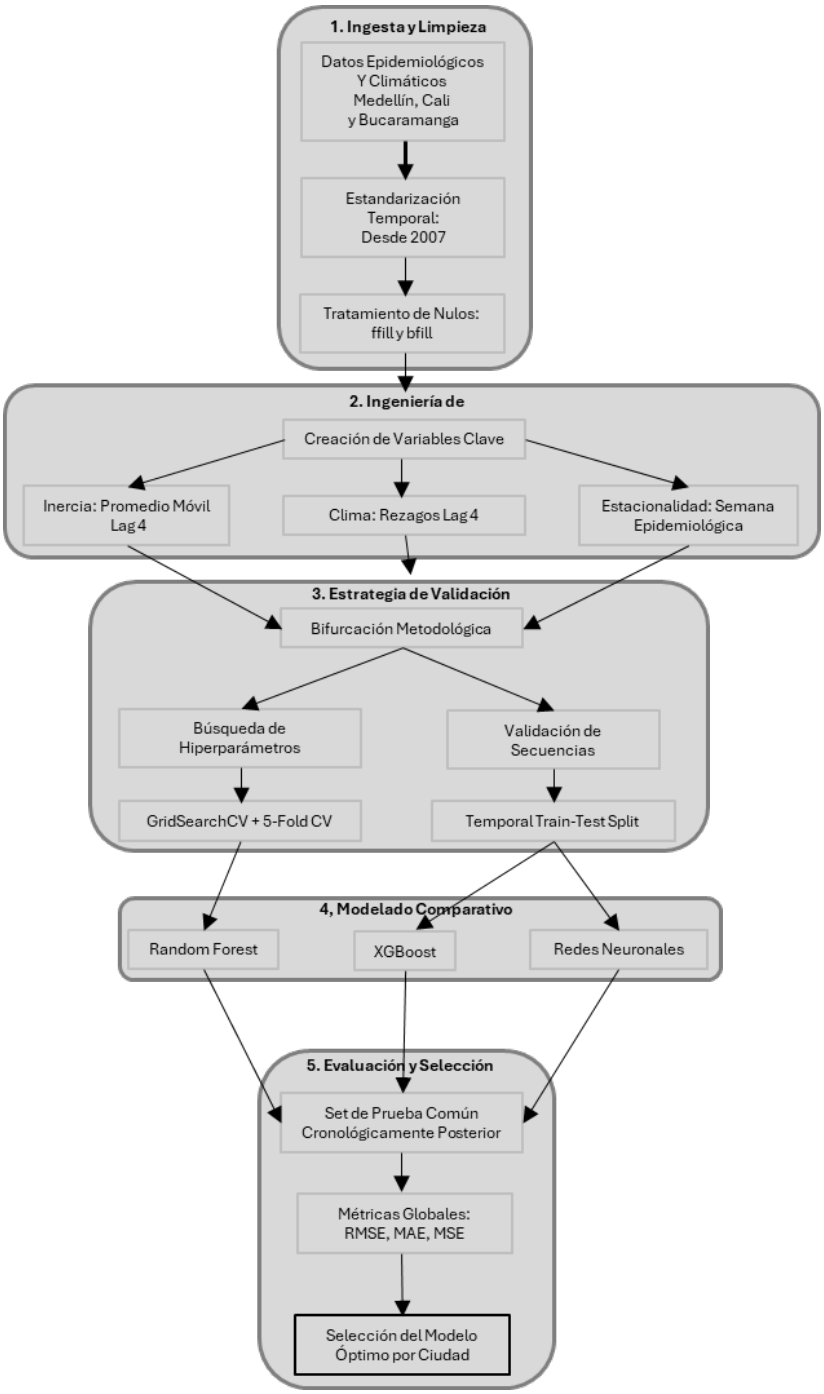
Se garantizó la homogeneidad temporal del conjunto de datos, abarcando el periodo 2008–2024. Asimismo, se aseguró la consistencia espacial a nivel municipal y urbano. En particular, se priorizó la disponibilidad de información con resolución semanal para cada ciudad.

Con el fin de garantizar la trazabilidad del proceso, se definió un único *pipeline* de preparación de datos, compuesto por las siguientes etapas:

- Recolección de datos de SIVIGILA y variables climáticas.
- Agregación de la información a resolución semanal.
- Unión de los conjuntos de datos por semana epidemiológica.
- Limpieza de datos, incluyendo eliminación de duplicados, tratamiento de valores nulos y detección de valores atípicos.
- Construcción de variables rezagadas e indicadores de inercia epidemiológica.
- Partición temporal de los datos para entrenamiento, validación y prueba de los modelos.

Este flujo de procesamiento se resume en la Figura 1 (*Pipeline de preparación de datos*).

Figura 1. Pipeline metodológico para la predicción de dengue por ciudad.



Elaboración propia.

2.3. Limpieza preliminar y consolidación inicial

Una vez obtenidos los datos, se realizó una **limpieza preliminar** orientada a unificar las estructuras y preparar los conjuntos para su análisis. Entre las principales acciones ejecutadas se destacan:

- **Normalización de nombres de variables y columnas**, utilizando el paquete janitor en Python.
- **Conversión y estandarización de formatos de fecha** (YYYY-MM-DD) para garantizar la alineación temporal entre series climáticas y epidemiológicas.
- **Homologación de unidades de medida**, especialmente en variables de precipitación (mm), temperatura (°C) y humedad (%).
- **Codificación unificada de ciudades y departamentos**, mediante los códigos territoriales del DANE.

Posteriormente, se consolidaron las distintas fuentes en **bases únicas por ciudad**, empleando como claves de integración las variables fecha, ciudad y código_municipio. El resultado de este proceso fue una base de datos estructurada con coherencia temporal y espacial, apta para los análisis exploratorios y el modelamiento predictivo posterior.

2.4. Limpieza avanzada: outliers, nulos, duplicados

La siguiente fase consistió en la **depuración avanzada** de las bases de datos. En esta etapa se aplicaron procedimientos estadísticos y visuales para identificar y corregir **valores atípicos**, **datos faltantes** y **registros duplicados**, con el fin de obtener un conjunto final consistente y reproducible para el análisis exploratorio y el modelamiento.

En primer lugar, se realizó la **detección de outliers** mediante el criterio del **rango intercuartílico (IQR)** y la inspección de **boxplots**, lo que permitió identificar observaciones climáticas anómalas. Los casos señalados se revisaron individualmente y

se contrastaron con promedios históricos reportados por el IDEAM para descartar errores instrumentales o de registro.

Posteriormente, se definió un **esquema de tratamiento de outliers** combinando criterios estadísticos (IQR) y criterios físicos plausibles. En el caso de la temperatura, se estableció un rango físicamente razonable (15 °C – 38 °C); los valores fuera de este intervalo se marcaron como atípicos y se trataron como faltantes con el objetivo de imputarlos de forma consistente, en lugar de eliminarlos de manera directa.

En cuanto a los **valores nulos**, se aplicaron estrategias de imputación reproducibles y compatibles con la estructura temporal semanal. Para variables continuas (temperatura, precipitación y humedad), se utilizaron cuatro mecanismos complementarios, según la naturaleza del faltante:

- Imputación por **promedio histórico** de la misma semana epidemiológica en distintos años (especialmente útil cuando faltaba una semana completa y no era posible aplicar métodos locales como ffill/bfill),
- **Interpolación spline** de orden 3 cuando existían tramos de faltantes prolongados, y
- Relleno hacia adelante (**ffill**) y
- Relleno hacia atrás (**bfill**) para asegurar continuidad semanal y eliminar NaNs residuales. En variables categóricas con baja proporción de nulos, se aplicó una eliminación controlada.

Adicionalmente, se realizó la **eliminación de duplicados** mediante verificaciones automatizadas (por ejemplo, drop_duplicates), garantizando la unicidad de cada registro por ciudad y semana epidemiológica.

Para monitorear el impacto del proceso, se construyó un **diagnóstico de calidad de datos** por ciudad y por variable, cuantificando el **número de semanas con valores faltantes (NaN)** y su **porcentaje respecto al total de semanas** en cada serie. Antes de

la imputación, la proporción de faltantes climáticos fue heterogénea entre ciudades. Por ejemplo, en **Cali** se identificaron **68 semanas con temperatura faltante (≈ 9,9 %)**, **61 semanas con humedad faltante (≈ 8,9 %)** y **9 semanas con precipitación faltante (≈ 1,3 %)**. En **Medellín**, se observaron **95 semanas con temperatura faltante (≈ 13,8 %)**, **109 semanas con humedad faltante (≈ 15,8 %)** y **9 semanas con precipitación faltante (≈ 1,3 %)**. En **Bucaramanga**, los faltantes fueron más pronunciados, con **289 semanas con temperatura faltante (≈ 42,0 %)**, **305 semanas con humedad faltante (≈ 44,3 %)** y **10 semanas con precipitación faltante (≈ 1,5 %)**. Con base en este diagnóstico, se definieron reglas de imputación consistentes y aplicables de manera uniforme, priorizando métodos reproducibles y coherentes con la estructura temporal semanal.

Para los modelos de redes neuronales (GRU), se aplicó **escalamiento Min-Max (0–1)**. El escalador se ajustó únicamente con el conjunto de entrenamiento y luego se aplicó a validación y prueba, evitando fuga de información (data leakage). Este paso permitió estabilizar el entrenamiento y prevenir que variables de mayor magnitud (por ejemplo, contagios) dominaran el gradiente frente a variables con variación más pequeña pero relevantes (como cambios sutiles en la temperatura). En contraste, los modelos basados en árboles no requirieron normalización.

Finalmente, se garantizó la **trazabilidad del preprocesamiento** mediante documentación en bitácoras y conservación de versiones intermedias de los datos antes y después de cada etapa. Como resultado, se obtuvo una base depurada y estructurada, lista para el análisis exploratorio y la implementación de modelos de aprendizaje automático orientados a la predicción de brotes.

Como recomendación de robustez, se propone realizar un **análisis de sensibilidad** comparando el desempeño de los modelos bajo distintas decisiones de preprocesamiento, por ejemplo: imputación por promedio semanal vs. interpolación, y detección/tratamiento de outliers mediante rango físico vs. IQR. Esto permitiría confirmar que las conclusiones no dependen de una única elección metodológica.

3. FACTORES CLIMÁTICOS, AMBIENTALES Y SOCIOECONÓMICOS CON MAYOR INFLUENCIA

3.1. Introducción

Una vez consolidada la base de datos integrada, el siguiente paso fue identificar los factores climáticos, ambientales y socioeconómicos con mayor influencia en la ocurrencia de brotes de dengue. El análisis se centró en las ciudades de **Cali, Medellín y Bucaramanga**.

El propósito de esta fase fue caracterizar el comportamiento histórico del dengue. Para ello, se examinó su relación con los determinantes más relevantes. El análisis se realizó mediante técnicas de **análisis exploratorio de datos (EDA)**, así como **análisis descriptivo** y **correlacional**.

3.2. Análisis descriptivo de los casos de dengue

El punto de partida del análisis fue la construcción de series temporales semanales de casos de dengue para Cali, Medellín y Bucaramanga. Para ello, se agregaron los registros individuales por año calendario y semana epidemiológica.

Esta organización permite observar la evolución temporal de la enfermedad en cada ciudad. Asimismo, facilita la comparación entre ellas bajo una misma unidad de medida.

En Cali se observan los niveles más altos de transmisión a lo largo del periodo de estudio. Se identifican picos epidémicos intensos, concentrados en determinados años, caracterizados por incrementos abruptos en el número de casos semanales.

Entre estos picos aparecen intervalos de menor actividad. No obstante, se mantiene una circulación persistente del virus. Esto indica una transmisión sostenida incluso fuera de los periodos de brote.

Medellín presenta una dinámica diferente. La magnitud absoluta de los casos es menor

que en Cali. Sin embargo, también se identifican años claramente epidémicos.

Estos años están separados por intervalos prolongados con baja notificación. Este patrón sugiere una transmisión más intermitente, en la que coexisten periodos de alta incidencia y fases de relativa calma epidemiológica.

En Bucaramanga la situación es intermedia. Los máximos de casos semanales no alcanzan los valores observados en Cali. Sin embargo, la recurrencia de brotes a lo largo de los años indica una carga importante de dengue.

Los incrementos de casos se concentran en bloques específicos de semanas dentro del año. Además, tienden a repetirse en varios ciclos. Esto refuerza la idea de una transmisión recurrente, aunque de menor magnitud relativa.

Además de describir las tendencias temporales y las variaciones anuales, este análisis descriptivo sirve como punto de partida para el análisis climático. En capítulos posteriores se presentan matrices de correlación entre los casos de dengue y variables ambientales rezagadas.

Estas matrices aportan una primera evidencia cuantitativa sobre los factores climáticos asociados a los aumentos de casos en cada ciudad. Asimismo, orientan la selección de predictores para los modelos de aprendizaje automático.

3.3. Relación con variables climáticas

Para explorar la relación entre los casos semanales de dengue (variable **Contagios**) y las condiciones ambientales, se construyeron matrices de correlación. Estas matrices utilizaron variables climáticas promedio semanales con un rezago de cuatro semanas. Se incluyeron humedad relativa, temperatura mínima, temperatura promedio, temperatura máxima y precipitación acumulada.

Este rezago busca aproximar el tiempo requerido para que cambios en temperatura, humedad o lluvia se traduzcan en variaciones en la incidencia. Esta decisión se

fundamenta en el ciclo de vida del *Aedes aegypti* y en los periodos de incubación viral.

En las tres ciudades se observaron correlaciones lineales de magnitud moderada o baja. Esto sugiere que el efecto del clima sobre el dengue es complejo. Además, probablemente no es lineal. Sin embargo, se identificaron patrones útiles para la selección de predictores en los modelos del Capítulo 4.

En **Cali**, la temperatura mínima promedio semanal con rezago de cuatro semanas presentó la asociación más clara con los casos de dengue. La correlación fue positiva y cercana a **r ≈ 0,39**. Le siguieron la temperatura promedio (**r ≈ 0,31**) y la temperatura máxima (**r ≈ 0,25**).

En contraste, la humedad relativa mostró una correlación negativa débil (**r ≈ –0,15**). La precipitación, por su parte, prácticamente no presentó relación lineal con los contagios (**r ≈ –0,04**).

Además, se evidenció una fuerte colinealidad entre las distintas medidas de temperatura. También se observó una correlación negativa moderada entre temperatura y humedad. Este patrón refuerza la necesidad de usar un número reducido de predictores climáticos para evitar redundancias.

Con base en estos resultados y en los patrones estacionales descritos en el Capítulo 4, se seleccionó la temperatura mínima promedio semanal rezagada cuatro semanas. Esta variable se adoptó como predictor climático principal para Cali. Su ventaja es que resume de manera parsimoniosa condiciones térmicas favorables para la supervivencia del vector.

En **Medellín**, la matriz de correlación mostró relaciones lineales simples muy débiles. Las correlaciones entre contagios y cada variable climática se ubicaron cerca de cero. En general, oscilaron entre aproximadamente **–0,10 y 0,04**.

No obstante, entre las propias variables climáticas se observó colinealidad marcada. Esta

colinealidad fue particularmente alta entre las tres medidas de temperatura, con coeficientes superiores a **0,73**.

En este contexto, y para mantener un marco homogéneo con Cali, se conservó la temperatura mínima promedio semanal rezagada cuatro semanas. Se utilizó como predictor climático principal. Esta variable captura de manera parsimoniosa el componente estacional de Medellín. Además, se integra de forma coherente con los escenarios descritos en la Sección 4.3.

Sin embargo, los resultados del Capítulo 5 muestran que, en esta ciudad, la inercia epidemiológica tiene un peso explicativo mayor que el componente climático.

En **Bucaramanga**, las correlaciones simples entre casos y variables climáticas también fueron de baja magnitud. La humedad relativa promedio semanal con rezago de cuatro semanas presentó la correlación más alta, aunque débil (**r ≈ 0,11**).

Las correlaciones con las medidas de temperatura y con la precipitación se mantuvieron cercanas a cero. En general, variaron entre aproximadamente **–0,00 y 0,10**.

De nuevo, se observó una fuerte colinealidad entre temperatura mínima, promedio y máxima, con coeficientes superiores a **0,73**. Este resultado desaconseja incluirlas simultáneamente en los modelos.

Sin embargo, el análisis de patrones estacionales y los diagramas de dispersión sugirieron un comportamiento diferente para la humedad. En particular, la humedad relativa mostró mayor variabilidad y una correspondencia más visible con los periodos de incremento de casos. Por ello, se seleccionó como variable climática principal rezagada cuatro semanas en los escenarios de modelado para Bucaramanga.

En conjunto, estos resultados sugieren que las variables climáticas no explican por sí solas la dinámica de los brotes. No obstante, aportan información relevante sobre el contexto ambiental que facilita o limita la transmisión del dengue.

La baja magnitud de las correlaciones lineales, junto con la presencia de rezagos, sugiere una relación no lineal y dependiente del tiempo. Por esta razón, en los modelos predictivos se complementó el componente climático con estacionalidad (semana epidemiológica). También se incluyó un término de inercia epidemiológica. Este término se construyó mediante el promedio móvil de cuatro semanas de contagios.

Esta integración, desarrollada en los Capítulos 4 y 5, permite representar de forma más realista la interacción entre el contexto climático y la historia reciente de transmisión en cada ciudad.

3.4. Factores ambientales y socioeconómicos

Además de las condiciones climáticas, el análisis incorporó variables ambientales y socioeconómicas. Estas variables se obtuvieron del DANE y de registros municipales. El objetivo fue explorar su influencia indirecta en la dinámica del dengue.

- **Densidad poblacional:** se observaron correlaciones positivas con los casos de dengue en las tres ciudades. El valor promedio fue **r ≈ 0,45**. Este resultado sugiere que una mayor concentración urbana incrementa la exposición al vector y dificulta las labores de control.

- **Índice de Pobreza Multidimensional (IPM):** presentó una asociación más marcada en Cali y Bucaramanga, con **r ≈ 0,40**. Esto indica que zonas con menor infraestructura sanitaria y deficiencias en el acceso a agua potable y recolección de residuos tienden a concentrar mayor incidencia.

- **Nivel educativo promedio y cobertura en salud:** ambas variables mostraron correlaciones negativas débiles, con valores cercanos a **r ≈ –0,25**. Este patrón puede interpretarse como un posible efecto protector. Dicho efecto estaría asociado a un mayor acceso a información preventiva y a atención médica oportuna.

En conjunto, estos resultados refuerzan la idea de que la dinámica del dengue no depende únicamente de condiciones ambientales favorables. También está modulada por determinantes sociales estructurales. Estos factores influyen en la vulnerabilidad de las comunidades urbanas frente a la enfermedad [7].

