



Maestría en:

INTELIGENCIA ARTIFICIAL APLICADA

Trabajo previo a la obtención de título de:

Magister en Inteligencia Artificial Aplicada

AUTORES

Segundo Ramiro Maila Asqui
Jair Jazmani Morales Martínez

TUTORES

Bryan Steven Vinueza Bustamante
Andrés Roberto Navas Castellanos

TEMA

Desarrollo y evaluación de un modelo de predicción espacial del dengue mediante machine learning geoespacial en la ciudad de Guayaquil.

Certificación de Autoría

Nosotros, Segundo Ramiro Maila Asqui y Jair Jazmani Morales Martínez, declaramos bajo juramento que el trabajo aquí descrito es de nuestra autoría; que no ha sido presentado anteriormente para ningún grado o calificación profesional y que se ha consultado la bibliografía detallada.

Cedemos nuestros derechos de propiedad intelectual a la Universidad Internacional del Ecuador (UIDE), para que sea publicado y divulgado en internet, según lo establecido en la Ley de Propiedad Intelectual, su reglamento y demás disposiciones legales.



Firma

Segundo Ramiro Maila Asqui



Firma

Jair Jazmani Morales Martínez

Autorización de Derechos de Propiedad Intelectual

Nosotros, Segundo Ramiro Maila Asqui y Jair Jazmani Morales Martínez, en calidad de autores del trabajo de investigación titulado **Desarrollo y evaluación de un modelo de predicción espacial del dengue mediante machine learning geoespacial en la ciudad de Guayaquil**, autorizamos a la Universidad Internacional del Ecuador (UIDE) para hacer uso de todos los contenidos que nos pertenecen o de parte de los que contiene esta obra, con fines estrictamente académicos o de investigación. Los derechos que como autores nos corresponden, lo establecido en los artículos 5, 6, 8, 19 y demás pertinentes de la Ley de Propiedad Intelectual y su Reglamento en Ecuador.

D. M. Quito, julio de 2026



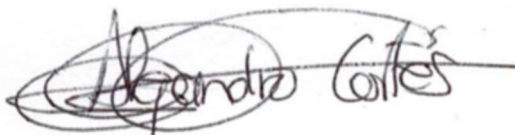
Firma
Segundo Ramiro Maila Asqui



Firma
Jair Jazmani Morales Martínez

Aprobación de dirección y coordinación del programa

Nosotros, **Alejandro Cortés López** Director EIG y **Karla Estefanía Mora Cajas** Coordinadora UIDE, declaramos que: **Segundo Ramiro Maila Asqui** y **Jair Jazmani Morales Martínez**, son los autores exclusivos de la presente investigación y que ésta es original, auténtica y personal de ellos.



Alejandro Cortés López Director de la
Maestría en Inteligencia Artificial
Aplicada



Karla Estefanía Mora Cajas
Coordinadora de la
Maestría en Inteligencia Artificial Aplicada

DEDICATORIA

A Dios, por darme la salud, la sabiduría y la fortaleza para alcanzar una meta más en mi vida profesional y personal.

A mi esposa Blanky Sofía, el amor de toda mi vida, por su paciencia y su apoyo incondicional en cada jornada de desvelo. Este logro también es tuyo.

A mis queridas hijas, Giomara y Emily, gracias por ser mi mayor orgullo y la motivación constante para demostrarles que el camino del conocimiento y la superación no tienen límites.

A mi adorada nieta, Cielito, la luz y la ternura de mis días; que este esfuerzo sea un legado de amor para su futuro.

A mis amados padres, por sus bendiciones, su amor eterno y por enseñarme desde la infancia el valor del trabajo honesto y la perseverancia.

A mis hermanos Rodrigo y Manuel, por el lazo inquebrantable de sangre, la complicidad de vida y por compartir la alegría de este nuevo paso.

A todas aquellas personas que, junto a mi yerno Andrés, me brindaron una palabra de aliento un consejo y un gesto de fe sincera.

Con profundo amor y gratitud, dedico este triunfo a todos ustedes.

Segundo Ramiro Maila Asqui

Dedico este trabajo a mis padres, por su amor incondicional, sus enseñanzas y el esfuerzo constante que realizaron para brindarme las herramientas necesarias para crecer personal y profesionalmente. Gracias por creer en mí, por impulsarme a seguir adelante en cada desafío y por ser siempre el pilar fundamental de mi vida.

A mis hermanos, por su apoyo, compañía y confianza a lo largo de este camino. Su respaldo y palabras de aliento han sido una motivación permanente para alcanzar cada una de mis metas.

A mi esposa, Dayra, por su amor, paciencia y comprensión durante cada etapa de este proceso. Gracias por acompañarme en los momentos más difíciles, celebrar cada logro y ser mi compañera incondicional en este camino de crecimiento y superación.

A mis hijos, Derek y Emma, quienes son mi mayor inspiración y el motor que impulsa cada uno de mis esfuerzos. Todo lo que hago busca ser un ejemplo para ustedes y construir un mejor futuro para nuestra familia. Su amor y sus sonrisas me recuerdan cada día la importancia de nunca rendirme.

Y de manera muy especial, a mi abuelita, quien hoy me acompaña desde el cielo. Sé que estaría profundamente orgullosa de verme alcanzar este logro, un sueño que siempre deseó presenciar. Aunque ya no esté físicamente a mi lado, su amor, sus consejos y los valores que sembró en mí continúan guiando mis pasos. Este logro también es para ella, con gratitud, cariño y eterna admiración.

Jair Jazmani Morales Martínez

AGRADECIMIENTOS

A la Universidad Internacional del Ecuador (UIDE - EIG) y a mi querido barrio San Miguel de Collacoto, por el convenio interinstitucional que hizo posible mi acceso a esta beca de estudios, brindándome la valiosa oportunidad de culminar con éxito esta Maestría en Inteligencia Artificial Aplicada.

Al Ministerio de Salud Pública (MSP) del Ecuador, por facilitar el acceso al conjunto de datos (*dataset*) esencial para el desarrollo práctico de esta investigación. Su valiosa contribución no solo hizo viable este proyecto, sino que fundamentó la aplicación de la inteligencia artificial al servicio del bienestar social.

A cada uno de los docentes de la maestría, por compartir su sabiduría y experiencia, impulsándome a expandir las fronteras de mi conocimiento profesional.

Finalmente, a mi compañero de cohorte Jair Morales por el intercambio de ideas, el trabajo en equipo y el apoyo mutuo durante este exigente camino académico.

Segundo Ramiro Maila Asqui

Quiero expresar mi más sincero agradecimiento a mi esposa, Dayra, por su amor, paciencia y apoyo incondicional durante todo este proceso académico. Gracias por acompañarme y motivarme a alcanzar esta meta.

A mis hijos, Derek y Emma, quienes son mi mayor inspiración y el motor que impulsa cada uno de mis esfuerzos. Este logro también es para ustedes.

A mi compañero de maestría, Segundo Maila, por el compañerismo, el trabajo en equipo y el intercambio de conocimientos que enriquecieron esta experiencia académica.