3.5. Integración de hallazgos y patrones comunes

Al comparar los resultados entre las tres ciudades, emergen patrones comunes que permiten comprender mejor los determinantes del dengue:

1. **Temperatura y humedad** son las variables climáticas más influyentes en todas las ciudades. Su efecto es especialmente relevante en los periodos previos a los brotes.

2. **La precipitación** actúa como un desencadenante indirecto. Favorece la generación de criaderos potenciales, aunque su impacto depende del contexto urbano y del manejo del saneamiento.

3. **Los factores socioeconómicos** amplifican el riesgo de transmisión. Este efecto es más evidente en zonas con alta densidad poblacional y menor cobertura de servicios públicos.

4. **Las relaciones temporales con desfase (lag)** entre variables climáticas y casos de dengue fueron claramente identificadas. Estos desfases serán aprovechados en los modelos predictivos presentados en el Capítulo 4.

En conjunto, estos resultados respaldan la hipótesis de que los brotes de dengue responden a un sistema **multifactorial**. El clima establece las condiciones biológicas para la propagación del vector. Por su parte, los factores sociales y urbanos determinan la magnitud del impacto en la población [7].

4. MODELADO PREDICTIVO DE BROTES DE DENGUE CON APRENDIZAJE AUTOMÁTICO Y VARIABLES CLIMATICAS EN CALI, MEDELLÍN Y BUCARAMANGA

4.1. Formulación del problema de predicción

4.1.1. Objetivo del modelado predictivo

En este capítulo se describe el proceso de construcción de modelos de aprendizaje automático para predecir el número semanal de casos de dengue. El análisis se centra en las ciudades de Cali, Medellín y Bucaramanga.

Los modelos utilizan como insumos datos históricos de vigilancia epidemiológica y variables climáticas. El propósito es **estimar, para cada semana epidemiológica, el nivel esperado de contagios**. Esta estimación se basa en la información climática de semanas previas y en la inercia reciente de la propia serie de dengue.

4.1.2. Horizonte temporal y periodo de estudio

El horizonte temporal del estudio abarcó, para Cali y Medellín, el periodo 2007–2019, mientras que para Bucaramanga se contó con información desde 2006–2019. A partir de los registros individuales del sistema de vigilancia, se construyeron series semanales de casos de dengue, indexadas por año calendario y semana epidemiológica. Se decidió no incluir años posteriores a 2019 con el fin de evitar la incorporación de choques externos no relacionados con la dinámica epidemiológica del dengue, en particular los cambios abruptos asociados a la pandemia de COVID-19 (alteraciones en movilidad, acceso a servicios, patrones de consulta y notificación), que podrían introducir sesgos estructurales y dificultar la comparabilidad temporal del modelo sin añadir variables externas adicionales.

Para el modelado se adoptó una partición temporal coherente con el enfoque de series de tiempo. Las semanas comprendidas entre 2007 y 2017 se utilizaron como conjunto de

entrenamiento (o 2006 – 2017 en Bucaramanga), mientras que las semanas de 2018 – 2019 se reservaron como conjunto de prueba. Esta estrategia permitió evaluar la capacidad de generalización de los modelos en un periodo reciente no utilizado durante el ajuste y representativo de condiciones operativas previas a 2020.

La resolución semanal se eligió porque coincide con el esquema de reporte y toma de decisiones de la vigilancia epidemiológica. Este nivel de agregación reduce el ruido diario, facilita la integración con variables climáticas y permite generar productos operativos para sistemas de alerta temprana. En términos prácticos, el modelo se alimenta de información reciente (incluyendo un rezago climático de cuatro semanas y un término de inercia epidemiológica basado en promedios móviles) para estimar el nivel esperado de casos en el corto plazo. Este horizonte es relevante en salud pública porque permite anticipar acciones con suficiente tiempo de preparación, tales como reforzar el control vectorial, planificar recursos asistenciales y orientar campañas de comunicación del riesgo.

Asimismo, aunque el estudio utiliza un rezago de cuatro semanas por sustento biológico y empírico, el horizonte predictivo útil en un sistema real puede implementarse como un esquema de pronóstico de 1 a varias semanas (por ejemplo, 1 a 4 semanas), manteniendo un balance entre anticipación operativa y precisión, y aprovechando que muchas intervenciones requieren al menos algunos días o semanas para desplegarse.

Tabla 1. Estructura de los datos por ciudad: tamaño de muestra y partición entrenamiento–prueba.

Ciudad	N° de semanas (observaciones)	Registros entrenamiento	Registros prueba	Años de entrenamiento	Años de prueba
Cali	689	583	106	2007–2017	2018–2019
Medellín	689	583	106	2007–2017	2018–2019
Bucaramanga	742	636	106	2006–2017	2018–2019

Elaboración propia.

La Tabla 1 sintetizó el número total de semanas disponibles por ciudad y el tamaño de las particiones de entrenamiento y prueba. Cali y Medellín presentaron un número similar de observaciones, mientras que Bucaramanga aportó un mayor número de semanas debido a la extensión temporal de su serie. En todos los casos, la proporción de datos asignada al entrenamiento fue suficiente para ajustar modelos complejos y dejó un bloque final de dos años para la evaluación independiente.

4.1.3. Metodología general de datos y modelado predictivo

El proceso metodológico integró de manera secuencial la construcción de la base de datos, la selección de variables explicativas y la implementación de modelos de aprendizaje automático. En primer lugar, tal como se describió en el Capítulo 2, se consolidaron fuentes epidemiológicas y climáticas para Cali, Medellín y Bucaramanga.

A partir de los registros individuales del SIVIGILA se obtuvieron los casos confirmados de

dengue. Del IDEAM y de portales de datos abiertos se descargaron series de temperatura mínima, media y máxima, humedad relativa y precipitación acumulada. Todas las fuentes se homogeneizaron en formatos de fecha, unidades de medida y códigos territoriales. Posteriormente, se integraron en bases únicas por ciudad.

Luego se desarrolló una fase de depuración y preparación de datos. Esta fase incluyó detección de valores atípicos, tratamiento de datos faltantes y eliminación de registros duplicados. Para las variables climáticas se aplicaron reglas basadas en rangos plausibles. También se utilizó análisis de *boxplots* y comparación con promedios históricos.

Los valores nulos se imputaron mediante interpolación temporal o promedios móviles, según la naturaleza de cada serie. Este proceso permitió disponer de una base integrada y consistente en términos temporales y espaciales. Además, dejó el conjunto apto para el análisis exploratorio y el modelamiento predictivo.

En una segunda etapa, descrita en detalle en los Capítulos 3 y 4, las series diarias se agregaron a resolución semanal. Para cada ciudad se construyó una estructura tipo panel. Esta estructura incluyó el número de contagios y variables climáticas asociadas a cada semana epidemiológica.

A partir de esta base se derivaron variables rezagadas e indicadores de inercia epidemiológica. El objetivo fue capturar los retrasos entre condiciones ambientales y aparición de casos. En particular, se calcularon promedios semanales de temperatura, humedad y precipitación con rezago de cuatro semanas. Asimismo, se construyó un promedio móvil de cuatro semanas del número de contagios.

La elección de este rezago se sustentó en las matrices de correlación entre contagios y clima. En dichas matrices, los valores máximos o más consistentes se observaron alrededor de cuatro semanas. Este resultado es coherente con la dinámica biológica del vector y del virus.

La tercera etapa correspondió al diseño de escenarios de modelado y a la selección de técnicas de aprendizaje automático. Se definieron dos configuraciones principales de predictores:

- un escenario basado en variables climáticas y semana epidemiológica, que representa el aporte del clima a la predicción; y
- un escenario ampliado que incorpora, además, la inercia epidemiológica mediante el promedio móvil de cuatro semanas de contagios.

Estas configuraciones se adaptaron a las particularidades de cada ciudad. En Cali y Medellín se utilizó como variable climática principal la temperatura mínima. En Bucaramanga se utilizó la humedad relativa. Esta selección se realizó con base en la tabla de correlación de Pearson estimada para cada ciudad.

Finalmente, sobre estas bases semanales se entrenaron y compararon tres familias de modelos: bosques aleatorios, XGBoost y redes neuronales. En este último grupo se incluyeron redes densas y redes recurrentes tipo GRU.

Los modelos se ajustaron utilizando el conjunto de entrenamiento definido en la Sección 4.1.2. Para seleccionar configuraciones con mejor desempeño, se emplearon particiones internas de validación o esquemas de validación cruzada.

La evaluación se realizó sobre un bloque temporal independiente (2018–2019). Se utilizaron métricas de error cuadrático medio (MSE), raíz del error cuadrático medio (RMSE) y error absoluto medio (MAE), como se detalla en el Capítulo 5. Esta estrategia permitió comparar de manera consistente el rendimiento de técnicas y escenarios. Además, facilitó evaluar su capacidad para predecir brotes de dengue en las tres ciudades estudiadas.

En los tres casos se utilizó un esquema de validación adecuado para series temporales. Este esquema se aplicó al ajuste de hiperparámetros del modelo Random Forest mediante *GridSearchCV*.

En lugar de emplear validación *k-fold* con particiones aleatorias, se usaron particiones por bloques temporales. Entre ellas se incluyeron enfoques como *TimeSeriesSplit* o *folds* cronológicos sin barajar. Este diseño garantizó que cada entrenamiento se realizara únicamente con semanas anteriores a las semanas de validación. De este modo, se evitó la fuga de información (*data leakage*).

Para XGBoost y las redes neuronales se utilizó un esquema de entrenamiento – validación basado en *train_test_split* sin aleatorización (*shuffle* = *False*). Además, se aplicó *early stopping* durante el entrenamiento. La evaluación final se llevó a cabo sobre un periodo temporal independiente, correspondiente a 2018 – 2019.

La validación cruzada tipo *k-fold* no se aplicó de forma sistemática a todos los modelos. Su uso se limitó específicamente al Random Forest, durante la etapa de ajuste de hiperparámetros.

4.2. Construcción de la base de datos semanal para el modelado

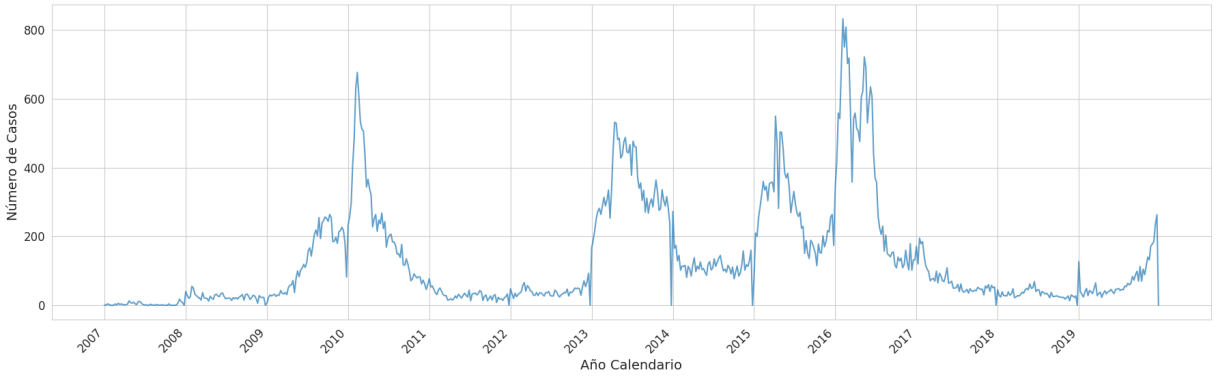
4.2.1. Agregación de casos de dengue y estacionalidad por ciudad

Los registros individuales de dengue se agregaron por combinación de **año calendario** y **semana epidemiológica**. De este modo, se obtuvo el número de contagios semanales en cada ciudad. Esta agregación permitió trabajar con una unidad de análisis homogénea entre fuentes. Además, resultó compatible con la resolución de las series climáticas.

Las Figuras 2, 3 y 4 presentan las series de casos semanales de dengue para Cali, Medellín y Bucaramanga. Estas series se construyeron a partir de la agregación por año calendario y semana epidemiológica.

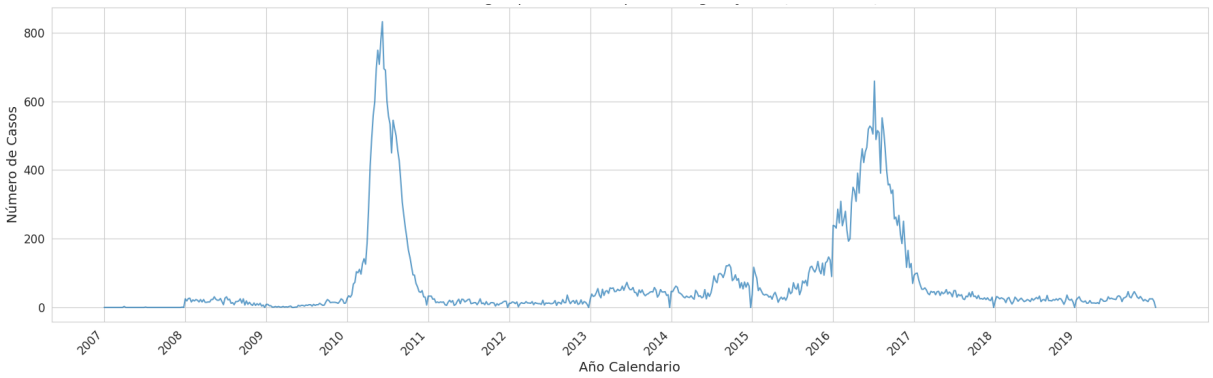
Una descripción detallada de las diferencias en magnitud, recurrencia y patrón temporal de los brotes entre las tres ciudades se presenta en la Sección 3.2. En el presente capítulo, dichas series se utilizan como insumo para la construcción de la base de datos semanal. Sobre esta base se entrenan los modelos predictivos.

Figura 2. Casos semanales de dengue en Cali (2007–2019).



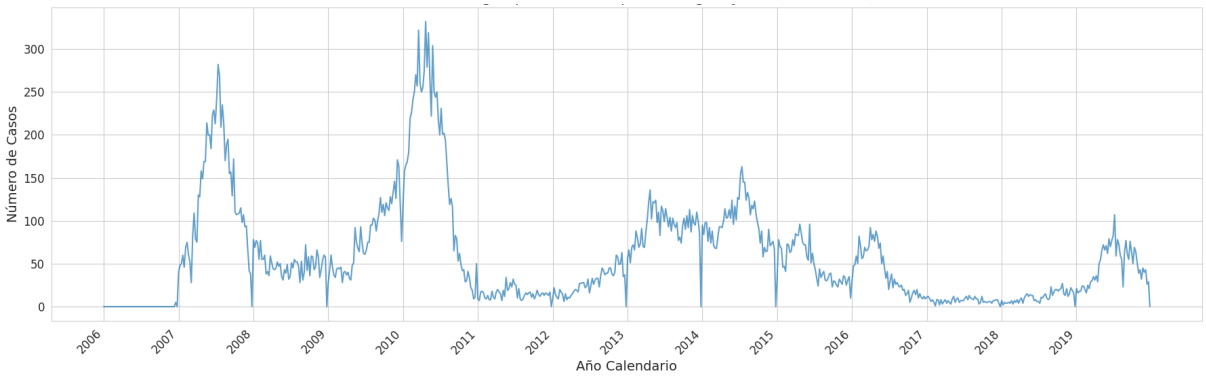
Elaboración propia.

Figura 3. Casos semanales de dengue en Medellín (2007–2019).



Elaboración propia.

Figura 4. Casos semanales de dengue en Bucaramanga (2006–2019).

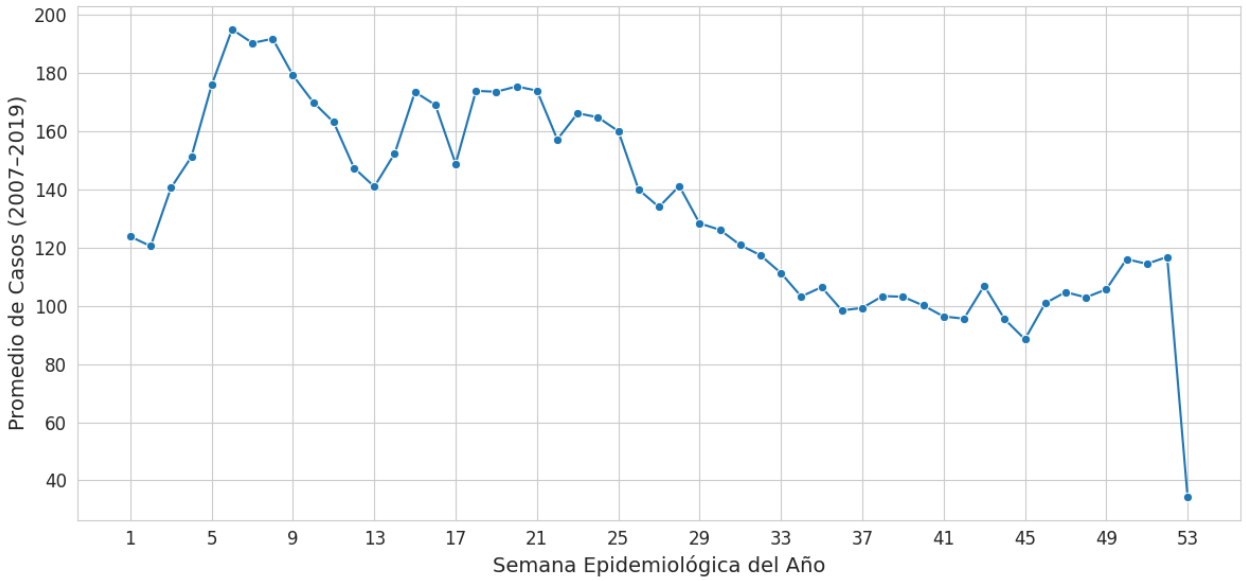


Elaboración propia.

Con el fin de caracterizar la estacionalidad, se calculó en las Figuras 5, 6 y 7 el promedio de casos para cada ciudad y por semana epidemiológica a lo largo de todo el periodo de estudio. Este cálculo permitió identificar el ciclo anual del dengue.

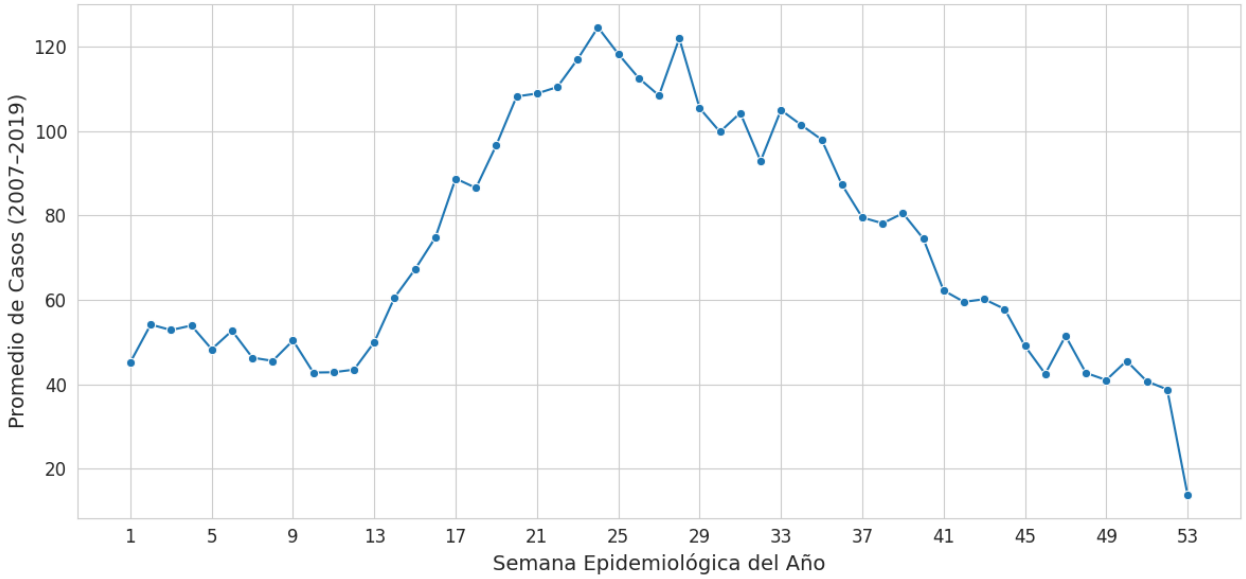
Estos patrones estacionales se utilizaron posteriormente para justificar la inclusión de la semana epidemiológica como variable numérica. Dicha variable se incorporó en los escenarios de modelado descritos en la Sección 4.3.

Figura 5. Patrón estacional promedio de casos de dengue en Cali.



Elaboración propia.

Figura 6. Patrón estacional promedio de casos de dengue en Medellín.



Elaboración propia.

Figura 7. Patrón estacional promedio de casos de dengue en Bucaramanga.

Elaboración propia.

De forma complementaria, la Tabla 2 presenta los estadísticos descriptivos de la variable objetivo para cada ciudad. Se incluyen la media, la desviación estándar, la mediana, los valores mínimos y máximos, así como el coeficiente de variación.

Estos resultados permiten constatar que Cali registró la media semanal más alta y los valores máximos más elevados. Este patrón es coherente con los grandes brotes observados en la Figura 2.

Bucaramanga presentó medias más bajas y epidemias de menor magnitud. En contraste, Medellín se caracterizó por una media intermedia y un coeficiente de variación más alto. Esto evidencia una dinámica más irregular, con largos periodos de baja transmisión alternados con episodios epidémicos intensos.

Tabla 2. Estadísticos descriptivos de los casos semanales de dengue por ciudad.

Ciudad	Media	DE	Mediana	Mínimo	Máximo	CV (%)
Cali	133.8	155.0	64	0	833	115.8
Medellín	72.6	132.3	25	0	832	182.3
Bucaramanga	58.1	60.0	41	0	332	103.3

Elaboración propia.

4.2.2. Integración de variables climáticas y patrones estacionales

La información climática se integró mediante la construcción de series semanales de:

- Humedad relativa promedio,
- Temperatura mínima, promedio y máxima, y
- Precipitación acumulada semanal.

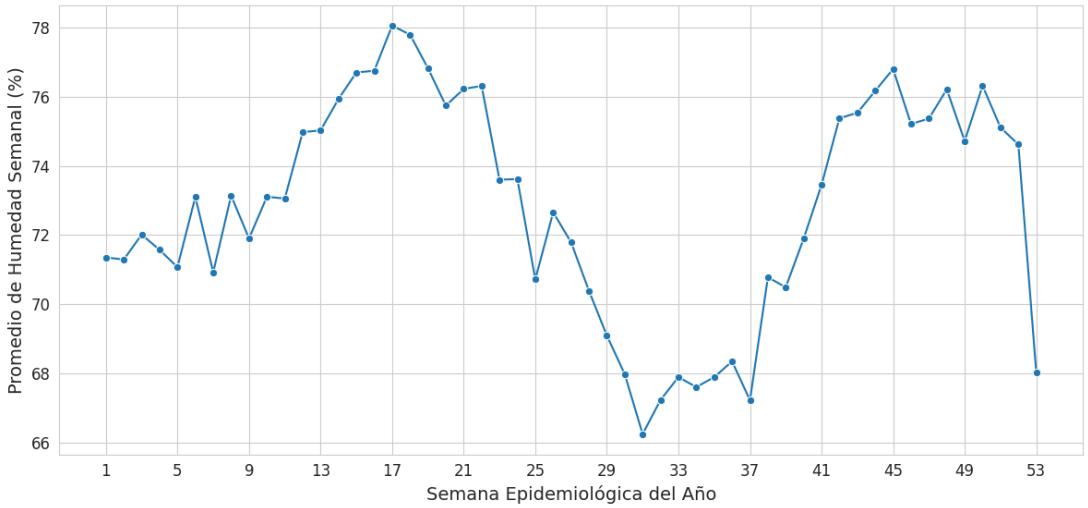
Para ello, las series diarias de las estaciones meteorológicas más representativas de cada ciudad se transformaron a frecuencia semanal. Según la naturaleza de cada variable, se calcularon promedios o acumulados.

En Cali, las Figuras 8 a 12 presentan los patrones estacionales promedio multianuales de humedad, temperaturas mínima, promedio y máxima, y precipitación. Se observó un incremento de la humedad y de la precipitación entre las semanas 10 y 20, asociado a uno de los periodos lluviosos.

Posteriormente, se registró una reducción relativa durante los meses intermedios del año. Un segundo aumento se observó hacia las semanas 40 a 48.

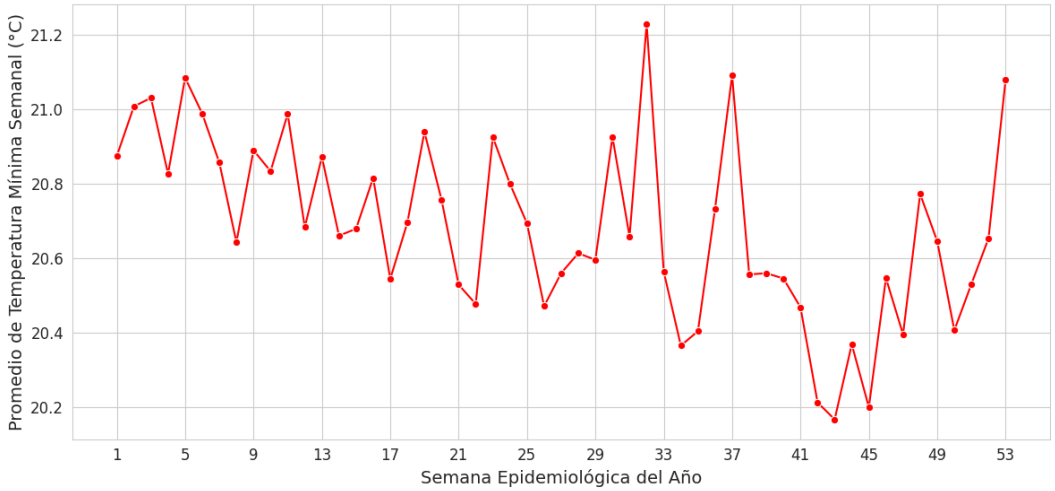
Las temperaturas oscilaron dentro de rangos relativamente estrechos. No obstante, se identificaron ligeras disminuciones durante los periodos más lluviosos.

Figura 8. Patrón estacional promedio de humedad relativa en Cali.



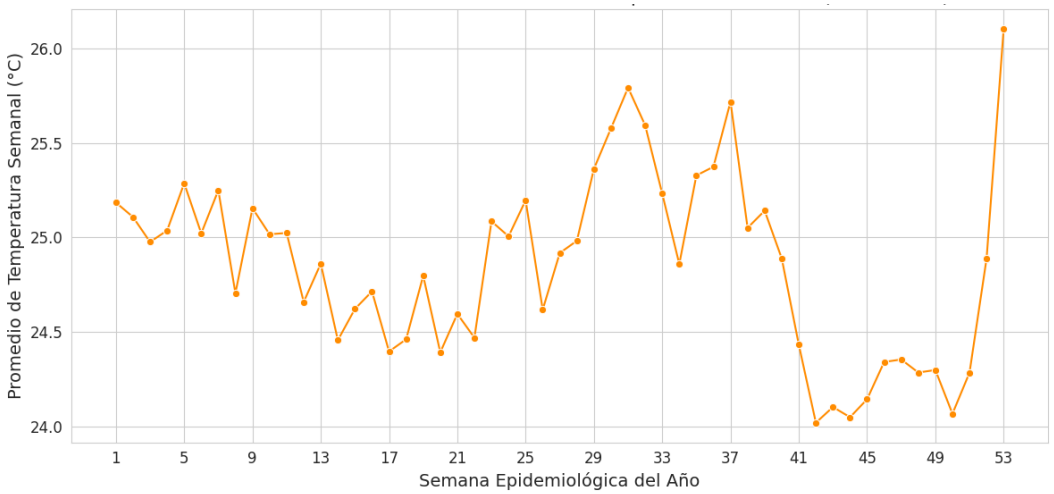
Elaboración propia.

Figura 9. Patrón estacional promedio de temperatura mínima en Cali.



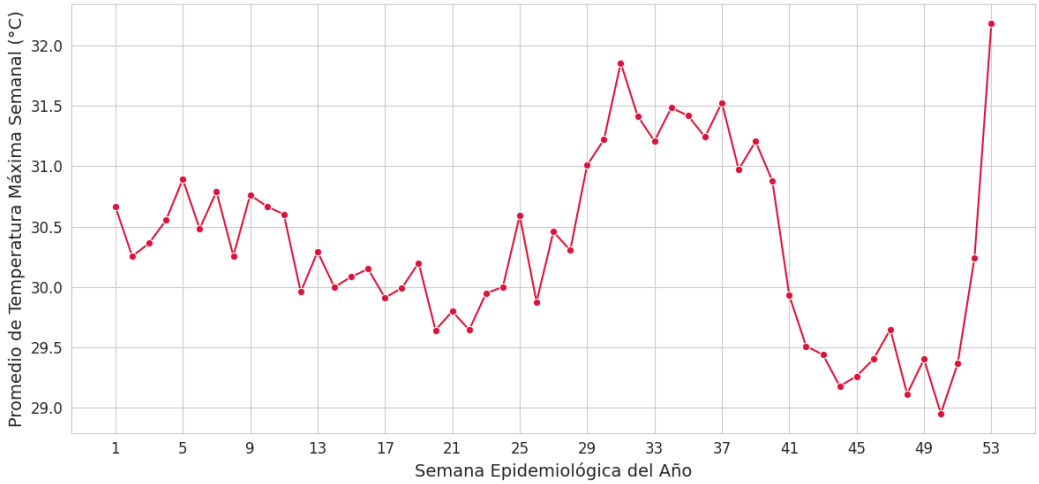
Elaboración propia.

Figura 10. Patrón estacional promedio de temperatura promedio en Cali.



Elaboración propia.

Figura 1. Patrón estacional promedio de temperatura máxima en Cali.



Elaboración propia.

Figura 2. Patrón estacional promedio de precipitación en Cali.

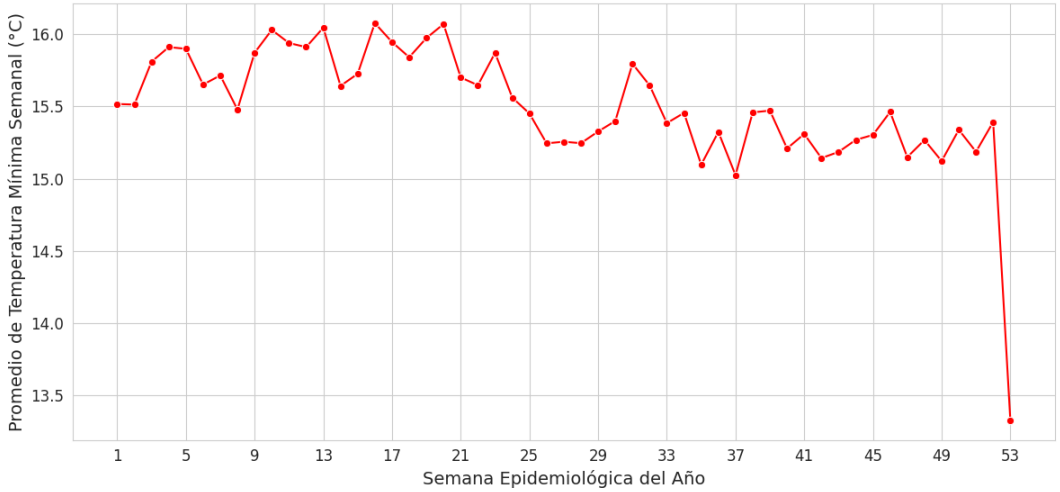
Elaboración propia.

En Medellín, las Figuras 13 a 17 reflejaron la coexistencia de dos estaciones lluviosas bien definidas, con picos de precipitación y humedad en torno a las semanas 10–20 y 35–45. Las temperaturas mínimas se mantuvieron en promedio por debajo de las registradas en Cali y Bucaramanga, mientras que las variaciones de las temperaturas promedio y máximas fueron más moderadas.

Figura 3. Patrón estacional promedio de humedad relativa en Medellín.

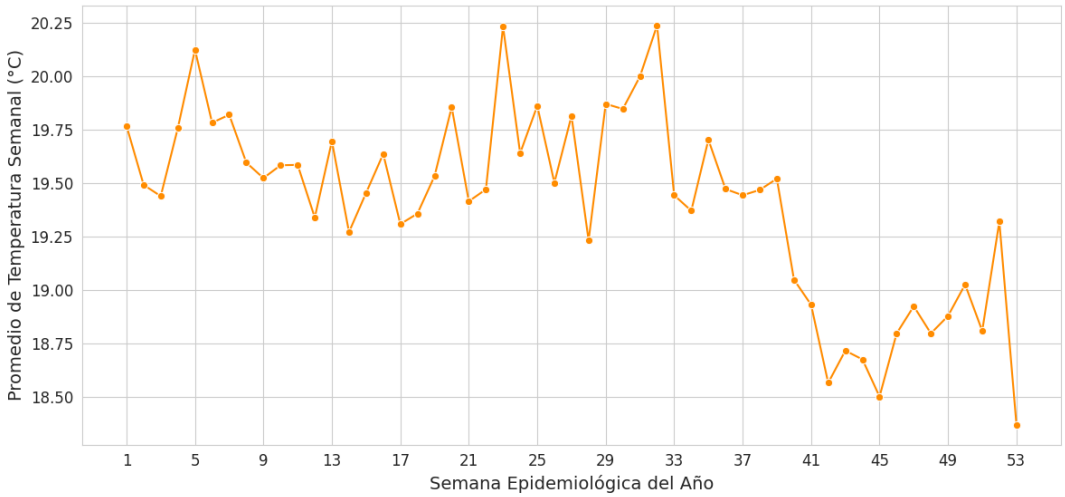
Elaboración propia.

Figura 4. Patrón estacional promedio de temperatura mínima en Medellín.



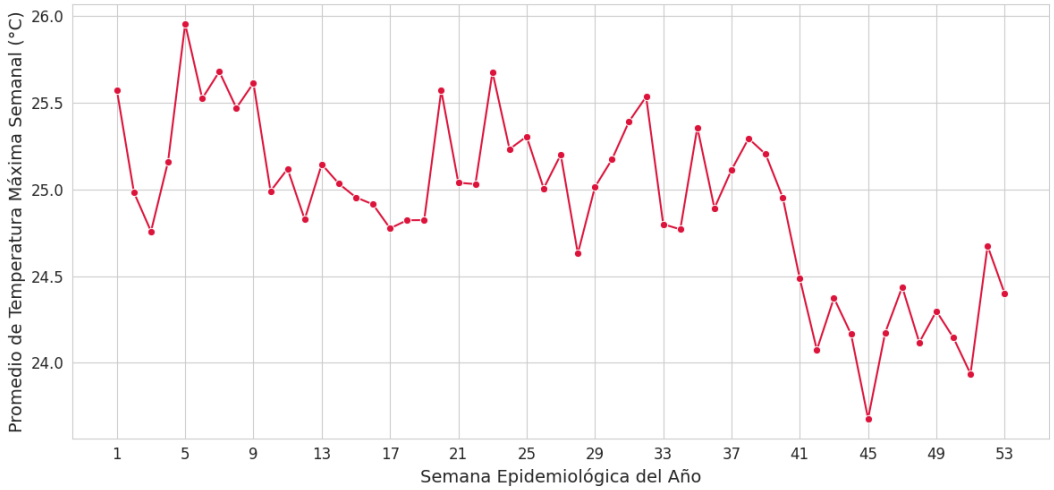
Elaboración propia.

Figura 5. Patrón estacional promedio de temperatura promedio en Medellín.



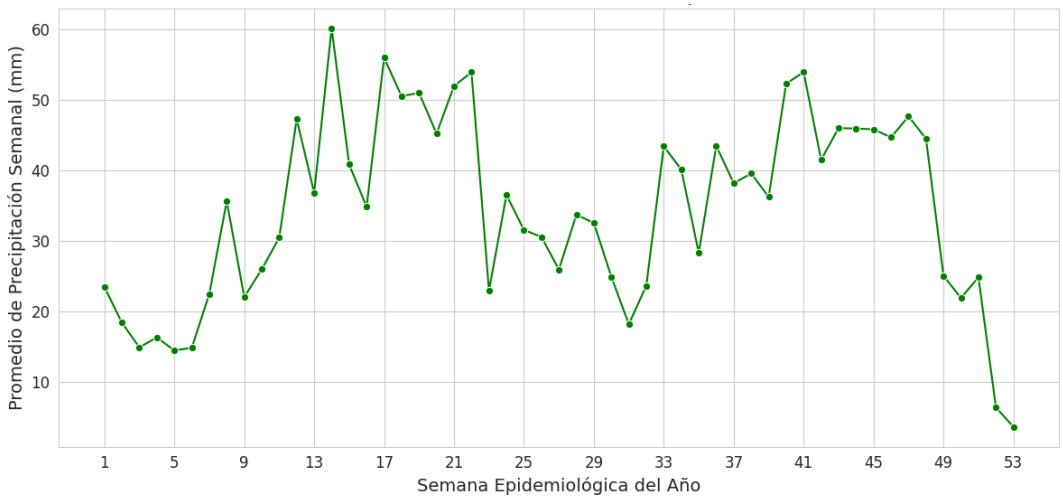
Elaboración propia.