A los docentes de la Universidad Internacional del Ecuador (UIDE) y de EIG, por compartir sus conocimientos, experiencia y orientación, fundamentales para mi formación profesional y el desarrollo de esta investigación.

Finalmente, agradezco a todas las personas que, de una u otra manera, contribuyeron con su apoyo y confianza para la culminación exitosa de esta maestría.

Jair Jazmani Morales Martínez

RESUMEN

El dengue constituye una de las enfermedades transmitidas por vectores con mayor incidencia en el Ecuador, representando un desafío constante para el sistema de salud pública debido a su relación con factores climáticos, ambientales y demográficos. La identificación temprana de zonas con alto riesgo de propagación resulta fundamental para la planificación de estrategias de prevención y control.

El presente proyecto propone el desarrollo de un sistema de análisis espacial de enfermedades mediante técnicas de Machine Learning Geoespacial, tomando como caso de estudio la distribución de casos de dengue en las parroquias urbanas de la ciudad de Guayaquil, Ecuador. Para ello, se integraron datos epidemiológicos, variables climáticas (temperatura, precipitación y humedad), información demográfica y datos geográficos provenientes de fuentes abiertas y organismos oficiales, construyendo un dataset espaciotemporal de 3,900 observaciones correspondientes a 15 parroquias urbanas durante el periodo 2021–2025.

La metodología contempla la recopilación, limpieza e integración de datos geoespaciales, seguida de un análisis exploratorio para identificar patrones de distribución y concentración de casos. Posteriormente, se aplicaron algoritmos de aprendizaje automático como OLS, Ridge, Lasso, ElasticNet, Random Forest y XGBoost para modelar y predecir zonas de riesgo epidemiológico. Adicionalmente, se utilizaron técnicas de análisis espacial (índice de Moran y Getis-Ord G_i^*) para detectar áreas críticas y visualizar los resultados mediante mapas.

Como resultado se obtuvo un modelo Random Forest que mejora el desempeño predictivo ($MAE = 0.6164$, $RMSE = 1.9432$, $R^2 = 0.6405$), identificando que la inercia epidemiológica y la temperatura mínima son los predictores más relevantes. Los mapas de zonas de riesgo y residuales permiten apoyar la toma de decisiones de entidades de salud pública, contribuyendo a una gestión más eficiente de los recursos destinados a la vigilancia y control epidemiológico.

Palabras Claves: Machine Learning Geoespacial, Dengue, Inteligencia Artificial, Análisis Espacial, Salud Pública, Sistemas de Información Geográfica (SIG), Predicción Epidemiológica, GeoAI, Guayaquil.

ABSTRACT

Dengue represents one of the vector-borne diseases with the highest incidence in Ecuador, posing a constant challenge to the public health system due to its correlation with climatic, environmental, and demographic factors. Early identification of high-risk propagation areas is fundamental for planning prevention and control strategies.

This project proposes the development of a spatial disease analysis system utilizing Geospatial Machine Learning techniques, **taking the distribution of dengue cases in the urban parishes of Guayaquil, Ecuador, as the case study**. To achieve this, epidemiological data, climatic variables (temperature, precipitation, and humidity), demographic information, and geographic data from open sources and official agencies were integrated, constructing a spatial-temporal dataset of 3,900 observations covering 15 urban parishes over the period 2021–2025.

The methodology encompasses geospatial data collection, cleaning, and integration, followed by an exploratory analysis to identify case distribution and concentration patterns. Subsequently, machine learning algorithms such as OLS, Ridge, Lasso, ElasticNet, Random Forest, and XGBoost were applied to model and predict epidemiological risk zones. Additionally, spatial analysis techniques (Moran's I and Getis-Ord Gi*) were utilized to detect critical areas and visualize the results through maps.

A Random Forest model was obtained, improving predictive performance (MAE = 0.6164, RMSE = 1.9432, $R^2 = 0.6405$), identifying that epidemiological inertia and minimum temperature are the most relevant predictors. The risk zone and residual maps can support decision-making for public health entities, contributing to more efficient management of resources allocated to epidemiological surveillance and control.

Keywords: Geospatial Machine Learning, Dengue, Artificial Intelligence, Spatial Analysis, Public Health, Geographic Information Systems (GIS), Epidemiological Prediction, GeoAI, Guayaquil.

TABLA DE CONTENIDOS

CAPÍTULO 1:.....	1
1 INTRODUCCIÓN	1
1.1 Definición del problema.....	1
1.2 Justificación.....	2
1.2.1 Justificación académica.	2
1.2.2 Justificación práctica.....	3
1.2.3 Justificación social.	3
1.2.4 Justificación tecnológica.....	3
1.3 Alcance del proyecto.....	3
1.4 Objetivos.....	6
1.4.1 Objetivo General.....	6
1.4.2 Objetivos específicos	6
1.5 Hipótesis y variables	7
CAPÍTULO 2.....	11
2 Fundamentos Epidemiológicos y Contexto de la Enfermedad	11
2.1 Definición y Agente Causal	11
2.2 El Vector: Aedes Aegypti y su Ecología Urbana.....	11
2.3 Formas de Transmisión y Dinámica del Brote.....	12
2.4 Impacto en la Salud Pública y Relevancia Urbana	13

2.5	Situación Epidemiológica en el Ecuador y Ciudades Comparables.....	13
2.6	Factores Ambientales, Climáticos y Territoriales Asociados al Dengue	13
2.6.1	Determinantes Climáticos y su Influencia Biológica.....	14
2.6.2	Factores Territoriales y de Infraestructura Urbana	14
2.6.3	Interacción Dinámica y Efectos Rezagados (Lags).....	15
2.6.4	Epidemiología Espacial y Análisis Geoespacial en Salud	16
2.6.5	Autocorrelación Espacial: La Primera Ley de Tobler	16
2.6.6	Clústeres, Hotspots y Zonas de Riesgo	17
2.6.7	El Rol de los SIG en la Vigilancia Epidemiológica Moderna	17
2.6.8	Utilidad en Salud Pública y Justificación del Modelo	18
2.7	Fuentes de Datos y Variables de Estudio	18
2.7.1	Datos Epidemiológicos (Variable Target)	18
2.7.2	Datos Climáticos y Ambientales (Predictores).....	19
2.7.3	Datos Demográficos y Territoriales (Normalización)	19
2.8	Modelos tradicionales para el análisis y predicción del dengue	20
2.8.1	Vigilancia descriptiva y análisis retrospectivo	20
2.8.2	Modelos de regresión para datos de conteo.....	21
2.8.3	Modelos de series temporales	22
2.8.4	Ventajas de los enfoques tradicionales	24
2.8.5	Limitaciones de los enfoques tradicionales.....	25
2.9	Machine learning aplicado a la predicción epidemiológica.....	26