Figura 6. Patrón estacional promedio de temperatura máxima en Medellín.



Elaboración propia.

Figura 7. Patrón estacional promedio de precipitación en Medellín.

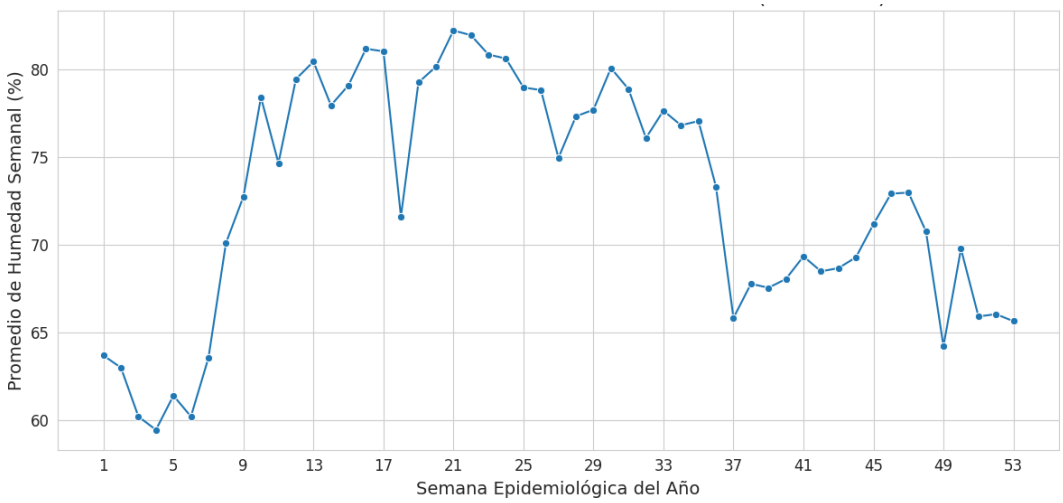


Elaboración propia.

En Bucaramanga, las Figuras 18 a 22 muestran un comportamiento similar en términos de doble estación lluviosa. Sin embargo, las temperaturas mínima, promedio y máxima fueron sistemáticamente superiores a las de Medellín. Además, se aproximaron a los valores observados en Cali.

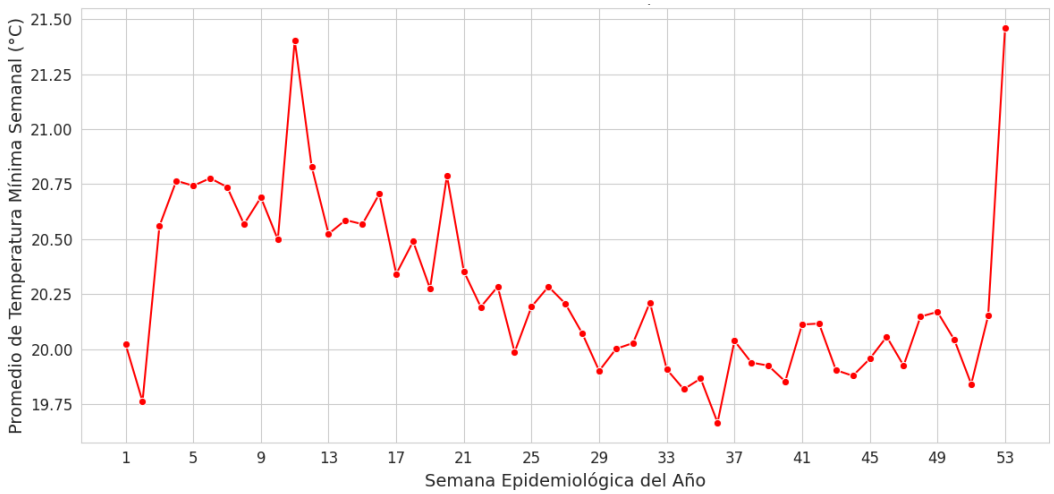
Las precipitaciones también se concentraron en dos bloques anuales. El primero se ubicó hacia el segundo trimestre del año. El segundo se observó aproximadamente entre las semanas 40 y 45.

Figura 88. Patrón estacional promedio de humedad relativa en Bucaramanga.



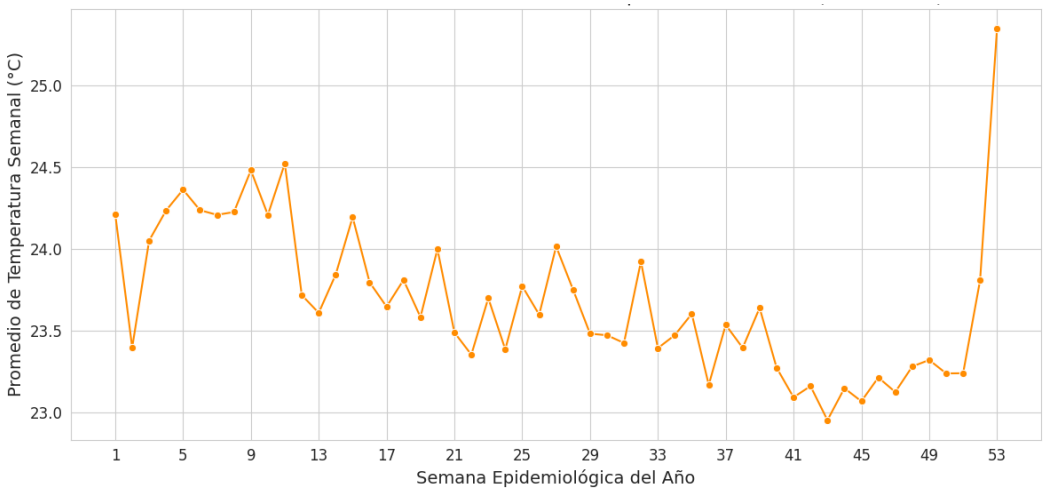
Elaboración propia.

Figura 9. Patrón estacional promedio de temperatura mínima en Bucaramanga.



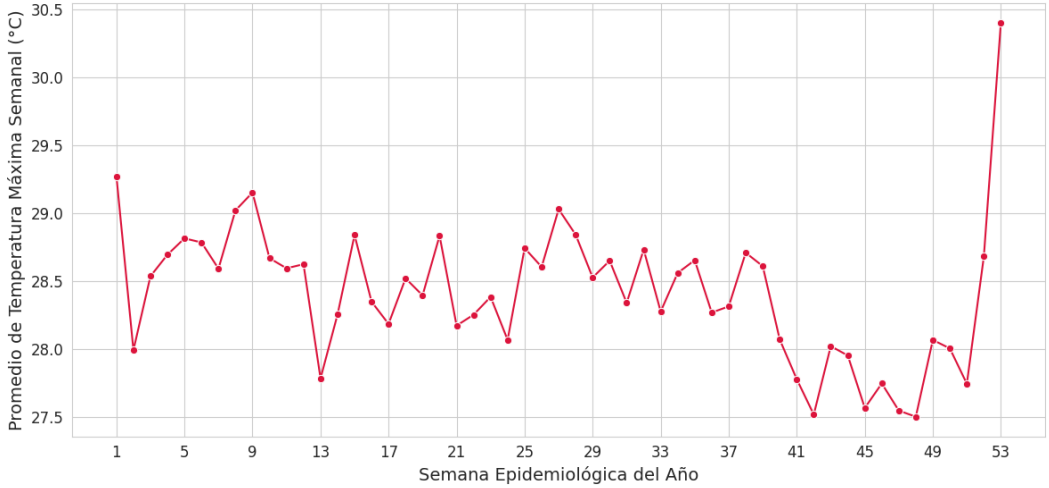
Elaboración propia.

Figura 2010. Patrón estacional promedio de temperatura promedio en Bucaramanga.



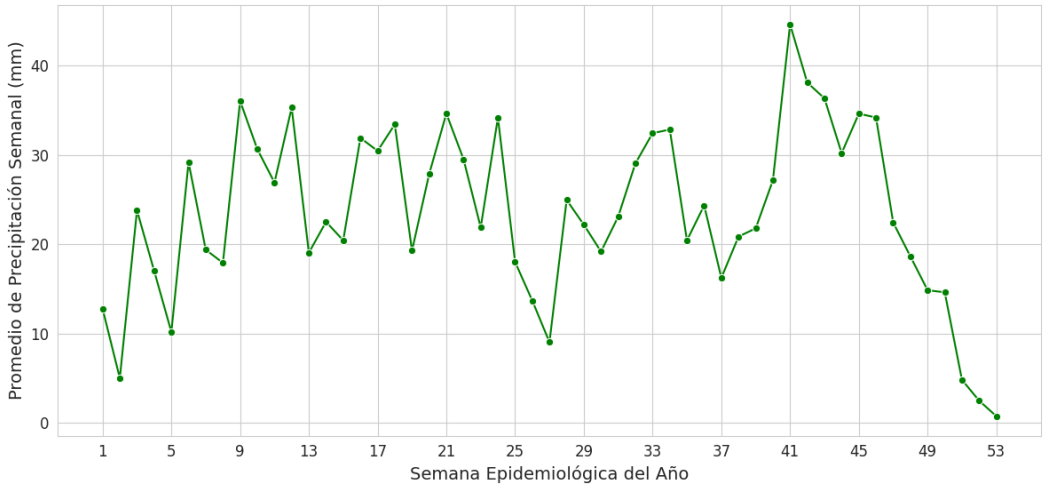
Elaboración propia.

Figura 21. Patrón estacional promedio de temperatura máxima en Bucaramanga.



Elaboración propia.

Figura 22. Patrón estacional promedio de precipitación en Bucaramanga.



Elaboración propia.

La **Tabla 3** sintetiza estos patrones climáticos mediante estadísticos descriptivos semanales por ciudad. Los resultados muestran que Medellín presentó, en promedio, temperaturas mínimas más bajas. En contraste, Cali y Bucaramanga registraron temperaturas más elevadas y mayores niveles de humedad.

Bucaramanga combinó, además, temperaturas y humedades altas durante gran parte del año. Este patrón configura condiciones potencialmente favorables para la proliferación del vector.

Estas diferencias en los rangos y en la variabilidad climática entre ciudades justifican el ajuste de modelos específicos para cada contexto urbano.

Tabla 3. Resumen de variables climáticas promedio semanales por ciudad.

Ciudad	Variable	Media	DE	Mínimo	Máximo
Cali	Humedad relativa promedio semanal (%)	73.00	5.12	58.83	93.62
Cali	Temperatura mínima promedio semanal (°C)	20.16	3.28	0.00	23.27
Cali	Temperatura promedio semanal (°C)	24.35	3.30	0.46	28.36
Cali	Temperatura máxima promedio semanal (°C)	29.95	3.20	1.90	39.77
Cali	Precipitación promedio semanal (mm)	23.38	28.33	0.00	272.80
Medellín	Humedad relativa promedio semanal (%)	73.59	13.53	0.16	90.42
Medellín	Temperatura mínima promedio semanal (°C)	15.54	1.30	10.43	19.03
Medellín	Temperatura promedio semanal (°C)	19.43	1.36	15.36	24.06
Medellín	Temperatura máxima promedio semanal (°C)	24.93	1.57	20.90	30.79
Medellín	Precipitación promedio semanal (mm)	34.77	30.15	0.00	148.80

Bucaramanga	Humedad relativa promedio semanal (%)	73.08	19.69	0.05	92.85
Bucaramanga	Temperatura mínima promedio semanal (°C)	20.16	0.99	11.20	22.67
Bucaramanga	Temperatura promedio semanal (°C)	23.55	1.07	15.20	26.18
Bucaramanga	Temperatura máxima promedio semanal (°C)	28.24	1.37	17.50	31.84
Bucaramanga	Precipitación promedio semanal (mm)	23.69	25.39	0.00	194.90

Elaboración propia.

4.2.3. Rezagos climáticos, inercia epidemiológica y evidencia empírica

Para explorar la relación entre los casos de dengue y las variables climáticas, se calcularon rezagos temporales de cuatro semanas para cada predictor climático. Este rezago busca aproximar el tiempo necesario para que cambios en temperatura, humedad o precipitación se traduzcan en variaciones en la incidencia de dengue.

La elección de este desfase se fundamenta en criterios biológicos. En particular, es coherente con:

- el desarrollo del vector (*huevo–larva–adulto*),
- el periodo de incubación extrínseca del virus en el mosquito y
- el periodo de incubación en humanos.

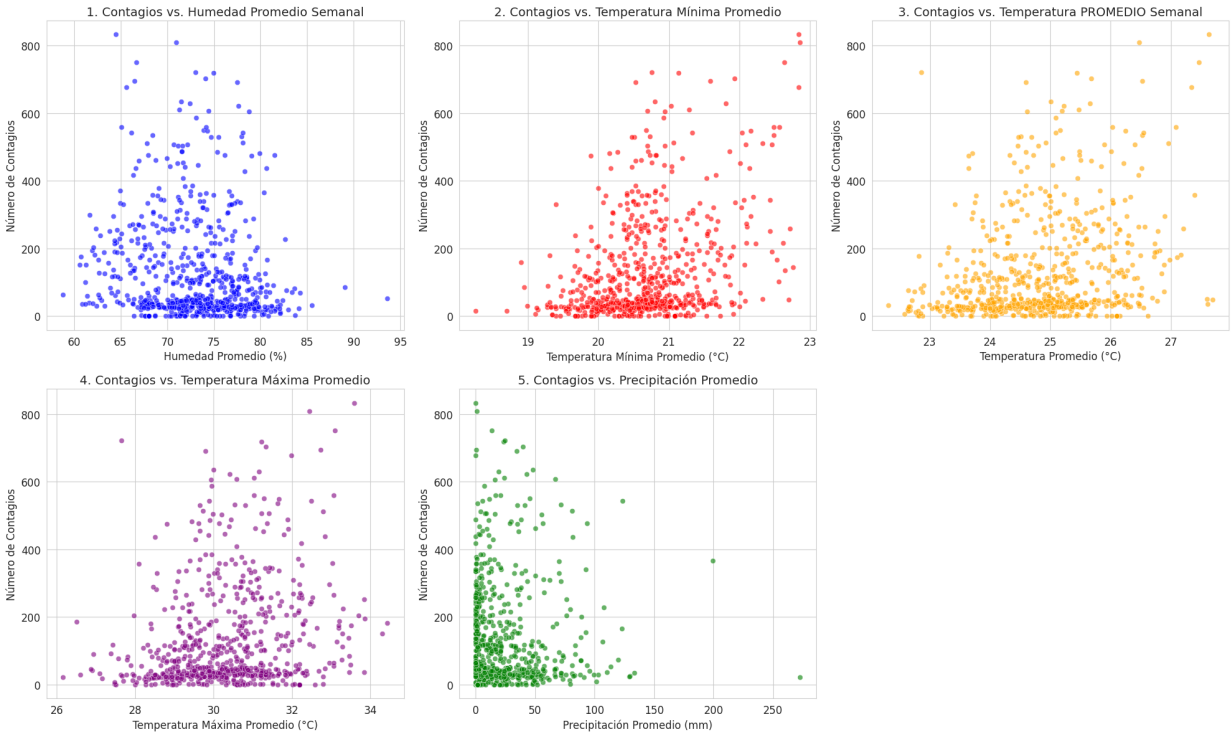
En conjunto, estos procesos suelen implicar varias semanas entre una condición climática favorable y el aumento observado de casos.

Adicionalmente, se incorporó un término de inercia epidemiológica. Este término se construyó mediante el cálculo de un promedio móvil de cuatro semanas del número de casos (*Contagios_promedio_movil_4*). Su objetivo es capturar la memoria reciente de la transmisión.

Las Figuras 23, 24 y 25 presentan diagramas de dispersión entre el número de casos semanales y las variables climáticas seleccionadas. Estas variables incluyen humedad, temperaturas mínima, promedio y máxima, así como precipitación, para cada ciudad.

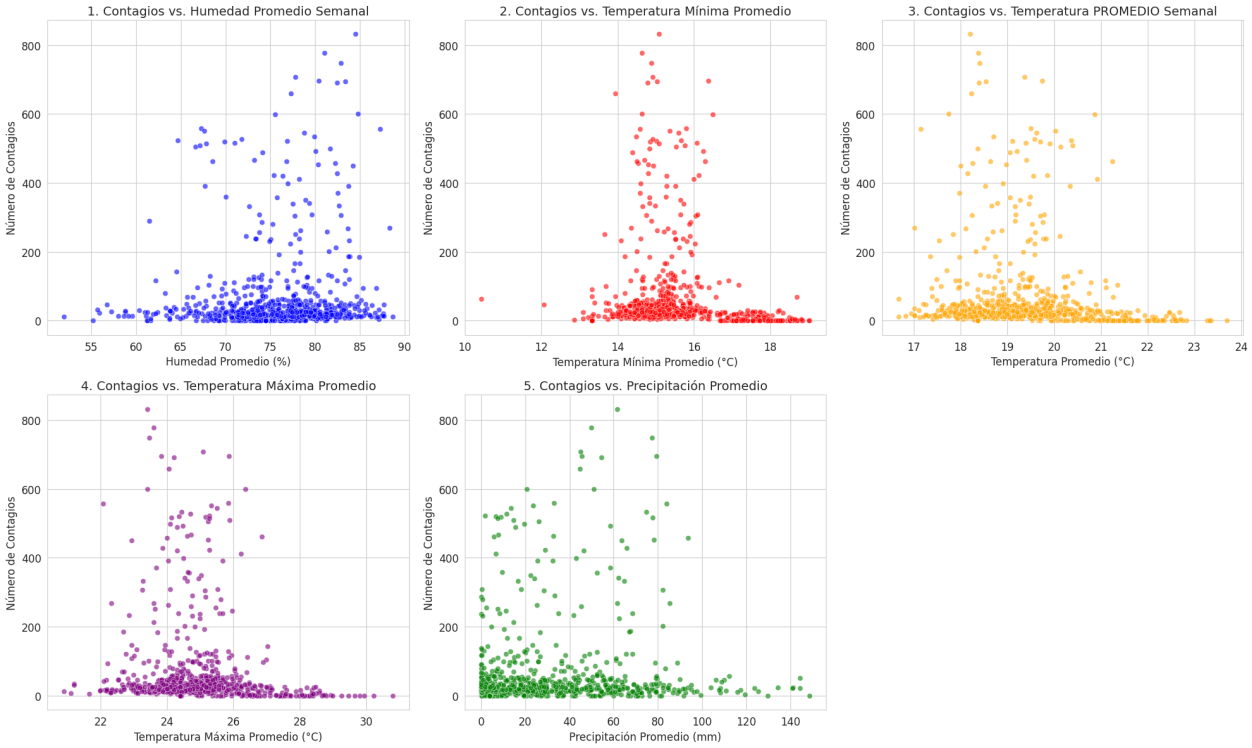
Aunque las nubes de puntos muestran una alta dispersión, se identificaron tendencias generales. En particular, se observó una asociación positiva entre ciertos rangos de humedad y temperatura y el incremento de casos. Asimismo, se evidenció un posible efecto de saturación a niveles elevados de precipitación.

Figura 11. Diagramas de dispersión entre casos y variables climáticas en Cali.



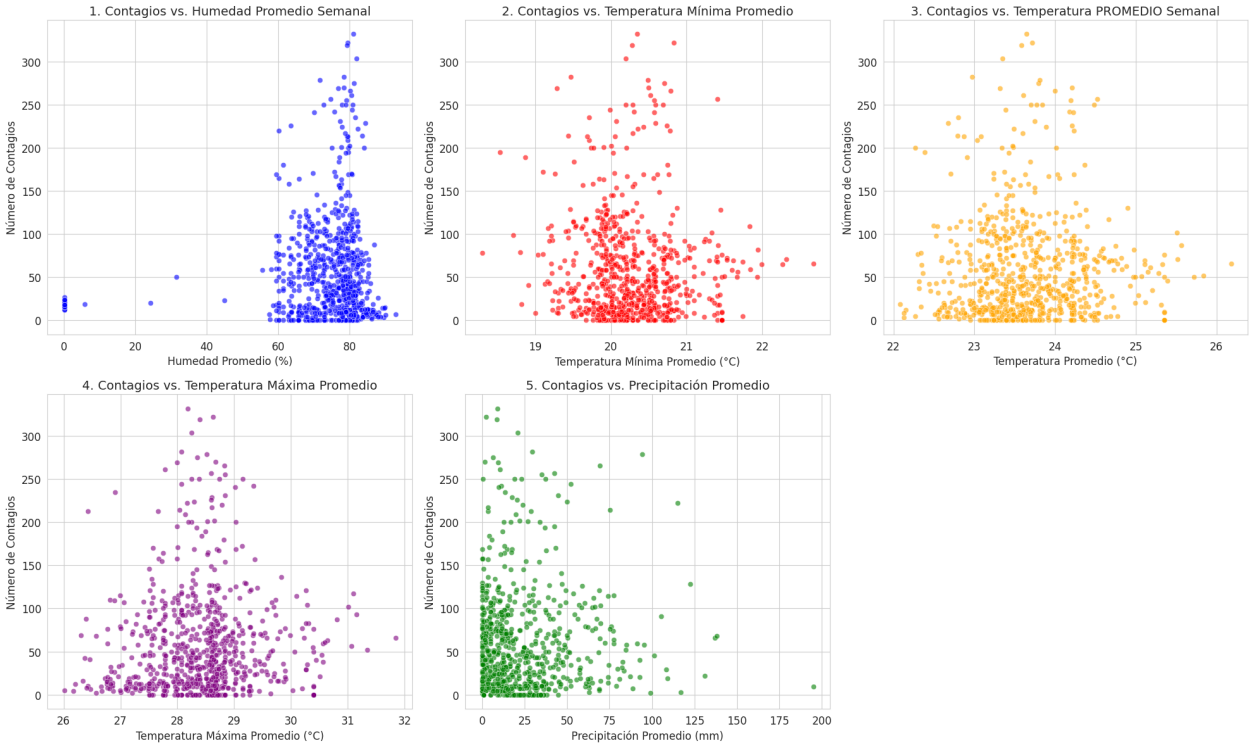
Elaboración propia.

Figura 124. Diagramas de dispersión entre casos y variables climáticas en Medellín.



Elaboración propia.

Figura 25. Diagramas de dispersión entre casos y variables climáticas en Bucaramanga.



Elaboración propia.

Esta evidencia exploratoria apoyó la definición de los predictores que se describen en la siguiente sección.

Con el fin de sintetizar estas relaciones, se construyeron matrices de correlación de Pearson entre los contagios semanales de dengue y las variables climáticas promedio con rezago de cuatro semanas. La Tabla 4 presenta los coeficientes obtenidos para cada ciudad.

En Cali se observa una correlación positiva moderada entre los contagios y la temperatura mínima (**r ≈ 0,39**). También se identifica una asociación positiva con la temperatura promedio (**r ≈ 0,31**). En contraste, la humedad relativa y la precipitación muestran asociaciones muy débiles con los casos.

En Medellín y Bucaramanga, las correlaciones entre contagios y todas las variables climáticas resultaron bajas (**|r| < 0,11**). Este patrón sugiere que la dinámica de los brotes en estas ciudades podría depender en mayor medida de otros factores. Alternativamente, podría responder a relaciones no lineales que serán capturadas por los modelos de aprendizaje automático.

Estos resultados son coherentes con los diagramas de dispersión. En particular, solo en Cali se aprecia una tendencia creciente más definida a medida que aumenta la temperatura.

Tabla 4. Coeficientes de correlación de Pearson entre los contagios de dengue y las variables climáticas rezagadas 4 semanas, por ciudad.

Ciudad	Humedad promedio (lag 4)	Temp. mínima promedio (lag 4)	Temp. promedio (lag 4)	Temp. máxima promedio (lag 4)	Precipitación promedio (lag 4)
Cali	-0,15	0,39	0,31	0,25	-0,04
Medellín	0,04	-0,1	-0,07	-0,07	-0,02
Bucaramanga	0,11	0	0,06	0,1	0

Elaboración propia.

4.3. Escenarios de modelado y variables predictoras

Con base en los patrones de estacionalidad, la estructura de los rezagos climáticos y los diagramas de dispersión, se definieron dos escenarios de modelado con complejidad creciente para cada ciudad.

El análisis se realizó a nivel semanal, ya que coincide con el esquema de notificación

epidemiológica y con la toma de decisiones operativas. Este nivel temporal permite activar alertas tempranas, intensificar el control vectorial y priorizar zonas de intervención.

El horizonte temporal seleccionado busca un equilibrio entre la utilidad práctica y la capacidad predictiva, considerando la disponibilidad de los datos.

4.3.1. Escenario 1: modelos basados en variables climáticas

En el **Escenario 1** se evaluó la capacidad de las variables climáticas rezagadas y de la estacionalidad anual para explicar la variación del número de contagios semanales. Para ello, se definió un conjunto de predictores compuesto por los siguientes elementos:

- **Semana epidemiológica (*semana_epi*):** codificada como variable numérica, con el fin de capturar el ciclo anual del dengue.

- **Variable climática principal con rezago de cuatro semanas (lag = 4):** esta variable se seleccionó de manera específica para cada ciudad, a partir del análisis de patrones estacionales y diagramas de dispersión.

La elección de un rezago de cuatro semanas se justifica por el ciclo epidemiológico del dengue. Este ciclo incluye el desarrollo del mosquito (≈10–14 días), la incubación extrínseca del virus en el vector (≈7–14 días) y la incubación en humanos (≈4–10 días). En conjunto, estos procesos suelen implicar un intervalo de entre cuatro y seis semanas desde un evento climático hasta la notificación de un caso.

No obstante, se reconoce que el rezago “típico” puede variar entre dos y seis semanas según el contexto local. Por esta razón, en el **Escenario 2** se ampliará el conjunto de predictores para incluir múltiples rezagos y capturar mejor la dinámica temporal.

En Cali y Medellín, la variable climática seleccionada fue la temperatura mínima promedio semanal con rezago de cuatro semanas. Esta variable mostró variaciones estacionales claras y una asociación plausible con la supervivencia del vector.

En Bucaramanga, en cambio, se seleccionó la humedad relativa promedio semanal con el mismo rezago. Esta decisión se basó en su mayor variabilidad y en su aparente relación con los incrementos de casos observados.

4.3.2. Escenario 2: modelos basados en clima e inercia epidemiológica

El **Escenario 2** amplió el conjunto de predictores mediante la incorporación de un término de **inercia epidemiológica**. Además de la semana epidemiológica y de la variable climática rezagada, se incluyó el **promedio móvil de cuatro semanas del número de contagios** (*Contagios_promedio_movil_4*).

Esta variable permitió capturar la dependencia temporal de corto plazo de la serie. En particular, refleja la tendencia de la incidencia de dengue a mantenerse elevada durante varias semanas consecutivas una vez iniciado un brote.

La **Tabla 5** detalla la configuración de ambos escenarios para cada ciudad. En ella se especifican la variable objetivo (contagios semanales) y el conjunto de predictores utilizados.

Los resultados muestran que, aunque el marco metodológico fue común, la variable climática principal difirió entre ciudades. En Cali y Medellín se priorizó la temperatura mínima. En Bucaramanga se utilizó la humedad relativa.

Tabla 5. Escenarios de modelado y conjunto de predictores por ciudad.

Ciudad	Escenario	Variable objetivo	Predictores incluidos
Cali	1	Contagios	semana_epi; Temperatura_PromedioMinimaSemanal_lag_4
Cali	2	Contagios	semana_epi; Temperatura_PromedioMinimaSemanal_lag_4; Contagios_promedio_movil_4
Medellín	1	Contagios	semana_epi; Temperatura_PromedioMinimaSemanal_lag_4
Medellín	2	Contagios	semana_epi; Temperatura_PromedioMinimaSemanal_lag_4; Contagios_promedio_movil_4
Bucaramanga	1	Contagios	semana_epi; Humedad_PromedioSemanal_lag_4
Bucaramanga	2	Contagios	semana_epi; Humedad_PromedioSemanal_lag_4; Contagios_promedio_movil_4

Elaboración propia.

Cabe resaltar que el término de inercia epidemiológica se definió de manera homogénea para las tres ciudades. En cada municipio se calculó un promedio móvil de cuatro semanas del número de contagios semanales (*Contagios_promedio_movil_4*).

Este indicador se incorporó como predictor adicional, junto con la semana epidemiológica

y la variable climática rezagada correspondiente.

4.4. Algoritmos de aprendizaje automático utilizados

Como se describió en la formulación de los escenarios de modelado, los predictores empleados se centraron en variables climáticas, la semana epidemiológica y un término de inercia epidemiológica. En particular, se utilizaron temperatura y humedad relativa con rezagos de cuatro semanas, junto con el promedio móvil de contagios.

Durante la revisión bibliográfica y la construcción de la base de datos se consideró inicialmente la inclusión de variables sociodemográficas y socioeconómicas a nivel municipal. Sin embargo, estas variables no se incorporaron en los modelos finales por varias razones metodológicas.

En primer lugar, la información disponible para estos indicadores presenta una resolución temporal anual o plurianual. Esta característica es poco compatible con la naturaleza semanal de las series de casos y de las variables climáticas. Su inclusión habría requerido supuestos fuertes de interpolación o de constancia temporal.

En segundo lugar, la variabilidad temporal de los indicadores sociodemográficos dentro de cada ciudad es limitada en el horizonte analizado. Por ello, su aporte marginal a la predicción semanal de brotes resulta reducido frente al efecto del clima y de la dinámica reciente de transmisión.

Finalmente, la incorporación de predictores adicionales con baja variabilidad, dado el tamaño muestral disponible, habría incrementado el riesgo de sobreajuste. Además, habría dificultado la interpretación de los modelos.

Por estas razones, se optó por una especificación parsimoniosa centrada en el componente climático y en la inercia epidemiológica. El análisis de los factores sociodemográficos se mantuvo como un elemento contextual y como una línea explícita de trabajo futuro.

Para cada combinación de ciudad y escenario se implementó un conjunto de algoritmos de aprendizaje automático. Estos algoritmos permiten modelar relaciones no lineales y posibles interacciones entre variables climáticas, estacionalidad e inercia epidemiológica.

4.4.1. Bosque aleatorio

El primer algoritmo empleado fue el **bosque aleatorio (*Random Forest*)**. Este método de *ensamble* se basa en la agregación de múltiples árboles de decisión entrenados sobre subconjuntos aleatorios de datos y predictores. Este enfoque ha demostrado ser robusto frente a relaciones no lineales y a la presencia de ruido en los datos.

En los experimentos se ajustaron bosques aleatorios con diferentes configuraciones. Se variaron el número de árboles, la profundidad máxima y el tamaño mínimo de muestra por nodo. El objetivo fue encontrar un equilibrio entre capacidad de ajuste y riesgo de sobreajuste. Los hiperparámetros seleccionados se resumen en la Tabla 6.

4.4.2. Modelos de gradiente reforzado (XGBoost)

El segundo algoritmo considerado fue **XGBoost**. Este método de gradiente reforzado construye secuencias de árboles de decisión. Cada nuevo árbol se ajusta para corregir los errores residuales de los árboles anteriores. Esta familia de modelos suele ofrecer un alto desempeño predictivo en problemas tabulares. En particular, resulta eficaz cuando existen interacciones complejas entre las variables.

Se exploraron distintas combinaciones de hiperparámetros. Entre ellas se incluyeron el número de árboles, la tasa de aprendizaje, la profundidad máxima y los parámetros de regularización. El objetivo fue obtener modelos parsimoniosos, pero con buen desempeño predictivo.

Al igual que en el caso del bosque aleatorio, la selección final de hiperparámetros se presenta en la Tabla 6.

4.4.3. Redes neuronales densas y recurrentes

Finalmente, se implementaron redes neuronales artificiales de dos tipos: redes densas (feedforward) y redes recurrentes basadas en unidades GRU o LSTM. Las redes densas permitieron modelar relaciones altamente no lineales entre los predictores en un espacio de características transformado.

Las redes recurrentes, por su parte, ofrecieron la posibilidad de capturar dependencias temporales explícitas a lo largo de la secuencia de semanas. Este enfoque resulta especialmente relevante en series epidemiológicas, donde la evolución reciente condiciona el comportamiento futuro.

En ambos casos se definieron arquitecturas con una o varias capas ocultas. Se emplearon funciones de activación no lineales y técnicas de regularización, como dropout, para mitigar el riesgo de sobreajuste. Los principales hiperparámetros —número de capas, neuronas por capa, funciones de activación y valores de dropout— se sintetizan en la Tabla 6.

La **Tabla 5** presenta un resumen comparativo de los algoritmos utilizados y de sus configuraciones base. Este resumen proporciona un marco de referencia para la replicación del estudio y para posibles extensiones futuras.

Los modelos basados en árboles, como Random Forest y XGBoost, se caracterizan por su robustez. Funcionan adecuadamente con variables heterogéneas y ofrecen un mayor grado de interpretabilidad mediante medidas de importancia de predictores. No obstante, estos modelos no incorporan explícitamente la dimensión temporal y requieren variables rezagadas para codificar la secuencia.

En contraste, las redes recurrentes tipo GRU permiten aprender dependencias temporales de forma más directa. Sin embargo, suelen requerir un mayor cuidado en el preprocesamiento, particularmente en el escalamiento y la regularización. Además, presentan una menor transparencia interpretativa y un mayor costo computacional

durante el entrenamiento.

Tabla 6. Algoritmos implementados y principales hiperparámetros utilizados.

Algoritmo	Tipo de modelo	Principales hiperparámetros usados	Comentario
Random Forest	Ensamble de árboles	n_estimators = 500; max_depth = 10	Captura no linealidades y permite importancia de variables.
XGBoost	Ensamble por boosting	n_estimators = 100; learning_rate = 0.1	Maneja bien interacciones complejas y datos ruidosos.

GRU (Cali)	Red neuronal recurrente	Architecture = GRU(128), Dense(64);
		learning_rate = 0.001;
		dropout = 0.2;
		batch_size = 25;
		epochs=100
			Captura dependencia temporal de la serie.

GRU (Medellín)		Architecture = GRU(128), Dense(64), Dense(32);
		learning_rate = 0.0005;
		dropout = 0.4;
		batch_size = 32;
		epochs = 100	

GRU (Bucaramanga)		Architecture = GRU(128), Dense(64), Dense(32);
		learning_rate=0.001;
		dropout=0.3;
		batch_size=16;
		epochs=100	

Elaboración propia.

4.5. Configuración de los modelos por ciudad

4.5.1. Especificación de los modelos para Cali

Para Cali se construyeron modelos bajo los dos escenarios definidos.

En el **Escenario 1**, el vector de características estuvo compuesto por la semana epidemiológica y la temperatura mínima promedio semanal con rezago de cuatro

semanas.

En el **Escenario 2**, se incorporó adicionalmente el promedio móvil de cuatro semanas del número de contagios.

Sobre estos conjuntos de predictores se ajustaron los cuatro tipos de modelos descritos en la Sección 4.4. Estos incluyeron bosque aleatorio, XGBoost, redes neuronales densas y recurrentes. El entrenamiento se realizó utilizando el conjunto definido en la Tabla 1.

El ajuste de los modelos se llevó a cabo mediante validación cruzada o particiones internas de entrenamiento y validación, según el algoritmo. Las configuraciones con mejor desempeño en el conjunto de validación se seleccionaron para su evaluación final, presentada en el capítulo siguiente.

4.5.2. Especificación de los modelos para Medellín

En Medellín se replicó la misma estructura de escenarios utilizada en Cali. En el **Escenario 1**, el conjunto de predictores incluyó la semana epidemiológica y la temperatura mínima promedio semanal con rezago de cuatro semanas.

En el **Escenario 2**, se incorporó adicionalmente el promedio móvil de cuatro semanas del número de contagios.

Los algoritmos de bosque aleatorio, XGBoost y redes neuronales se entrenaron utilizando configuraciones análogas a las empleadas en Cali. Los hiperparámetros se ajustaron a las características específicas de la serie temporal de Medellín.

La selección de los modelos finales se basó, de igual manera, en su desempeño sobre el conjunto de validación interno.

4.5.3. Especificación de los modelos para Bucaramanga

Para Bucaramanga se siguió la misma lógica metodológica, sustituyendo únicamente la variable climática principal. En el **Escenario 1**, el vector de características incluyó la semana epidemiológica y la humedad relativa promedio semanal con rezago de cuatro semanas.

En el **Escenario 2**, se incorporó adicionalmente el promedio móvil de cuatro semanas del número de contagios.

Los algoritmos de aprendizaje automático se entrenaron utilizando el conjunto de entrenamiento correspondiente a Bucaramanga. La comparación entre escenarios y entre algoritmos, así como la evaluación de su capacidad para reproducir los patrones observados en las Figuras 3, 6 y 17 a 21, se presentan en detalle en el **Capítulo 5**.