2.9.1	Definición general de machine learning	26
2.9.2	Machine learning en epidemiología	27
2.9.3	Ventajas del machine learning frente a enfoques tradicionales	28
2.9.4	Aplicación del Machine Learning al Dengue	30
2.9.5	Limitaciones y Precauciones en el uso del Machine Learning.....	32
2.9.6	Relación con la presente investigación.....	33
2.10	Random Forest y XGBoost como modelos de predicción	34
2.10.1	Random Forest	34
2.10.2	XGBoost	36
2.10.3	Justificación de selección	38
2.10.4	Comparación breve con otras alternativas.....	40
2.11	Estado del arte comparativo	41
2.11.1	Estudios con modelos tradicionales	41
2.11.2	Estudios con machine learning.....	43
2.11.3	Estudios con análisis espacial del dengue	46
2.12	Vacío de investigación	48
2.13	Síntesis del capítulo	50
3	METODOLOGÍA Y DESARROLLO	51
3.1	Enfoque de la investigación	51
3.2	Tipo, nivel y diseño de investigación.....	51
3.3	Ámbito y unidad de estudio.....	52

3.4	Fuentes de datos.....	53
3.4.1	Fuente epidemiológica: Ministerio de Salud Pública (MSP).....	53
3.4.2	Fuente ambiental: NASA POWER	55
3.4.3	Fuente demográfica: INEC	57
3.5	Descripción del dataset.....	57
3.6	Variables del estudio.....	58
3.6.1	Variable dependiente (target)	58
3.6.2	Variables independientes (predictores)	59
3.7	Preprocesamiento e integración de datos	60
3.8	Análisis exploratorio de datos.....	62
3.8.1	Distribución de la variable dependiente	62
3.8.2	Patrones temporales	63
3.8.3	Variabilidad espacial entre parroquias	64
3.8.4	Correlaciones entre variables	66
3.9	Evaluación y selección de modelos predictivos	68
3.9.1	Evaluación de la linealidad y selección de modelos	68
3.9.2	Estrategia de validación	68
3.9.3	Modelos evaluados.....	69
3.9.4	Métricas de evaluación.....	70
3.9.5	Procedimiento de optimización de hiperparámetros.....	71
3.9.6	Criterio de selección del modelo final	72

3.10	Técnicas de análisis espacial	73
3.11	Herramientas tecnológicas	74
3.12	Consideraciones éticas y limitaciones metodológicas	75
3.12.1	Consideraciones éticas	75
3.12.2	Limitaciones metodológicas	76
3.13	Síntesis del Capítulo 3	79
CAPÍTULO 4:.....		81
4	RESULTADOS	81
4.1	Caracterización del dataset de modelado	81
4.2	Análisis exploratorio de datos (EDA).....	82
4.2.1	Distribución de la variable dependiente	82
4.2.2	Patrones temporales	83
4.2.3	Variabilidad espacial entre parroquias	84
4.2.4	Correlaciones entre variables	85
4.2.5	Síntesis del análisis exploratorio.....	87
4.3	Modelos baseline (Modelos lineales)	87
4.3.1	Modelo OLS (Baseline)	88
4.3.2	Modelo Ridge	90
4.3.3	Modelo Lasso.....	93
4.3.4	Modelo ElasticNet.	95
Resumen de Modelos Baseline		98

4.4	Evaluación de la linealidad (Prueba de Ramsey RESET).....	98
4.5	Modelos basados en árboles	101
4.5.1	Random Forest: optimización y evaluación.....	101
4.5.2	Evaluación de XGBoost	103
4.5.3	Selección del modelo final.....	104
4.5.4	Prueba de Diebold-Mariano (comparación de precisión predictiva)	105
4.5.5	Importancia de variables	106
4.6	Validación temporal adicional.....	108
4.7	Análisis espacial	110
4.7.1	Índice de Moran global.....	110
4.7.2	Detección de hotspots (Getis-Ord Gi*)	111
4.7.3	Interpretación de los resultados espaciales	113
4.8	Análisis de residuales y capacidad predictiva	114
4.8.1	Comparación de valores observados vs predichos.....	114
4.8.2	Análisis de residuales de Random Forest	115
4.8.3	Coefficientes de Lasso (selección de variables).....	115
4.8.4	Serie temporal de predicción (Ayacucho)	116
4.8.5	Mapa de residuales espaciales para el modelo Random Forest	117
4.9	Análisis de riesgo predicho	119
4.10	Análisis de incertidumbre	122
4.10.1	Intervalos de confianza de las predicciones	122

4.10.2	Análisis de sensibilidad de la importancia de variables	123
4.11	Discusión de resultados en el contexto de la literatura regional.....	125
4.11.1	Comparación con estudios en Ecuador	125
4.11.2	Comparación con estudios en Latinoamérica	126
CAPÍTULO 5:.....		128
5	Conclusiones y Recomendaciones.....	128
5.1	Conclusiones generales.....	128
5.1.1	Sobre la integración de fuentes de datos	128
5.1.2	Sobre la caracterización de la enfermedad	128
5.1.3	Sobre la evaluación de la linealidad y selección de modelos.....	129
5.1.4	Sobre los predictores más importantes.....	129
5.1.5	Sobre la validación temporal y robustez interanual	130
5.1.6	Sobre el análisis espacial y residuales	131
5.1.7	Sobre la respuesta a la pregunta de investigación	131
5.2	Implicaciones para la vigilancia epidemiológica	132
5.2.1	Recomendaciones operativas para el Ministerio de Salud Pública	133
5.3	Limitaciones del estudio	135
5.4	Líneas de investigación futura	136
5.5	Reflexión final	137

LISTA DE TABLAS

Tabla 1	Características principales de los modelos de series temporales aplicadas al dengue	23
Tabla 2	Evidencia reciente sobre modelos de series temporales en pronóstico del dengue.....	23
Tabla 3	Limitaciones de los enfoques tradicionales en la predicción del dengue.....	25
Tabla 4	Ventajas del machine learning frente a enfoques tradicionales en predicción epidemiológica.	30
Tabla 5	Aplicaciones del machine learning en la predicción del dengue	32
Tabla 6	Características de XGBoost y su relevancia.....	38
Tabla 7	Justificación de selección de Random Forest y XGBoost.	40
Tabla 8	Comparación breve con otras alternativas.	41
Tabla 9	Síntesis comparativa de estudios con modelos tradicionales en dengue.....	43
Tabla 10	Síntesis comparativa de estudios con machine learning en dengue.	45
Tabla 11	Síntesis comparativa de estudios con análisis espacial del dengue.....	48
Tabla 12	Fuentes de datos Ministerio de Salud Pública (MSP)	53
Tabla 13	Datos: NASA POWER.....	55
Tabla 14	Descripción de variables del Dataset	57
Tabla 15	Definición de variables independientes.....	59
Tabla 16	Resultados de análisis exploratorio.	67
Tabla 17	División de datos - Train & Test.....	69
Tabla 18	Evaluación de modelos y justificación	69
Tabla 19	Evaluación de métricas	70
Tabla 20	Definición de modelos	72
Tabla 21	Técnicas aplicadas para el análisis espacial	73
Tabla 22	Frameworks y Herramientas utilizadas.....	74
Tabla 23	Limitaciones metodológicas	76