5. EVALUACIÓN DEL DESEMPEÑO Y LA ROBUSTEZ DE LOS MODELOS PREDICTIVOS DE DENGUE

5.1. Diseño de la evaluación

5.1.1. Esquema de partición temporal

Tal como se describió en el capítulo anterior, para cada ciudad se dispuso de una serie semanal de casos de dengue. En Cali y Medellín, el horizonte temporal abarcó el periodo 2007–2019. En Bucaramanga, se contó con información desde 2006 hasta 2019.

Se adoptó una partición temporal diseñada para preservar la estructura de serie de tiempo:

- **Conjunto de entrenamiento:** semanas comprendidas entre 2007 y 2017 (o 2006–2017 en el caso de Bucaramanga).
- **Conjunto de prueba:** semanas correspondientes a los años 2018–2019.

Este esquema garantizó que la evaluación se realizara sobre un bloque temporal reciente, no utilizado durante el ajuste de los modelos. De este modo, fue posible valorar la capacidad de generalización de los modelos frente a condiciones futuras.

5.1.2. Métricas de desempeño

La evaluación cuantitativa de los modelos se basó principalmente en dos métricas:

- **Error Cuadrático Medio (MSE)**

$$\text{MSE} = \frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2$$

- **Raíz del Error Cuadrático Medio (RMSE)**

$$\text{RMSE} = \sqrt{\text{MSE}}$$

donde $y_t$representó el número observado de contagios en la semana ty $\hat{y}_t$la predicción correspondiente del modelo sobre el conjunto de prueba. El **Error Cuadrático Medio (MSE)** penaliza con mayor peso los errores grandes, mientras que la **Raíz del Error Cuadrático Medio (RMSE)** permite interpretar el desempeño en las mismas unidades que la variable objetivo, es decir, casos de dengue por semana.

En los tres *notebooks* de modelado, el **RMSE** se utilizó como métrica central para comparar algoritmos y escenarios. Esta elección se debe a que el RMSE proporciona una medida directa del error promedio de predicción semanal y penaliza con mayor severidad los errores grandes. Este aspecto resulta especialmente relevante en el contexto de la salud pública, ya que fallos de predicción durante semanas de pico epidémico pueden afectar la activación de alertas tempranas y la asignación eficiente de recursos sanitarios.

El **Error Absoluto Medio (MAE)** se incluyó como métrica complementaria, dado que ofrece una interpretación más directa del error promedio en términos de número de casos por semana. El **MSE**, por su parte, se reportó principalmente con fines de optimización y comparación técnica, aunque su interpretación en términos de magnitud es menos intuitiva.

En el informe se presenta el RMSE como métrica principal de desempeño, mientras que el MAE y el MSE se utilizan como métricas de apoyo para evaluar la estabilidad de los modelos y posibles sesgos en la predicción.

5.1.3. Procedimiento de comparación y criterios de robustez

Para cada ciudad se siguió el siguiente procedimiento:

1. **Entrenamiento de los modelos** en el conjunto de entrenamiento, bajo dos escenarios de entrada:

- Escenario 1: **variables climáticas + semana epidemiológica**.
- Escenario 2: **variables climáticas + semana epidemiológica + inercia epidemiológica** (promedio móvil de contagios).

2. **Predicción sobre el conjunto de prueba** (2018–2019), generando series semanales de predicciones.
3. **Cálculo de MSE y RMSE** para cada modelo, algoritmo y escenario.
4. **Comparación de desempeño** entre:
   - algoritmos dentro de cada escenario, y
   - escenarios (1 vs 2) dentro de cada ciudad.
5. **Análisis gráfico de robustez temporal**, contrastando las series observadas y predichas, con énfasis en:
   - semanas con picos epidémicos, y
   - semanas con niveles bajos o moderados de transmisión.

5.2. Resultados de desempeño por escenario

5.2.1. Escenario 1: modelos basados en variables climáticas

En el **Escenario 1** se utilizaron como predictores la **semana epidemiológica** y una **variable climática rezagada cuatro semanas** (temperatura mínima promedio en Cali y Medellín, humedad relativa en Bucaramanga). Sobre este conjunto reducido de características se entrenaron modelos de **Random Forest**, **XGBoost** y **red neuronal densa (DNN)**.

En una primera etapa se ajustó un modelo base de Random Forest con hiperparámetros por defecto. Los resultados iniciales sobre el conjunto de prueba arrojaron valores de RMSE de:

- **Cali**: RMSE ≈ **154,77**.
- **Medellín**: RMSE ≈ **113,78**.
- **Bucaramanga**: RMSE ≈ **38,66**.

Estos valores representaron el punto de partida para el resto de los modelos. Posteriormente, la optimización de hiperparámetros del Random Forest (mediante ajuste de número de árboles, profundidad y tamaño mínimo de muestra) permitió reducir el RMSE a:

- **132,19** en Cali,
- **101,16** en Medellín,
- **36,69** en Bucaramanga.

La red neuronal densa (DNN), entrenada con las mismas variables de entrada, alcanzó un desempeño comparable o mejor que el Random Forest optimizado, especialmente en Cali y Medellín:

- **Cali**: RMSE de la DNN ≈ **126,11**, ligeramente inferior al del Random Forest optimizado.
- **Medellín**: RMSE de la DNN ≈ **71,17**, con una mejora sustantiva frente al Random Forest (101,16).
- **Bucaramanga**: RMSE de la DNN ≈ **37,01**, muy similar al Random Forest optimizado (36,69).

En cambio, en este escenario el modelo XGBoost no superó de manera consistente a los anteriores, particularmente en Cali, donde obtuvo un RMSE alrededor de 151,89, por encima de la DNN y del Random Forest optimizado.

Tabla 7. Desempeño de los modelos en el Escenario 1 (RMSE en el conjunto de prueba por ciudad y algoritmo).

Ciudad	Algoritmo	RMSE
Cali	Random Forest	132.19
Cali	XGBoost	151.89
Cali	DNN	126.11
Medellín	Random Forest	101.16
Medellín	XGBoost	116.19
Medellín	DNN	71.17
Bucaramanga	Random Forest	36.69
Bucaramanga	XGBoost	49.38
Bucaramanga	DNN	37.01

Elaboración propia.

La Tabla 7 sintetizó el RMSE de los principales algoritmos en el Escenario 1. En términos generales, los resultados indicaron que:

- La **optimización de hiperparámetros** del Random Forest mejoró de forma sistemática el desempeño respecto al modelo base.

- La **DNN** obtuvo el mejor RMSE en Cali y Medellín, mientras que en Bucaramanga el Random Forest y la DNN mostraron desempeños muy similares.
- El uso exclusivo de variables climáticas y de la estacionalidad anual resultó suficiente para capturar parte de la variabilidad semanal de los contagios, pero dejó espacio para mejoras importantes.

5.2.2. Escenario 2: modelos con clima e inercia epidemiológica

En el **Escenario 2** se incorporó la variable de **inercia epidemiológica** Contagios_promedio_movil_4, junto con la semana epidemiológica y la variable climática rezagada. Este enriquecimiento del vector de características se tradujo en una mejora drástica del desempeño.

En Cali, la inclusión de la inercia permitió que:

- El **Random Forest (Inercia y Clima)** alcanzara un RMSE ≈ **16,22**.
- El modelo **XGBoost (Inercia y Clima)** mejorara aún más, con RMSE ≈ **14,94**.
- La red recurrente GRU, aunque competitiva, se situara por encima con un RMSE ≈ **25,35**.

En Medellín se observó un patrón similar de mejora, aunque con un liderazgo diferente entre algoritmos:

- El **Random Forest (Inercia y Clima)** obtuvo un RMSE ≈ **5,20**.
- El **XGBoost (Inercia y Clima)** tuvo un RMSE muy próximo, ≈ **5,27**.
- La GRU alcanzó un RMSE algo mayor, ≈ **8,13**.

En Bucaramanga, la inercia epidemiológica también desempeñó un papel clave, y el modelo ganador fue la red recurrente:

- El **Random Forest (Inercia y Clima)** presentó un RMSE ≈ **8,79**.
- El **XGBoost (Inercia y Clima)** obtuvo un RMSE ≈ **9,29**.
- La **GRU (Inercia y Clima)** logró el mejor desempeño, con RMSE ≈ **8,35**.

Tabla 8. Desempeño de los modelos en el Escenario 2 (RMSE en el conjunto de prueba por ciudad y algoritmo).

Ciudad	Algoritmo	RMSE
Cali	RandomForest	16.22
Cali	XGBoost	14.94
Cali	GRU	25.35
Medellín	RandomForest	5.20
Medellín	XGBoost	5.27
Medellín	GRU	8.13
Bucaramanga	RandomForest	8.79
Bucaramanga	XGBoost	9.29
Bucaramanga	GRU	8.35

Elaboración propia.

La Tabla 8 resumió estos resultados y permitió identificar, para cada ciudad, el modelo con menor RMSE en el escenario con clima e inercia:

- **Cali**: modelo ganador **XGBoost**.
- **Medellín**: modelo ganador **Random Forest**.
- **Bucaramanga**: modelo ganador **GRU**.

5.3. Comparación entre escenarios y selección de modelos

5.3.1. Mejora global al incorporar la inercia epidemiológica

La comparación entre el Escenario 1 (solo clima) y el Escenario 2 (clima + inercia) puso de manifiesto la importancia de incorporar la dinámica interna de la serie de contagios. Si se toma como referencia el modelo base de Random Forest del Escenario 1, la reducción relativa de RMSE al pasar al mejor modelo del Escenario 2 fue aproximadamente de:

- **Cali**: reducción cercana al **90 %** (de 154,77 a 14,94).
- **Medellín**: reducción cercana al **95 %** (de 113,78 a 5,20).
- **Bucaramanga**: reducción cercana al **78 %** (de 38,66 a 8,35).

El panorama general de los resultados entre escenarios puede apreciarse en la Tabla 9.

Tabla 9. Comparación de RMSE entre escenarios (por ciudad, modelo base del Escenario 1 y modelo ganador del Escenario 2, con mejora porcentual).

Ciudad	Modelo base (Esc.1)	RMSE Esc.1	Modelo ganador (Esc.2)	RMSE Esc.2	Mejora (%)
Cali	RandomForestbase	154.77	XGBoost(InerciayClima)	14.94	90.35
Medellín	RandomForestbase	113.78	RandomForest(InerciayClima)	5.20	95.43
Bucaramanga	RandomForestbase	38.66	GRU(InerciayClima)	8.35	78.41

Elaboración propia.

Esto resultados mostraron que:

- La **información climática por sí sola** aportó capacidad explicativa, pero fue insuficiente para reproducir la compleja dinámica de los brotes.
- La **inercia epidemiológica** resultó decisiva para mejorar el ajuste, lo que sugiere que los brotes recientes contienen información crítica para anticipar la evolución inmediata de la serie.

5.3.2. Comparación entre algoritmos por ciudad

La comparación entre algoritmos dentro del Escenario 2 permitió extraer las siguientes conclusiones para cada ciudad:

- Cali: XGBoost fue el modelo con menor RMSE, superando ligeramente al Random Forest. Este resultado indica que, para esta ciudad, un método de gradiente reforzado basado en árboles logró capturar de manera más eficiente las interacciones entre estacionalidad, variables climáticas e inercia epidemiológica.
- Medellín: el Random Forest presentó el mejor desempeño, aunque con una ventaja marginal frente a XGBoost y la GRU. La estructura de los datos y la combinación de

predictores parecen favorecer un ensamble de árboles independientes, con mayor robustez y menor riesgo de sobreajuste que los métodos de boosting o las arquitecturas recurrentes.

- Bucaramanga: la red neuronal recurrente tipo GRU obtuvo el menor RMSE, seguida de cerca por Random Forest y XGBoost. Este resultado sugiere que, en esta ciudad, la modelación explícita de dependencias temporales mediante una arquitectura recurrente aportó ventajas adicionales, posiblemente asociadas a la persistencia de niveles de transmisión relativamente estables.

En conjunto, los resultados evidencian que no existe un único algoritmo dominante para todas las ciudades. Este hallazgo respalda la decisión metodológica de seleccionar el modelo final de forma diferenciada según el contexto urbano. La Tabla 9 presenta el modelo seleccionado para cada ciudad bajo el Escenario 2.

Tabla 10. Modelo seleccionado por ciudad en el Escenario 2.

Ciudad	Modelo seleccionado	RMSE (Esc.2)
Cali	XGBoost	14.94
Medellín	RandomForest	5.20
Bucaramanga	GRU	8.35

Elaboración propia.

5.4. Análisis gráfico de errores y robustez temporal

5.4.1. Ajuste temporal en el periodo de prueba

Además de las métricas numéricas, se realizaron análisis gráficos comparando las series de casos observados y las predicciones de los modelos seleccionados en el periodo de prueba (2018–2019). Para cada ciudad se construyeron gráficos de líneas donde se

superpusieron:

- la curva de **contagios reales**, y
- la curva de **contagios predichos** por el modelo ganador del Escenario 2.

Los resultados cuantitativos globales del Escenario 2 (inercia y clima) se resumen en la Tabla 11, donde se presentan los valores de RMSE y MAE para cada combinación ciudad–modelo.

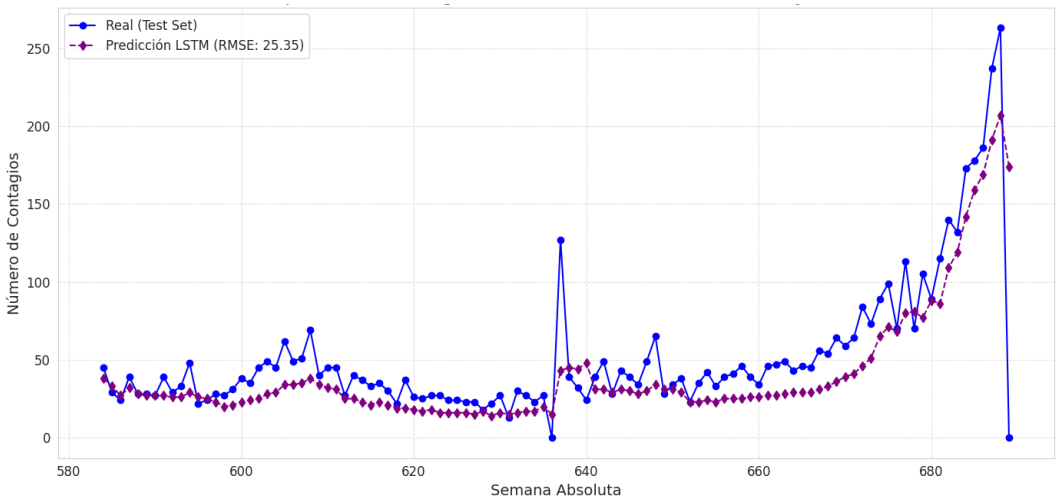
Tabla 11. Rendimiento final de los modelos en el Escenario 2 (inercia y clima) por ciudad.

Ciudad	Modelo	RMSE	MAE
Cali	XGBoost	14.94	8.78
	Random Forest	16.22	9.24
	GRU	24.67	15.84
Medellín	Random Forest	5.20	3.92
	XGBoost	5.27	4.06
	GRU	9.19	7.48
Bucaramanga	GRU	8.13	6.04
	Random Forest	8.79	5.91
	XGBoost	9.29	6.32

Elaboración propia.

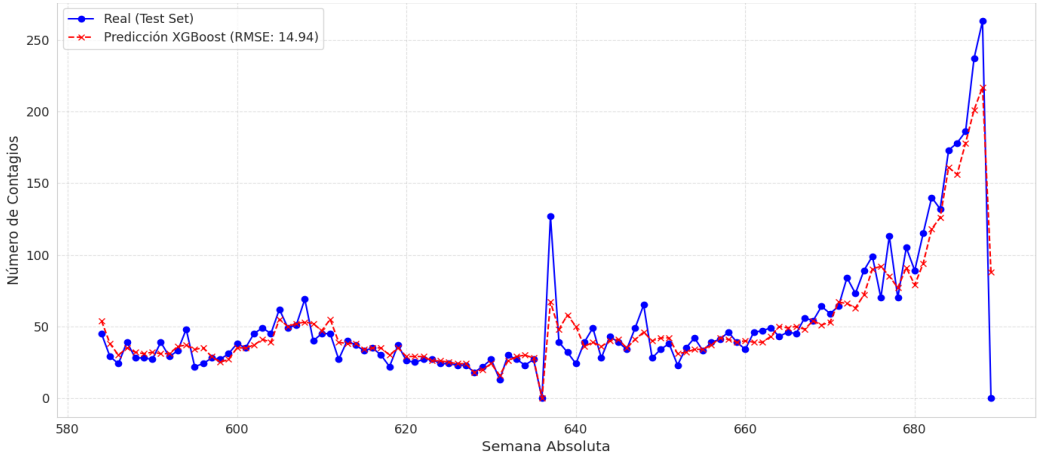
En la Figura 26 a 35, se puede apreciar el comportamiento del modelo seleccionado con relación a las observaciones reales, para las 3 ciudades.

Figura 136. Casos observados y predichos por el modelo GRU (Escenario 2) en Cali (2018–2019).



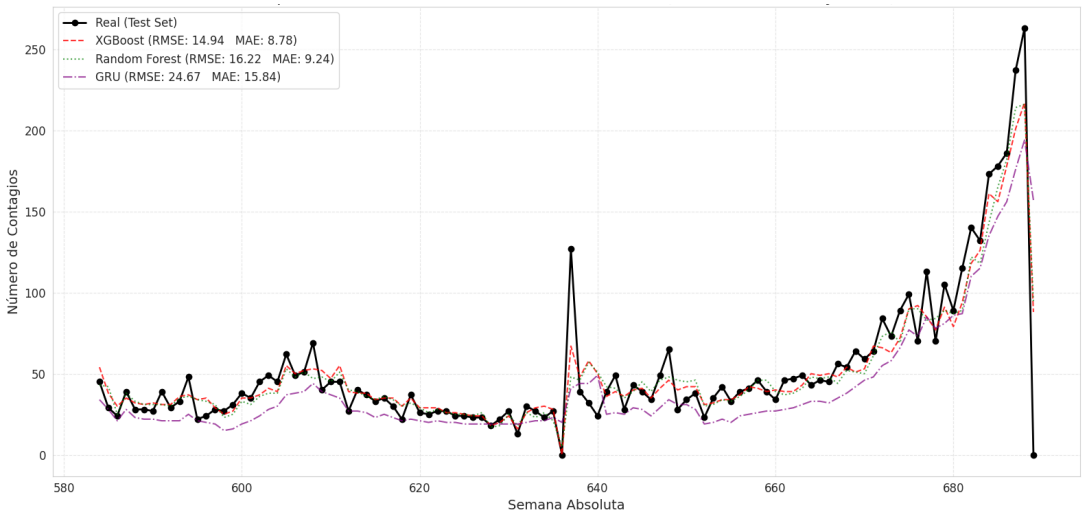
Elaboración propia.