Tabla 24 Síntesis del análisis exploratorio.....	87
Tabla 25 Métricas del modelo OLS (baseline).	88
Tabla 26 Coeficientes del modelo OLS (Top 10 por magnitud absoluta).....	88
Tabla 27 Validación cruzada temporal para la selección de α en Ridge.....	90
Tabla 28 Métricas del modelo Ridge.	91
Tabla 29 Coeficientes de Ridge vs OLS (Top 10 por magnitud en Ridge).....	92
Tabla 30 Validación cruzada temporal para la selección de α en Lasso	93
Tabla 31 Métricas del modelo Lasso	94
Tabla 32 Variables seleccionadas por Lasso.	94
Tabla 33 Mejores resultados de la búsqueda de hiperparámetros para ElasticNet.....	96
Tabla 34 Métricas del modelo ElasticNet	96
Tabla 35 Variables seleccionadas por ElasticNet.....	96
Tabla 36 Comparación de modelos baseline.....	97
Tabla 37 Resultados de la prueba de Ramsey RESET	98
Tabla 38 Correlación de Pearson vs Spearman (variables seleccionadas)	99
Tabla 39 Desempeño de los modelos por bloques de variables	102
Tabla 40 Métricas del modelo Random Forest.....	103
Tabla 41 Métricas de XGBoost y comparación con Random Forest.....	104
Tabla 42 Comparativa con modelos baseline.....	104
Tabla 43 Resultados de la prueba de Diebold-Mariano (Random Forest vs modelos baseline)...	105
Tabla 44 Importancia de variables en Random Forest (Top 10)	106
Tabla 45 Resultados de la validación temporal adicional.....	108
Tabla 46 Resultados del estadístico Getis-Ord Gi por parroquia.....	112
Tabla 47 Residuales promedio por parroquia (2025).....	117
Tabla 48 Categorías de riesgo por parroquia	121

Tabla 49 Intervalos de confianza (95%) para parroquias con mayor incidencia predicha (2025).. 122

Tabla 50 Análisis de sensibilidad de la importancia de variables (desviación estándar tras 100 repeticiones)..... 123

LISTA DE FIGURAS

Figura 1 Ciclo de vida de <i>Aedes Aegypti</i>	12
Figura 2 Distribución de la tasa de incidencia (casos por 10,000 habitantes).	62
Figura 3 Incidencia semanal promedio de dengue en Guayaquil (2021-2025).....	63
Figura 4 Distribución de la tasa de incidencia por parroquia (boxplot).	64
Figura 5 Incidencia acumulada por parroquia (2021-2025).....	65
Figura 6 Matriz de correlaciones (Pearson) entre variables climáticas y tasa de incidencia.	66
Figura 7 Distribución de la tasa de incidencia semanal de dengue en Guayaquil (2021-2025).....	82
Figura 8 Incidencia semanal promedio de dengue en Guayaquil (2021-2025).....	83
Figura 9 Distribución de la tasa de incidencia por parroquia (boxplot)	84
Figura 10 Incidencia acumulada por parroquia (2021-2025).....	84
Figura 11 Matriz de correlaciones (Pearson) entre variables climáticas y tasa de incidencia	86
Figura 12 Valores observados vs predichos por OLS (2025).....	89
Figura 13 Diagnóstico de linealidad del modelo OLS	100
Figura 14 Comparativa de métricas (MAE, RMSE, R^2) por modelo.	105
Figura 15 Importancia de variables en Random Forest	107
Figura 16 Mapa de hotspots y coldspots de incidencia de dengue en Guayaquil (Getis-Ord G_i^*)	111
Figura 17 Valores observados vs predichos para (a) OLS, (b) Lasso y (c) Random Forest.	114
Figura 18 Análisis de residuales de Random Forest: (a) residuales vs valores ajustados, (b) Q-Q plot.	115
Figura 19 Coeficientes de Lasso (variables seleccionadas).	116
Figura 20 Serie temporal de incidencia observada vs predicha por Random Forest (Ayacucho, 2025).	117
Figura 21 Mapa de residuales promedio por parroquia (2025)	119

Figura 22	Mapa de zonas de riesgo de dengue por parroquia (2025)	120
------------------	--	-----

CAPÍTULO 1:

1 INTRODUCCIÓN

En regiones tropicales y subtropicales, el dengue es considerado una de las enfermedades transmitidas por vectores de mayor impacto en la salud pública (World Health Organization, 2025).

En el Ecuador, la transmisión vinculada al mosquito *Aedes aegypti* depende de varios factores biológicos, ambientales, climáticos y territoriales que condicionan la aparición y expansión de brotes (Real-Cotto et al., 2017). Guayaquil representa un contexto de especial interés epidemiológico debido a su densidad urbana; las condiciones climáticas de la ciudad la hacen propensa a la proliferación del vector y a la presencia permanente de casos de dengue (Castillo et al., 2022).

Por consiguiente, resulta necesario fortalecer y proponer nuevos mecanismos que permitan predecir de manera ágil zonas de transmisión, integrando datos históricos y factores que inciden directamente en la propagación del vector.

En este contexto, la incorporación de técnicas de machine learning geoespacial aplicadas a datos epidemiológicos y variables ambientales constituye una alternativa metodológica adecuada para mejorar la capacidad predictiva y generar evidencia útil para la atención epidemiológica y la toma de decisiones en salud pública.

1.1 Definición del problema

La gestión de la vigilancia epidemiológica del dengue en la ciudad de Guayaquil enfrenta actualmente tres limitaciones metodológicas que reducen la eficacia de las intervenciones sanitarias:

- **Complejidad ambiental y epidemiológica:** La incidencia del dengue no responde a un comportamiento aleatorio, sino a la interacción de factores climáticos, ambientales y demográficos que favorecen la proliferación del mosquito *Aedes aegypti*. Sin embargo,

estas relaciones suelen abordarse de forma descriptiva y no mediante modelos predictivos integrados

- **Limitación de los enfoques tradicionales:** La vigilancia epidemiológica se apoya principalmente en análisis descriptivos y retrospectivos de los casos reportados. Aunque estos métodos son útiles para caracterizar el comportamiento histórico de la enfermedad, presentan limitaciones para anticipar la aparición de zonas de riesgo y variaciones futuras de la incidencia.
- **Desarticulación de fuentes de datos:** Existe una limitada integración entre los registros epidemiológicos y las variables ambientales de alta resolución, como temperatura, precipitación y humedad, disponibles en plataformas como NASA EarthData. Esta desarticulación dificulta modelar relaciones no lineales y espacio temporales relevantes para la predicción del dengue.

Pregunta de investigación

¿En qué medida la integración de variables ambientales provenientes de NASA EarthData y registros epidemiológicos históricos, mediante modelos de machine learning geoespacial, mejora la predicción espacial de la incidencia del dengue en Guayaquil, Ecuador, en comparación con modelos baseline tradicionales?

1.2 Justificación

1.2.1 Justificación académica.

La investigación aporta al campo de la epidemiología espacial y de la ciencia de datos aplicada a la salud pública mediante la integración de datos epidemiológicos históricos y variables ambientales en un enfoque de predicción espacial del dengue. Aunque la literatura reciente ha desarrollado modelos estadísticos, series temporales y técnicas de aprendizaje automático para el análisis de esta enfermedad, aún existen vacíos en la comparación sistemática entre modelos tradicionales y enfoques de machine learning geoespacial, especialmente en contextos urbanos del Ecuador. (Zhao et al., 2020). Por ello, el estudio busca aportar evidencia metodológica aplicada al caso de Guayaquil.

1.2.2 Justificación práctica.

El desarrollo de un modelo de predicción espacial puede fortalecer la capacidad de análisis de la vigilancia epidemiológica al facilitar la identificación de zonas de riesgo y patrones territoriales asociados a la incidencia del dengue. Esto permitiría orientar con mayor precisión las acciones preventivas y la asignación de recursos sanitarios.

1.2.3 Justificación social.

El dengue representa un problema de salud pública que afecta especialmente a sectores urbanos vulnerables expuestos a condiciones ambientales favorables para la proliferación del vector. En este sentido, mejorar la capacidad de anticipación del riesgo puede contribuir indirectamente a la prevención de brotes y a la reducción del impacto social de la enfermedad en la población de Guayaquil.

1.2.4 Justificación tecnológica.

El uso de algoritmos como Random Forest y XGBoost se justifica por su pertinencia para modelar relaciones no lineales, interacciones complejas y datos heterogéneos en problemas de predicción (Breiman, 2001). Su aplicación permitirá evaluar si estos enfoques ofrecen mejoras frente a modelos baseline tradicionales. Además, el proyecto es técnicamente viable gracias a la articulación de tres factores: (i) los datos epidemiológicos, obtenidos mediante una solicitud formal al Ministerio de Salud Pública del Ecuador (MSP) y autorizados específicamente para esta investigación, (ii) los datos climáticos, accesibles públicamente a través de la plataforma NASA POWER, y (iii) el uso exclusivo de herramientas de software libre (Python, pandas, scikit-learn, geopandas, QGIS) para el análisis geoespacial y el modelado predictivo, lo que garantiza la transparencia, reproducibilidad y bajo costo del estudio.

1.3 Alcance del proyecto

La presente investigación se orienta al desarrollo y evaluación de un modelo de predicción espacial del dengue en la ciudad de Guayaquil, Ecuador, a partir de la integración de datos

epidemiológicos históricos y variables ambientales provenientes de fuentes secundarias. El estudio se plantea como una prueba de concepto analítica con enfoque aplicado, cuyo propósito es analizar la capacidad de técnicas de machine learning geoespacial para mejorar la predicción de la incidencia del dengue en un contexto urbano de alta relevancia epidemiológica.

Dentro del alcance del proyecto se contempla, en primer lugar, la recopilación, integración y preprocesamiento de un dataset unificado que combine registros semanales de incidencia del dengue del Ministerio de Salud Pública (MSP) con variables ambientales (temperatura, precipitación y humedad) provenientes de NASA POWER, y datos poblacionales del INEC (INEC, 2022). El dataset resultante cubre 15 parroquias urbanas de Guayaquil para el periodo 2021-2025, con un total de 3,900 observaciones.

En segundo lugar, se realiza un análisis exploratorio de datos (EDA) para caracterizar la distribución de la variable dependiente, identificar patrones de estacionalidad y heterogeneidad espacial, y evaluar las correlaciones entre variables predictoras. Este análisis revela una alta proporción de ceros (68.74%), una asimetría extrema y relaciones no lineales, lo que condiciona la selección de los modelos predictivos.

En tercer lugar, se aplica la prueba de Ramsey RESET sobre un modelo OLS baseline para verificar la linealidad de las relaciones (Ramsey, 1969). Al rechazar la hipótesis nula de especificación lineal ($p < 0.001$), se justifica la exploración de dos familias de modelos: (a) modelos lineales regularizados (Ridge, Lasso, ElasticNet) como baseline comparativo, y (b) modelos basados en árboles (Random Forest y XGBoost), capaces de capturar relaciones no lineales e interacciones complejas.

En cuarto lugar, se evalúa el aporte incremental de las variables climáticas mediante la comparación de cuatro versiones del modelo Random Forest (solo inercia, inercia + clima contemporáneo, inercia + clima con rezagos, y modelo completo). Adicionalmente, se realiza una

validación temporal con dos splits (2021-2023→2024 y 2021-2024→2025) para evaluar la estabilidad interanual del modelo.

En quinto lugar, se aplican técnicas de análisis espacial, incluyendo el índice de Moran y el estadístico Getis-Ord G_i^* , para evaluar la autocorrelación espacial de la incidencia del dengue e identificar posibles hotspots o zonas de riesgo. Los resultados se complementan con un análisis de residuales geoespaciales del modelo final.

Finalmente, se genera una representación geoespacial de los resultados (mapas de incidencia, hotspots y residuales) como recurso de apoyo para la interpretación técnica de los hallazgos y para la toma de decisiones en salud pública.

En cuanto a su delimitación, el estudio se circunscribe geográficamente a la ciudad de Guayaquil, con énfasis en su zona urbana. Se trabajará con una serie histórica de cinco años, desde el año 2021 al 2025. Esta selección histórica se la toma para evitar el ruido introducido por la COVID-19 en la notificación de casos y en los patrones de movilidad.

Temáticamente, la investigación se concentra en la predicción espacial de la incidencia del dengue mediante la integración de variables epidemiológicas y ambientales (Cabrera et al., 2022).

Desde el punto de vista tecnológico, el proyecto se apoya en:

- **Machine learning:** algoritmos de ensamble como Random Forest y XGBoost, junto con modelos lineales regularizados (OLS, Ridge, Lasso, ElasticNet), implementados con las librerías scikit-learn y xgboost en Python.
- **Análisis geoespacial:** técnicas de autocorrelación espacial (índice de Moran global) y detección de hotspots (Getis-Ord G_i^*), implementadas con la librería esda (PySAL).
- **Sistemas de Información Geográfica (SIG):** manejo de archivos GeoJSON y shapefiles, manipulación de geometrías, proyecciones cartográficas y generación de mapas temáticos, utilizando las librerías geopandas, shapely y matplotlib en Python.

- **Series temporales:** construcción de rezagos temporales (lags de 1 a 4 semanas) para variables climáticas y epidemiológicas, y validación del modelo mediante splits temporales que preservan el orden cronológico de los datos.

El estudio no incluye la recolección de datos primarios mediante encuestas, entrevistas o trabajo de campo, ni contempla el diagnóstico clínico de pacientes. Tampoco forma parte del alcance la implementación operativa en tiempo real dentro de la infraestructura institucional del Ministerio de Salud Pública, el desarrollo de una plataforma productiva de uso oficial o la construcción de hardware especializado. En consecuencia, los resultados del trabajo deben entenderse como una propuesta metodológica y analítica orientada al apoyo de la vigilancia epidemiológica y la toma de decisiones.

1.4 Objetivos

1.4.1 Objetivo General

Desarrollar y evaluar un modelo de predicción espacial del dengue aplicando técnicas de machine learning geoespacial, utilizando datos epidemiológicos históricos y variables ambientales, con el fin de identificar patrones espaciales y mejorar la capacidad predictiva de la incidencia del dengue en la ciudad de Guayaquil.