Figura 27. Casos observados y predichos por el modelo XGBoost (Escenario 2) en Cali (2018–2019).



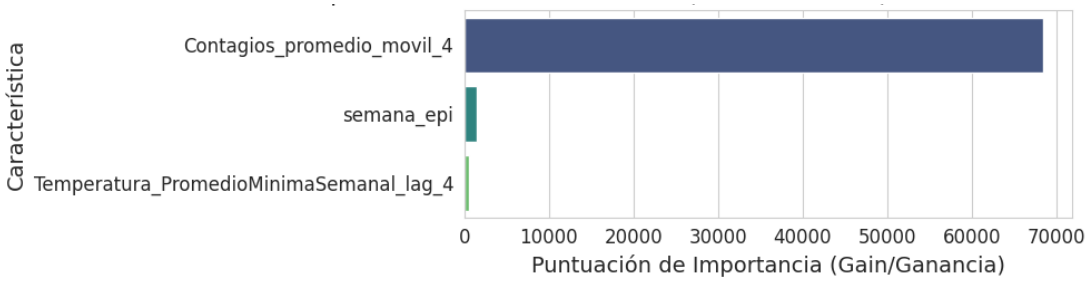
Elaboración propia.

Figura 148. Comparación de Rendimiento de Modelos Finales (Escenario Inercia y Clima) Cali.



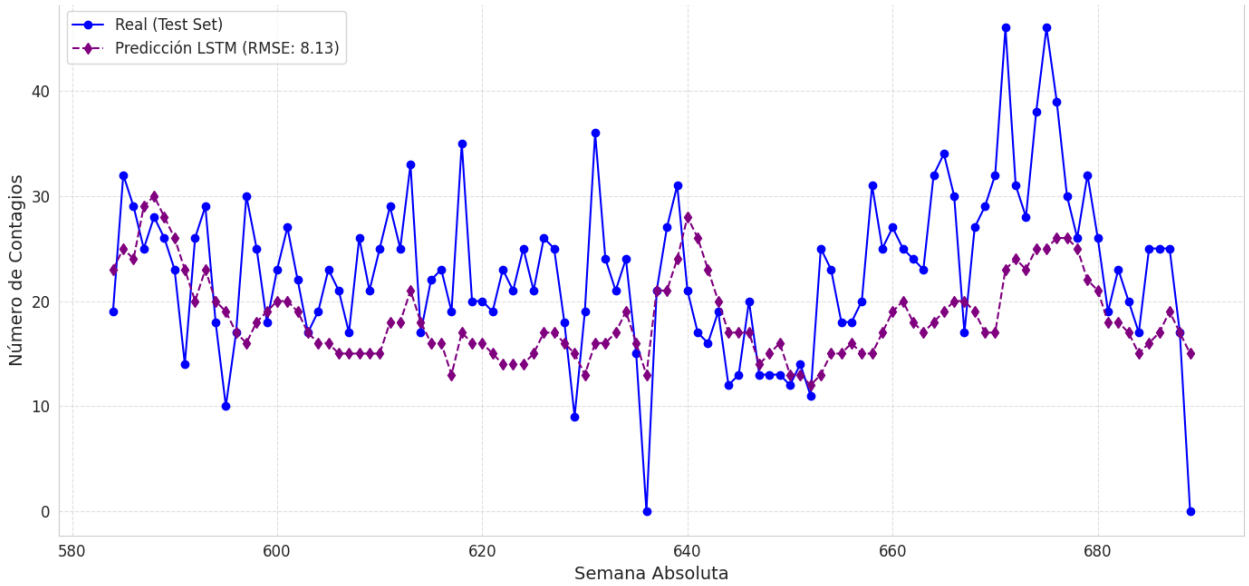
Elaboración propia.

Figura 29. Importancia de características (Ganancia) modelo ganador Cali: XGBoost.



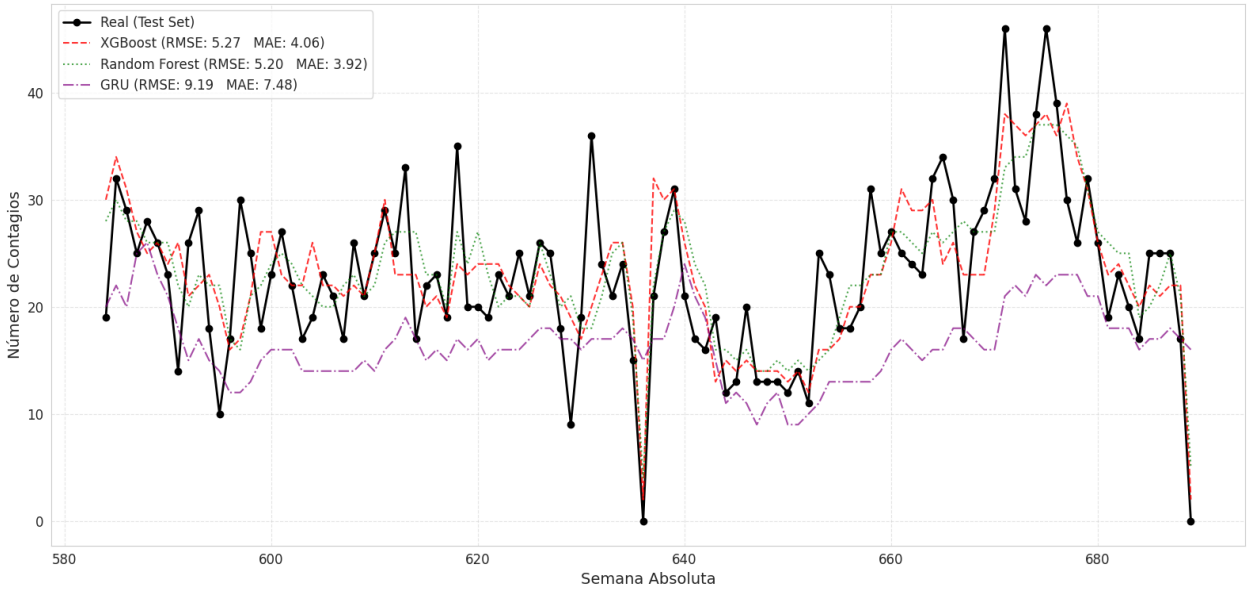
Elaboración propia.

Figura 30. Casos observados y predichos por el modelo GRU (Escenario 2) en Medellín (2018–2019).



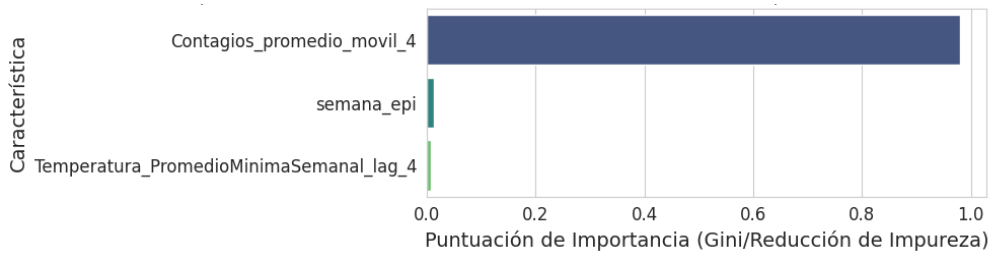
Elaboración propia.

Figura 31. Comparación de Rendimiento de Modelos Finales (Escenario Inercia y Clima) Medellín.



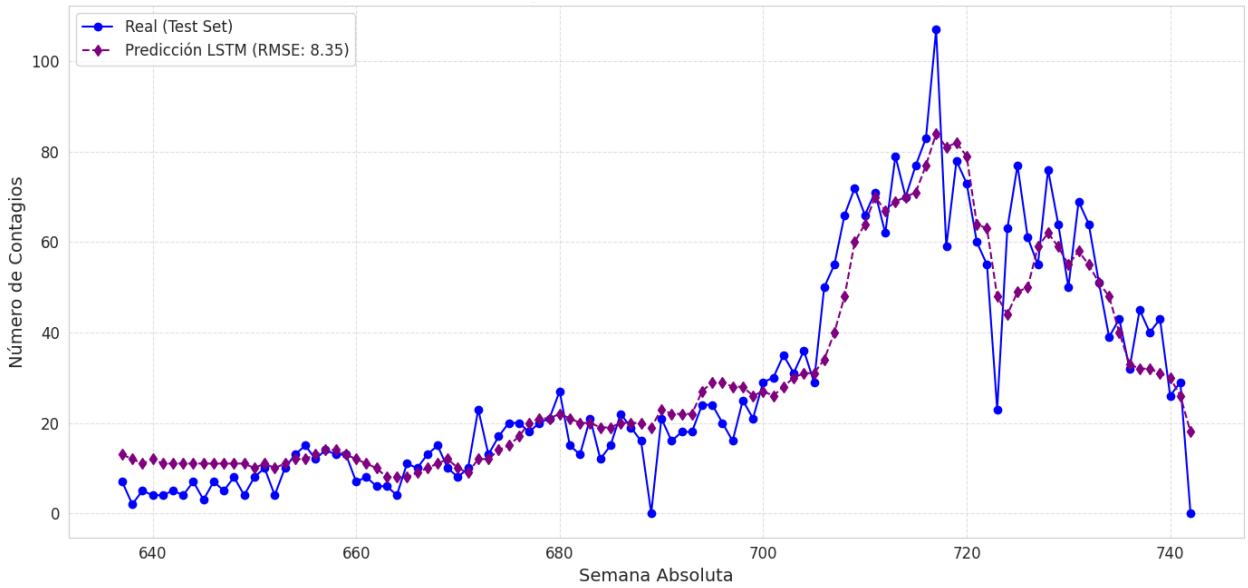
Elaboración propia.

Figura 32. Importancia de características (Ganancia) modelo ganador Medellín: RandomForest.



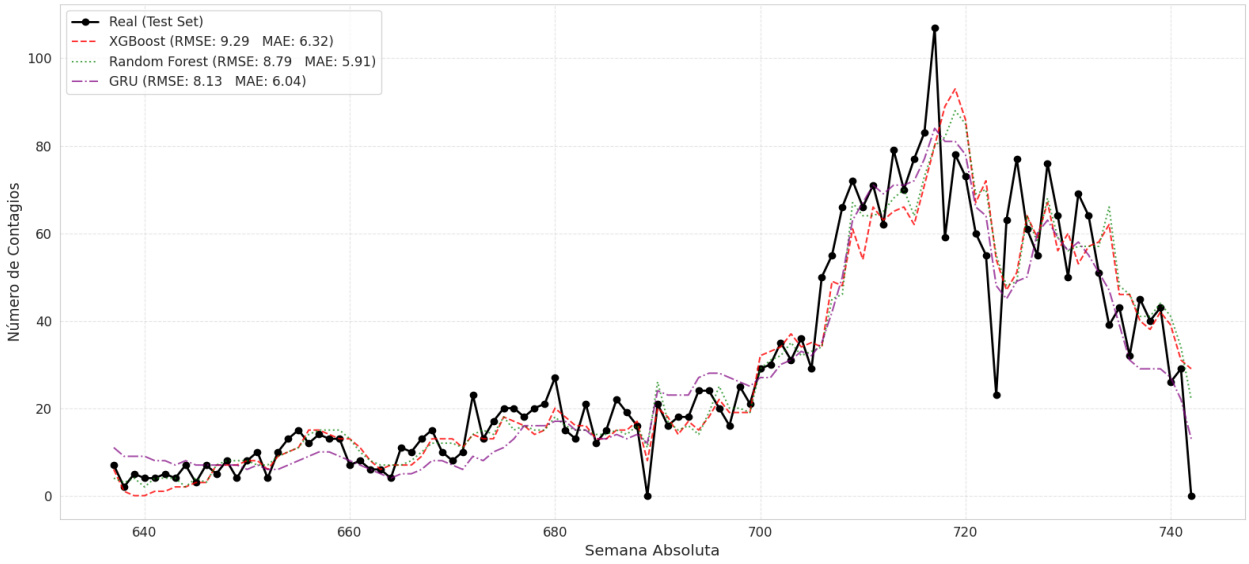
Elaboración propia.

Figura 153. Casos observados y predichos por el modelo GRU (Escenario 2) en Bucaramanga (2018–2019).



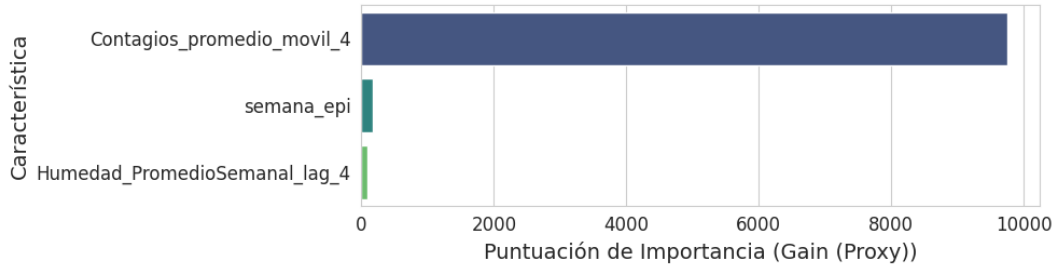
Elaboración propia.

Figura 34. Comparación de Rendimiento de Modelos Finales (Escenario Inercia y Clima) Bucaramanga.



Elaboración propia.

Figura 16. Importancia de características (Ganancia) modelo ganador Bucaramanga: GRU.



Elaboración propia.

Además de las métricas numéricas, se realizaron análisis gráficos comparando las series de casos observados y las predicciones en el periodo de prueba (2018–2019). Para cada

ciudad se construyeron gráficos de líneas en los que se superpuso la curva de contagios reales con las curvas de contagios predichos por los tres modelos evaluados en el Escenario 2 (GRU, Random Forest y XGBoost). Las Figuras 27 (Cali), 30 (Medellín) y 33 (Bucaramanga), junto con las tablas de rendimiento final por ciudad, sintetizan esta comparación.

En Cali, el modelo XGBoost (clima + inercia) presentó el mejor desempeño global, con RMSE ≈ 14,94 y MAE ≈ 8,78. Le siguió el Random Forest (RMSE ≈ 16,22; MAE ≈ 9,24). La red GRU mostró errores mayores (RMSE ≈ 24,67; MAE ≈ 15,84) y tendió a suavizar los picos epidémicos más intensos.

En Medellín, el Random Forest fue el modelo más preciso (RMSE ≈ 5,20; MAE ≈ 3,92). XGBoost obtuvo resultados muy cercanos (RMSE ≈ 5,27; MAE ≈ 4,06), mientras que la GRU mostró un ajuste claramente inferior (RMSE ≈ 9,19; MAE ≈ 7,48).

En Bucaramanga, la GRU se consolidó como el modelo con menor RMSE (RMSE ≈ 8,13; MAE ≈ 6,04). El Random Forest y XGBoost presentaron errores algo mayores (Random Forest: RMSE ≈ 8,79; MAE ≈ 5,91; XGBoost: RMSE ≈ 9,29; MAE ≈ 6,32).

En conjunto, las Figuras 27, 30 y 33 muestran que los modelos reproducen adecuadamente la forma general de los brotes en cada ciudad. En particular, capturan las fases ascendentes y descendentes de las epidemias y distinguen entre semanas de baja y alta transmisión. Las discrepancias más notorias se concentran en algunos picos muy agudos, especialmente en Cali y Bucaramanga. En estos casos, las predicciones tienden a subestimar ligeramente la magnitud máxima. Este patrón es coherente con los valores de RMSE y MAE reportados en las tablas de rendimiento final.

En el Escenario 2, la variable con mayor importancia fue el promedio móvil de cuatro semanas de contagios. Este resultado es consistente con la autocorrelación temporal de la serie. Además, sugiere que la inercia epidemiológica resume no solo la dinámica reciente de transmisión, sino también el efecto acumulado de factores no incluidos

explícitamente. Entre ellos se encuentran condiciones sociodemográficas, dinámicas vectoriales, comportamientos, y medidas de control.

Por tanto, la menor importancia relativa de las variables climáticas no implica que el clima sea irrelevante. Más bien, indica que su efecto se superpone con determinantes estructurales relativamente estables. Estos determinantes quedan parcialmente capturados por el término autorregresivo incluido en el modelo.

5.4.2. Robustez en periodos epidémicos y no epidémicos

La robustez de los modelos se evaluó analizando su comportamiento en dos tipos de periodos:

- **Semanas epidémicas**, definidas como aquellas con niveles de contagios por encima de un umbral alto (por ejemplo, percentil 75 o 90 de la distribución).
- **Semanas no epidémicas**, con niveles bajos o moderados de transmisión.

En las tres ciudades, los modelos del Escenario 2 mantuvieron errores moderados tanto en semanas epidémicas como en semanas no epidémicas. No obstante, la magnitud absoluta del error tendió a ser mayor durante los picos, simplemente porque los niveles de contagios fueron más altos. Aun así, el RMSE logrado en estos periodos resultó consistente con los valores globales reportados en las Tablas 7 y 8.

Este análisis apoyó la idea de que los modelos seleccionados fueron **robustos en el tiempo**, en el sentido de que no solo ajustaron promedios globales, sino que también capturaron de manera razonable la dinámica de ascenso y descenso de los brotes.

5.5. Síntesis y discusión de los modelos seleccionados

En síntesis, la evaluación realizada en este capítulo permitió establecer que:

1. El uso exclusivo de **variables climáticas y estacionalidad anual** (Esc. 1) produjo modelos con capacidad explicativa limitada, aunque ya útiles como línea base.

2. La incorporación de la **inercia epidemiológica** mediante el promedio móvil de contagios (Esc. 2) condujo a **reducciones drásticas del RMSE** en las tres ciudades, con mejoras del orden del 78–95 % frente al modelo base del Esc. 1.

3. Los **modelos ganadores** en el Esc. 2 fueron:

    - **XGBoost** en Cali,
    - **Random Forest** en Medellín, y
    - **GRU** en Bucaramanga,

lo que demuestra la conveniencia de una selección diferenciada de modelos por ciudad.

4. Los análisis gráficos indicaron que los modelos seleccionados fueron capaces de reproducir adecuadamente la **estructura temporal de los brotes**, tanto en periodos epidémicos como no epidémicos, lo que refuerza su utilidad como herramientas de apoyo para la vigilancia y la planificación de respuesta.