1.4.2 Objetivos específicos

- **Recopilar y caracterizar** los registros epidemiológicos del dengue (MSP) y las variables meteorológicas de la NASA POWER API para el periodo 2021-2025 para la ciudad de Guayaquil.
- **Integrar y preprocesar** un dataset geoespacial unificado que incorpore variables epidemiológicas históricas, variables ambientales y componentes espaciales y temporales para el modelado predictivo.
- **Desarrollar y entrenar** modelos de predicción basados en Random Forest y XGBoost para estimar la incidencia del dengue en la ciudad de Guayaquil.
- **Evaluar el desempeño de los modelos** utilizando métricas de error como el Error Absoluto Medio (MAE) y la Raíz del Error Cuadrático Medio (RMSE).

- **Visualizar e interpretar** los resultados mediante mapas y recursos geoespaciales de apoyo al análisis epidemiológico y a la toma de decisiones.
- **Definir e implementar** un esquema de comparación con modelos baseline tradicionales para evaluar la mejora predictiva de los modelos propuestos.
- **Aplicar técnicas de análisis espacial** para identificar patrones de distribución y zonas de riesgo asociadas a la incidencia del dengue en Guayaquil.

1.5 Hipótesis y variables

La aplicación de modelos de machine learning geoespacial, mediante la integración de datos epidemiológicos históricos y variables ambientales de alta resolución, permitirá captar patrones espaciotemporales no lineales y mejorará la precisión en la predicción de la incidencia del dengue en la ciudad de Guayaquil, en comparación con modelos tradicionales de referencia.

Para contrastar esta hipótesis, se evaluará el desempeño de los modelos mediante métricas de error predictivo, principalmente el Mean Absolute Error (MAE) y el Root Mean Squared Error (RMSE) (Willmott & Matsuura, 2005; Chai & Draxler, 2014). Se espera que los modelos propuestos reduzcan el error de predicción frente a los enfoques tradicionales utilizados como línea base.

- **Mean Absolute Error (MAE):**

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

- **Root Mean Squared Error (RMSE):**

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

Identificación de Variables:

a) Variable dependiente (Y)

Tasa de incidencia semanal de dengue (TASA_INCIDENCIA): definida como el número de casos confirmados y probables de dengue (DSSA y DCSA) registrados en una parroquia durante una semana epidemiológica, expresado como casos por 10,000 habitantes. La tasa se calcula mediante la siguiente fórmula:

$$TI_{i,t} = \frac{Casos_{i,t}}{Poblacion_i} \times 10,000$$

Donde:

- $TI_{i,t}$: tasa de incidencia en la parroquia i durante la semana t
- $Casos_{i,t}$: suma de casos DSSA y DCSA en la parroquia i durante la semana t
- $Poblacion_i$: población de la parroquia i según el censo INEC 2022

Esta variable fue seleccionada porque normaliza los conteos absolutos por el tamaño poblacional, permitiendo comparaciones justas entre parroquias de distinta magnitud demográfica (WHO, 2022). El denominador de 10,000 habitantes (en lugar de 100,000) se eligió porque los conteos semanales por parroquia son pequeños, facilitando la interpretación de las tasas.

b) Variables independientes (X)

- **Variables epidemiológicas rezagadas:** conteos históricos de casos de dengue con retardos temporales de 1 semana ($TASA_INCIDENCIA_lag_1$). Se utilizó únicamente el rezago de 1 semana, ya que es el que captura la inercia epidemiológica inmediata (Johansson et al., 2016).
- **Variables ambientales actuales y rezagadas:**
 - Temperatura mínima y máxima a 2 metros (°C): $T2M_MIN$, $T2M_MAX$
 - Precipitación acumulada corregida (mm): $PRECTOTCORR$

- Humedad relativa a 2 metros (%): *RH2M*

Estas variables fueron extraídas de la plataforma NASA POWER (NASA POWER, 2025) para los centroides de cada parroquia. Para capturar los efectos retardados de las condiciones ambientales sobre la transmisión del dengue, se generaron rezagos temporales de 1 a 4 semanas para cada variable climática (*T2M_MIN_lag_1* a *lag_4*, etc.), siguiendo la literatura que documenta desfases entre las condiciones climáticas y la aparición de casos (Tuladhar, 2019; Lambrechts et al., 2011).

- **Variables espaciales y temporales complementarias:**

- Coordenadas geográficas (latitud y longitud) de los centroides parroquiales (*LAT, LON*)
- Año epidemiológico (*EPI_YEAR*) y semana epidemiológica (*EPI_WEEK*) como controles de estacionalidad y tendencia.

Criterio de contraste de la hipótesis

La hipótesis se considerará respaldada si se cumple al menos una de las siguientes condiciones:

1. El modelo de machine learning geoespacial seleccionado (Random Forest) presenta un MAE menor que el modelo baseline OLS en el conjunto de prueba (2025) en la predicción de la tasa de incidencia semanal de dengue, indicando una reducción del error absoluto promedio.
2. El modelo de machine learning geoespacial seleccionado presenta un RMSE menor que el modelo baseline OLS en el conjunto de prueba (2025), indicando una mejor capacidad para capturar brotes epidémicos (errores grandes).
3. El modelo de machine learning geoespacial seleccionado presenta un R^2 mayor que el modelo baseline OLS en el conjunto de prueba (2025), indicando una mayor proporción de varianza explicada.

Adicionalmente, se evaluará el aporte incremental de las variables climáticas mediante la comparación de cuatro versiones del modelo Random Forest (M1: solo inercia, M2: inercia + clima contemporáneo, M3: inercia + clima con rezagos, M4: modelo completo). Si M2, M3 o M4 superan a M1 en al menos dos de las tres métricas (MAE, RMSE, R^2), se considerará que las variables climáticas aportan información relevante más allá de la inercia epidemiológica.

La hipótesis se considerará no respaldada si el modelo Random Forest no supera a OLS en al menos dos de las tres métricas, o si el aporte incremental del clima no es consistente a través de las métricas evaluadas.

CAPÍTULO 2

2 Fundamentos Epidemiológicos y Contexto de la Enfermedad

2.1 Definición y Agente Causal

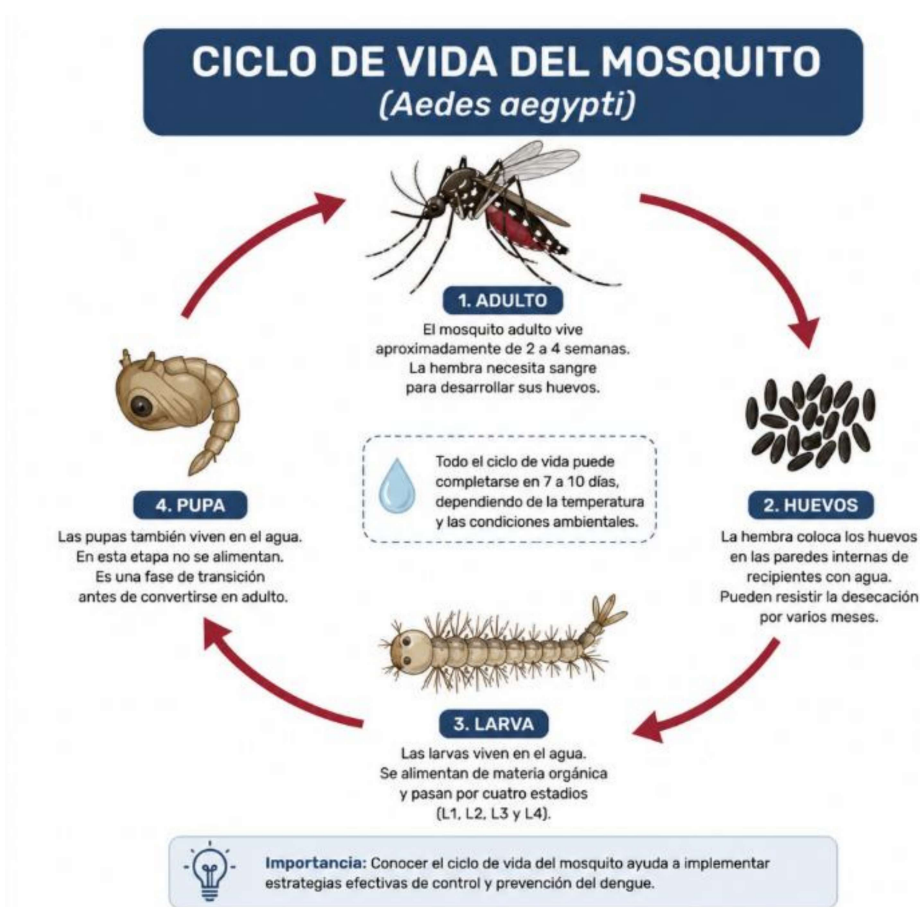
El dengue es una enfermedad febril infecciosa, de etiología viral y carácter sistémico, transmitida por la picadura de mosquitos del género *Aedes*. El agente causal es el virus del dengue (DENV), un arbovirus perteneciente a la familia *Flaviviridae*, género *Flavivirus*. El complejo DENV se compone de cuatro serotipos inmunológicamente distantes (DENV-1, DENV-2, DENV-3 y DENV-4). La infección por un serotipo proporciona inmunidad homóloga de por vida, pero solo una protección temporal y parcial contra los otros tres, lo que incrementa el riesgo de formas graves de la enfermedad (dengue grave) ante infecciones secundarias por serotipos heterólogos (World Health Organization, 2025).

2.2 El Vector: *Aedes Aegypti* y su Ecología Urbana

El principal vector en el hemisferio occidental, y específicamente en la región costa del Ecuador, es el mosquito *Aedes aegypti*. Este insecto se ha especializado evolutivamente en entornos antrópicos; es un mosquito predominantemente doméstico y peridoméstico. Su ciclo de vida —huevo, larva, pupa y adulto— está estrechamente ligado a la disponibilidad de agua estancada en recipientes artificiales (Ministerio de Salud Pública del Ecuador, 2026).

La relevancia de este vector para el modelado geoespacial reside en su sensibilidad térmica: temperaturas óptimas entre 25 °C y 30 °C aceleran su metabolismo y reducen el periodo de incubación extrínseca (PIE), mientras que humedades relativas superiores al 60 % incrementan su tasa de supervivencia y frecuencia de alimentación (picadura), favoreciendo la transmisión del virus del dengue y su distribución espacial en zonas urbanas (Dom et al., 2013; Tuladhar et al., 2019; Tuan et al., 2024).

Figura 1
Ciclo de vida de *Aedes Aegypti*



Fuente: Elaboración propia con base en World Health Organization (2025) y Ministerio de Salud Pública del Ecuador (2026).

2.3 Formas de Transmisión y Dinámica del Brote

La transmisión ocurre a través de un ciclo humano-vector-humano. Cuando un mosquito hembra pica a una persona virémica, el virus ingresa al intestino del insecto y, tras un periodo de 8 a 12 días (dependiente del clima), migra a las glándulas salivales. Una vez infectado, el mosquito transmite el virus durante el resto de su vida funcional. Es importante destacar que el radio de vuelo del *Aedes aegypti* es limitado (aproximadamente 50 a 100 metros), lo que justifica el análisis de la enfermedad a escalas micro-espaciales o parroquiales, ya que la dispersión a larga distancia suele ser producto de la movilidad humana y no del desplazamiento del vector.

2.4 Impacto en la Salud Pública y Relevancia Urbana

El dengue representa una carga económica y social masiva. El impacto se mide no solo en la mortalidad, sino en los años de vida ajustados por discapacidad (AVAD) y el costo por hospitalización, que puede saturar los sistemas de salud pública durante picos epidémicos.

En contextos tropicales y urbanos como Guayaquil, la enfermedad adquiere una dimensión de "riesgo socio-espacial". La rápida urbanización no planificada, la acumulación de desechos sólidos y las deficiencias en el suministro de agua potable obligan a la población al almacenamiento de agua en contenedores, creando criaderos permanentes. Esto convierte a las ciudades en "islas de calor" y reservorios virales donde el clima y la infraestructura urbana interactúan para perpetuar la endemidad.

2.5 Situación Epidemiológica en el Ecuador y Ciudades Comparables

Ecuador es un país hiperendémico donde circulan simultáneamente los cuatro serotipos del virus. Históricamente, la región litoral concentra más del 70% de los casos nacionales, siendo la provincia del Guayas el epicentro recurrente. Guayaquil, debido a su densidad poblacional y condiciones hidrometeorológicas, presenta una dinámica comparable a metrópolis como Medellín (Colombia) o Fortaleza (Brasil), donde la transmisión no se detiene durante la época seca, sino que se mantiene de forma basal para explotar exponencialmente al inicio de la temporada de lluvias.

Durante el periodo 2021-2024, el país experimentó fluctuaciones significativas relacionadas con fenómenos climáticos como El Niño-Oscilación del Sur (ENOS), lo que subraya la necesidad de modelos predictivos que, como el propuesto en esta investigación, integren variables climáticas de alta resolución para anticipar estos eventos extremos.

2.6 Factores Ambientales, Climáticos y Territoriales Asociados al Dengue

La dinámica de transmisión del dengue en entornos urbanos no precisa a procesos aleatorios, sino que responde a una compleja red de determinantes ambientales y territoriales que condicionan la aptitud del hábitat para el vector *Aedes Aegypti*.

2.6.1 Determinantes Climáticos y su Influencia Biológica.

La literatura científica coincide en que el clima es el principal impulsor de la variabilidad temporal de la enfermedad. En el contexto de ciudades tropicales de llanura, tres variables dictan el ritmo de los brotes:

2.6.1.1 Temperatura. Actúa como el regulador metabólico del mosquito. Temperaturas óptimas (24°C - 30°C) reducen el Periodo de Incubación Extrínseca (PIE), permitiendo que el virus se replique más rápido en el vector. Estudios en la región litoral demuestran que incrementos en la temperatura mínima son predictores críticos de la expansión de la frontera epidemiológica.