6. CONCLUSIONES Y TRABAJOS FUTUROS

6.1. CONCLUSIONES

6.1.1. Conclusiones generales

El trabajo permitió desarrollar y evaluar modelos de aprendizaje automático para la predicción semanal de casos de dengue en Cali, Medellín y Bucaramanga, integrando información histórica de vigilancia epidemiológica y variables climáticas. En conjunto, los resultados muestran que:

1. **Es posible predecir con buena precisión la carga semanal de dengue a corto plazo** cuando se combinan tres componentes clave:

    - la estacionalidad anual, representada por la semana epidemiológica;
    - una variable climática principal con rezago (temperatura mínima o humedad relativa, según la ciudad); y
    - un término de inercia epidemiológica, basado en el promedio móvil de contagios.

2. **Los modelos demostraron capacidad para reproducir la dinámica de los brotes**, capturando de manera razonable las fases de aumento, pico y descenso de la transmisión. Asimismo, lograron diferenciar entre semanas de baja, media y alta incidencia.

3. **No se identificó un único algoritmo dominante para todas las ciudades**. Por el contrario, el modelo con mejor desempeño varió según el contexto urbano, lo que refuerza la necesidad de enfoques diferenciados por ciudad dentro de un marco metodológico común.

6.1.2. Conclusiones respecto a la construcción de los modelos

El objetivo específico de construir modelos de aprendizaje automático utilizando datos históricos de casos de dengue y variables climáticas se cumplió de manera sistemática a través de los siguientes componentes:

1. **Construcción de series semanales integradas por ciudad**, combinando:
   • casos de dengue agregados por año calendario y semana epidemiológica;
   • variables climáticas semanales (humedad relativa, temperaturas mínima, promedio y máxima, y precipitación);

   • variables derivadas, incluyendo rezagos climáticos y promedios móviles de contagios.

2. **Definición de dos escenarios de modelado**, a partir del análisis exploratorio y de estacionalidad:

   • *Escenario 1:* clima + semana epidemiológica.

   • *Escenario 2:* clima + semana epidemiológica + inercia epidemiológica.

3. **Selección de una variable climática principal por ciudad:**

   • temperatura mínima promedio semanal con rezago de cuatro semanas en Cali y Medellín;
   • humedad relativa promedio semanal con rezago de cuatro semanas en Bucaramanga.

4. **Implementación y ajuste de distintas familias de modelos** en las tres ciudades:

   • bosques aleatorios (Random Forest);

   • modelos de gradiente reforzado (XGBoost);

• redes neuronales densas;

• redes neuronales recurrentes tipo GRU.

5. **Coherencia metodológica entre ciudades**, manteniendo la misma lógica de escenarios y la misma partición temporal entrenamiento–prueba. Este diseño permitió realizar comparaciones cruzadas bajo criterios homogéneos.

En síntesis, el Capítulo 4 establece un marco de modelado reproducible y extensible a otras ciudades o periodos. Además, integra de manera explícita la dimensión temporal, climática y epidemiológica del dengue, fortaleciendo la solidez metodológica del estudio.

6.1.3. Conclusiones respecto a la evaluación y robustez

El objetivo de evaluar la precisión y robustez de los modelos predictivos mediante métricas estadísticas apropiadas se abordó utilizando el RMSE sobre un periodo de prueba reciente (2018–2019), complementado con el análisis gráfico de las series observadas y predichas. De este proceso se derivan las siguientes conclusiones:

1. **Escenario 1 (clima + semana epidemiológica).**

   Los modelos lograron capturar parcialmente la variabilidad de los casos, pero presentaron errores elevados. El RMSE superó los 100 casos semanales en Cali y Medellín y se situó alrededor de 37 casos en Bucaramanga. Estos resultados, aunque útiles como línea base, evidencian que la información climática por sí sola no explica completamente la dinámica de los brotes.

2. **Escenario 2 (clima + semana epidemiológica + inercia epidemiológica).**

   La incorporación del término de inercia produjo una mejora sustancial en el desempeño:
   • En Cali, el RMSE se redujo de aproximadamente 155 casos (Random Forest base, Escenario 1) a cerca de 15 casos con XGBoost en el Escenario 2.

• En Medellín, el RMSE pasó de alrededor de 114 casos a aproximadamente 5 casos con Random Forest.

• En Bucaramanga, el RMSE disminuyó de cerca de 39 casos a un rango de 8–9 casos, siendo la GRU el modelo con mejor desempeño.

3. **Mejora porcentual tras incorporar la inercia epidemiológica.**

   • Aproximadamente 90 % de reducción del RMSE en Cali.

   • Cerca de 95 % en Medellín.

   • Más de 78 % en Bucaramanga.

4. **Modelos con mejor desempeño por ciudad (Escenario 2).**

   • Cali: XGBoost (clima + inercia).

   • Medellín: Random Forest (clima + inercia).

   • Bucaramanga: red recurrente GRU (clima + inercia).

5. **Evaluación gráfica de robustez temporal.**

   Los gráficos de series observadas vs. predichas muestran que los modelos seleccionados:
   • reproducen adecuadamente la forma general de los brotes;

   • capturan las fases de crecimiento y descenso;

   • mantienen errores relativamente acotados tanto en semanas epidémicas como en semanas de baja transmisión.

De manera global, la evaluación confirma que la combinación de estacionalidad, clima e inercia epidemiológica, articulada mediante modelos de aprendizaje automático,

constituye una estrategia eficaz para la predicción semanal de dengue.

Desde una perspectiva práctica, el modelo podría ejecutarse semanalmente con datos actualizados y generar proyecciones para las siguientes cuatro semanas. Si el valor esperado supera un umbral histórico previamente definido, podría activarse una alerta para intensificar el control vectorial, fortalecer campañas comunitarias y preparar la capacidad hospitalaria.

En este sentido, el modelo se concibe como una herramienta de apoyo a la toma de decisiones, y no como un sustituto del sistema de vigilancia epidemiológica.

6.2. TRABAJOS FUTUROS

6.2.1. Implicaciones para la vigilancia y la gestión del riesgo

Los hallazgos del estudio tienen implicaciones relevantes para la vigilancia epidemiológica y la gestión del riesgo de dengue:

1. **Integración en sistemas de alerta temprana.**

   Los modelos desarrollados pueden incorporarse como componente de sistemas de alerta temprana, al proporcionar predicciones de corto plazo sobre la carga esperada de casos. Esto permitiría:

   • anticipar el refuerzo de acciones de control vectorial;

   • planificar recursos asistenciales (insumos, camas y personal);

   • orientar campañas de comunicación y educación en salud.

2. **Factibilidad operativa.**

   Los modelos requieren variables relativamente accesibles: casos semanales, una variable climática agregada y promedios móviles simples. Esta característica facilita

su implementación en entornos institucionales con capacidades técnicas limitadas, siempre que existan rutinas claras de actualización y validación de datos.

3. **Adaptación a contextos locales.**

   La necesidad de seleccionar modelos específicos por ciudad evidencia que las dinámicas locales del dengue —relacionadas con clima, urbanización, movilidad y condiciones socioeconómicas— son determinantes. En consecuencia, las herramientas predictivas deben adaptarse a estos contextos para resultar realmente útiles en la práctica.

6.2.2. Limitaciones del estudio

A pesar de los resultados favorables, el estudio presenta varias limitaciones que es importante reconocer:

1. **Alcance de las variables explicativas.**

   Los modelos se centraron en variables climáticas y en la inercia de la serie de casos. No se incorporaron de manera explícita otros factores relevantes, tales como:

   • indicadores socioeconómicos;

   • condiciones de vivienda y acceso a servicios básicos;

   • información detallada sobre intervenciones de control vectorial;

   • movilidad de la población.

2. **Parcimonia en la selección climática.**

   La elección de una única variable climática principal por ciudad, aunque metodológicamente parsimoniosa y respaldada por el análisis exploratorio, pudo omitir interacciones entre múltiples variables climáticas. La inclusión simultánea de

varios predictores ambientales podría enriquecer los modelos en futuros desarrollos.

3. **Horizonte del periodo de prueba.**

   Los modelos se evaluaron sobre un periodo relativamente corto (2018–2019). Si bien esta decisión garantiza independencia temporal, sería deseable contar con periodos de prueba más extensos, incluyendo años con epidemias de distinta magnitud y características.

4. **Búsqueda de hiperparámetros.**

   La calibración de hiperparámetros se realizó de manera sistemática. No obstante, no se exploró exhaustivamente todo el espacio de búsqueda posible. En consecuencia, podrían existir configuraciones alternativas con mejoras marginales en el desempeño.

5. **Predominio del término de inercia epidemiológica.**

   La alta importancia del promedio móvil de contagios sugiere la necesidad de incorporar variables adicionales en futuros trabajos. Estas deberían capturar de forma más explícita determinantes estructurales como factores sociodemográficos, presencia vectorial, intervenciones de control y movilidad poblacional. De este modo sería posible desagregar mejor qué parte de la "memoria" de la serie se asocia al clima y cuál responde a otros determinantes contextuales.

6. **Limitaciones en variables socioeconómicas.**

   La incorporación de indicadores socioeconómicos estuvo restringida por su disponibilidad, periodicidad y comparabilidad entre ciudades. Muchas variables demográficas finas (por ejemplo, edad o sexo desagregado) no están disponibles con resolución semanal o presentan dificultades de integración. Asimismo, pueden

existir barreras institucionales para el acceso y consolidación de datos provenientes de EPS u otras fuentes dispersas. Por estas razones, en esta fase se priorizó un núcleo robusto centrado en información epidemiológica y climática, dejando la integración demográfica detallada como línea futura de investigación.

6.2.3. Recomendaciones y líneas de trabajo futuro

A partir de las conclusiones y limitaciones identificadas, se proponen las siguientes recomendaciones y líneas de investigación futura:

1. **Ampliación del marco de modelado.**

   Se recomienda extender el conjunto de predictores incorporando:

   • variables socioeconómicas y ambientales (densidad poblacional, cobertura de servicios públicos, uso del suelo);

   • indicadores operativos de control vectorial;

   • información sobre circulación simultánea de otros arbovirus (dengue, Zika, chikunguña).

   Esta ampliación permitiría capturar determinantes estructurales actualmente absorbidos por el término de inercia epidemiológica.

2. **Modelos multiciudad y componente espacial.**

   Se propone explorar enfoques multivariados que integren información de varias ciudades de forma conjunta. Esto permitiría capturar posibles correlaciones espaciales y efectos de movilidad entre territorios conectados.

3. **Integración en plataformas de vigilancia en tiempo real.**

   Un paso clave consiste en evaluar la incorporación de los modelos en sistemas operativos de vigilancia, incluyendo:

• automatización de la actualización de datos;

• generación rutinaria de pronósticos semanales;

• desarrollo de interfaces visuales para tomadores de decisión.

Esto permitiría trasladar el modelo desde el ámbito académico hacia un uso aplicado en salud pública.

4. **Calibración operacional y definición de umbrales.**

   Se recomienda realizar estudios específicos para convertir predicciones continuas en niveles de riesgo (bajo, medio, alto). La definición de umbrales de alerta permitiría evaluar el impacto potencial del modelo en la reducción de la carga de enfermedad y optimizar la activación de intervenciones.

5. **Comparación con enfoques metodológicos emergentes.**

   Futuras investigaciones podrían contrastar los modelos desarrollados con:
   • modelos híbridos (ARIMA–Machine Learning);

   • arquitecturas profundas más complejas (LSTM bidireccionales, Transformers);

   • modelos probabilísticos que incorporen estimaciones explícitas de incertidumbre.

6. **Métricas de error relativo y comparabilidad entre ciudades.**

   Además del reporte numérico de RMSE, MAE y MSE, es importante interpretar qué representan estas diferencias en términos operativos para la vigilancia epidemiológica. En la práctica, una mejora del error (por ejemplo, una reducción del RMSE) implica que las predicciones semanales se aproximan más a los valores observados, lo cual puede apoyar decisiones como **anticipar refuerzos de control vectorial**, **dimensionar recursos asistenciales** y **ajustar la comunicación del riesgo**. Esto es particularmente crítico en semanas de alta transmisión, donde una

subestimación puede retrasar la activación de alertas, mientras que una sobreestimación puede inducir asignaciones de recursos innecesarias.

No obstante, estas métricas globales resumen el desempeño promedio y no describen de forma explícita la capacidad del modelo para responder ante eventos extremos ni la estabilidad del rendimiento a lo largo del tiempo. Por ello, como línea de trabajo futuro se propone complementar la evaluación con:

- Análisis específicos en **semanas pico** (por ejemplo, error condicionado a semanas por encima de umbrales históricos o por cuantiles),
- Esquemas de validación temporal más exigentes (ventanas deslizantes o *rolling-origin*) para evaluar estabilidad, y
- Métodos para cuantificar **incertidumbre** (intervalos de predicción o técnicas de remuestreo) que permitan interpretar los pronósticos con márgenes de confianza.

Asimismo, se reconoce la importancia de métricas de **error relativo** para facilitar la comparabilidad entre ciudades con diferentes escalas de contagio. En principio, el **MAPE (Mean Absolute Percentage Error)** permitiría expresar el desempeño en términos porcentuales; sin embargo, en este estudio su uso no se priorizó porque las series semanales incluyen valores cercanos a cero, situación en la que el MAPE puede volverse inestable o indefinido y distorsionar la interpretación en periodos interepidémicos. Como extensión futura, se plantea incorporar alternativas más robustas como el **sMAPE (Symmetric Mean Absolute Percentage Error)** o el **WMAPE (Weighted Mean Absolute Percentage Error)**, que permiten normalizar el error sin los inconvenientes numéricos de la división por cero y favorecen comparaciones interurbanas más consistentes [19, 20].

7. REFERENCIAS BIBLIOGRÁFICAS

[1] Instituto Nacional de Salud (INS). (2024). **Dengue. Informe de evento, Colombia, 2024.** Bogotá: INS

https://www.ins.gov.co/buscador-eventos/Informesdeevento/DENGUE%20INFORME%20DE%20EVENTO%202024.pdf

[2] Instituto Nacional de Salud (INS). (2024). **Dengue. Informe de evento. Periodo epidemiológico II, 2024.** Bogotá: INS

https://www.ins.gov.co/buscador-eventos/Informesdeevento/DENGUE%20PE%20II%202024.pdf

[3] Instituto Nacional de Salud (INS). (2024). **Boletín Epidemiológico Semanal. Semana epidemiológica 47, año 2024.** Bogotá: INS

https://www.ins.gov.co/buscador-eventos/BoletinEpidemiologico/2024_Boletin_epidemiologico_semana_47.pdf

[4] **Alcaldía de Santiago de Cali – Secretaría de Salud Pública.** «Descenso continuo de casos de dengue en Cali refleja éxito de medidas de prevención y control de la Alcaldía». Boletín/nota de prensa, 2024.

https://www.cali.gov.co/salud/publicaciones/182894/la-curva-del-dengue-sigue-en-descenso-en-cali-gracias-a-la-intensificacion-de-acciones-para-controlar-el-zancudo-transmisor

[5] **Universidad Icesi.** *Estrategias para combatir el dengue: recomendaciones de política pública para Colombia*. Políticas en Breve No. 11. Cali: Universidad Icesi; 2023.

https://www.icesi.edu.co/proesa/images/publicaciones/politicas-en-breve/pdf/proesa-pb-11.pdf

[6] World Health Organization. (2025, August 21). *Dengue and severe dengue* [Fact sheet]. World Health Organization.

https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue

[7] Instituto de Hidrología, Meteorología y Estudios Ambientales. (s. f.). *Temperatura ambiente del aire* [Conjunto de datos]. Datos Abiertos Colombia.

https://www.datos.gov.co/Ambiente-y-Desarrollo-Sostenible/Temperatura-Ambiente-del-Aire/sbwg-7ju4

[8] Instituto de Hidrología, Meteorología y Estudios Ambientales. (s. f.). *Humedad del aire* [Conjunto de datos]. Datos Abiertos Colombia.

https://www.datos.gov.co/Ambiente-y-Desarrollo-Sostenible/Humedad-del-Aire/uext-mhny

[9] Instituto de Hidrología, Meteorología y Estudios Ambientales. (s. f.). *Precipitación* [Conjunto de datos]. Datos Abiertos Colombia.

https://www.datos.gov.co/Ambiente-y-Desarrollo-Sostenible/Precipitaci-n/s54a-sgyg

[10] Universidad de Antioquia – Facultad Nacional de Salud Pública.

«*Dengue en Colombia: brotes de enfermedades, entre las alertas y la respuesta*». Artículo web de divulgación; 2025.

https://www.udea.edu.co/wps/portal/udea/web/generales/interna/Unidades%2BAcad%21c3%21a9micas/Salud%2BP%21c3%21bablica/asContenidos/listado/Dengue_2024

[11] Morin, C. W., Comrie, A. C. & Ernst, K. (2013). *Climate and dengue transmission: Evidence and implications. Environmental Health Perspectives*, 121(11–12): 1264-1272.

[12] Lambrechts, L., Paaijmans, K.P., Fansiri, T., Carrington, L.B., Kramer, L.D., Thomas, M.B. & Scott, T.W. (2011). *Impact of daily temperature fluctuations on dengue virus transmission by Aedes aegypti. Proceedings of the National Academy of Sciences of the United States of America*, 108(18): 7460–7465.

[13] Zeledón, L.C. (2025). *Determinantes sociales en la incidencia de dengue en América Latina*. Revista Científica de Salud BIOSANA, 5(1), 35–49.

[14] Sexton J, Russell T, Burkot TR, Craig A, Hickson RI (2025*) Investigating linkages between human movement and meteorological variables on dengue outbreaks in the Pacific Islands*. PLOS Neglected Tropical Diseases 19(10): e0013607

[15] Hethcote, H. W. (2000). *The mathematics of infectious diseases. SIAM Review, 42*(4), 599–653.

[16] Zhao, N., Charland, K., Carabali, M., Nsoesie, E. O., Maheu-Giroux, M., Rees, E., Winskell, K. (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.

[17] Montenegro Torres, J. F. (2023). *Sistema de predicción de casos sub-estacional de dengue en Colombia utilizando un modelo LSTM Seq2Seq.* Universidad de los Andes.

[18] Leung, X. Y., Islam, R. M., Adhami, M., Ilic, D., McDonald, L., Palawaththa, S., Diug, B., Munshi, S. U., & Karim, M. N. (2023). *A systematic review of dengue outbreak prediction models: Current scenario and future directions.* PLoS Neglected Tropical Diseases, 17(2), e0010631.

[19] Hyndman, R. J., & Koehler, A. B. (2006). *Another look at measures of forecast accuracy*. International Journal of Forecasting, 22(4), 679–688.

[20] Goodwin, P., & Lawton, R. (1999). *On the asymmetry of the symmetric MAPE.* International Journal of Forecasting, 15(4), 405–408.