2.6.1.2 Precipitación. Proporciona los sitios de cría necesarios para la fase acuática del vector. Sin embargo, su relación es no lineal: lluvias moderadas y persistentes crean criaderos, mientras que precipitaciones extremas pueden generar un efecto de "lavado" (*flushing*), eliminando las larvas de los recipientes.

2.6.1.3 Humedad Relativa. Es el factor determinante para la longevidad del mosquito adulto. Humedades superiores al 70% evitan la desecación del insecto, extendiendo su vida funcional y, por ende, aumentando la probabilidad de que realice múltiples picaduras infectivas.

2.6.2 Factores Territoriales y de Infraestructura Urbana

Mientras el clima explica el "cuándo", el territorio explica el "dónde". La configuración urbana de Guayaquil presenta particularidades que potencian la endemidad:

2.6.2.1 Densidad Poblacional. En parroquias de alta densidad como Febres Cordero o Ximena, la proximidad entre huéspedes susceptibles y vectores infectados incrementa la tasa de contacto. El *Aedes aegypti*, al tener un rango de vuelo corto, se beneficia de entornos donde el desplazamiento humano actúa como el principal vehículo de dispersión viral inter-parroquial.

2.6.2.2 Condiciones de Saneamiento y Ambiente. Las deficiencias en el suministro continuo de agua potable obligan al almacenamiento doméstico en tanques y reservorios, los cuales funcionan como criaderos permanentes, independientemente de la temporada de lluvias. Asimismo, la gestión inadecuada de desechos sólidos (neumáticos, envases) proporciona micro-hábitats ideales para la oviposición.

2.6.3 Interacción Dinámica y Efectos Rezagados (Lags)

Uno de los pilares de esta investigación es el reconocimiento de que la incidencia de dengue observada en un período determinado es el resultado de condiciones ambientales ocurridas semanas antes. Diversos estudios han demostrado que variables climáticas como la temperatura, la precipitación y la humedad influyen en la dinámica poblacional del vector y en la transmisión del virus mediante efectos rezagados, generando desfases temporales entre las condiciones ambientales y la aparición de casos de dengue (Tuladhar, 2019).

Semana 1-2: Precipitación y temperatura activan la eclosión y el desarrollo larvario.

Semana 3: Emergencia de adultos y periodo de incubación extrínseca del virus.

Semana 4: Aparición de síntomas en la población humana y reporte al sistema de salud.

Esta jerarquía temporal justifica técnicamente la inclusión de rezagos de hasta 4 semanas en el modelo predictivo, permitiendo que los algoritmos de *Machine Learning* capturen la "memoria" del ecosistema urbano antes de que se manifieste el brote.

2.6.3.1 Estacionalidad en Contextos Urbanos Tropicales. En Guayaquil, la interacción entre el clima y el territorio produce una estacionalidad bimodal. Durante la estación lluviosa, el riesgo es generalizado debido a los criaderos naturales. No obstante, durante la estación seca, la persistencia de la enfermedad se vincula a factores territoriales (saneamiento y almacenamiento de agua), lo que convierte a ciertas parroquias en reservorios basales que mantienen el virus circulando hasta el siguiente ciclo pluvial.

2.6.4 Epidemiología Espacial y Análisis Geoespacial en Salud

La integración del componente geográfico en el estudio de las enfermedades infecciosas ha permitido transitar de una visión puramente clínica a una comprensión sistémica del riesgo. La epidemiología espacial se define como la descripción y el análisis de la variación geográfica de las enfermedades con respecto a factores de riesgo demográficos, ambientales y socioeconómicos (Elliott & Wartenberg, 2004).

2.6.4.1 Concepto y alcance de la epidemiología espacial. En el contexto del dengue, la epidemiología espacial no se limita a la localización de casos en un mapa, sino que busca entender los procesos que generan dicha distribución. Se fundamenta en tres pilares:

- **Maapeo de enfermedades:** Identificación de patrones de propagación.
- **Estudios de correlación geográfica:** Vinculación de la incidencia con variables del entorno (clima, densidad).
- **Detección de clústeres:** Localización de áreas con un riesgo significativamente mayor al promedio.

2.6.5 Autocorrelación Espacial: La Primera Ley de Tobler

El sustento teórico para incluir coordenadas (LAT, LON) en el modelo de Machine Learning es la Primera Ley de la Geografía de Tobler, la cual establece que "todas las cosas están relacionadas entre sí, pero las cosas cercanas están más relacionadas que las distantes". En epidemiología, esto se traduce en autocorrelación espacial: la probabilidad de que

una parroquia presente un brote aumenta si sus parroquias vecinas ya presentan una alta incidencia. El modelo de *Random Forest* captura esta dependencia espacial al analizar cómo la ubicación geográfica interactúa con los rezagos climáticos.

2.6.6 Clústeres, Hotspots y Zonas de Riesgo

Un concepto crítico es la identificación de Hotspots (puntos calientes), definidos como áreas geográficas que presentan una concentración inusual de eventos en comparación con su entorno. Mientras que el clima dicta la idoneidad biológica, la estructura urbana de Guayaquil genera clústeres de riesgo persistente.

- **Clústeres espaciales:** Grupos de parroquias con altas tasas de incidencia (ej. el cinturón del suburbio).
- **Análisis de Residuales:** La utilidad del análisis espacial en salud pública radica en identificar el **riesgo residual**. Si el modelo predice un riesgo bajo basado en el clima, pero la incidencia observada es alta, se identifica un hotspot de vulnerabilidad estructural que requiere intervención inmediata.

2.6.7 El Rol de los SIG en la Vigilancia Epidemiológica Moderna

Los Sistemas de Información Geográfica (SIG) han evolucionado de herramientas de visualización estática a plataformas de soporte para la toma de decisiones. En la vigilancia del dengue, el uso de SIG permite:

- **Integración de datos heterogéneos:** Fusión de microdatos del MSP, sensores remotos de la NASA y censos poblacionales.
- **Normalización Espacial:** Transformar conteos absolutos en tasas por habitante, permitiendo comparaciones justas entre parroquias de distinto tamaño poblacional.
- **Modelado Predictivo:** La transición hacia la Inteligencia Artificial Geoespacial (Geo-AI) permite que el SIG no solo muestre "donde está" la enfermedad, sino que anticipe "donde estará" en función de la dinámica ambiental detectada.

2.6.8 Utilidad en Salud Pública y Justificación del Modelo

La relevancia de este componente geoespacial reside en la capacidad de optimizar la gestión de recursos. En lugar de realizar fumigaciones masivas e ineficientes en todo el cantón, el análisis espacial permite dirigir las brigadas de control vectorial hacia las parroquias que el modelo identifica como zonas con mayor riesgo predictor. De este modo, la metodología propuesta se justifica como una herramienta de precisión que combina la biología del vector con la inteligencia territorial.

2.7 Fuentes de Datos y Variables de Estudio

La construcción del dataset multivariante para este modelo predictivo se fundamenta en la integración de tres pilares de información oficial, asegurando la validez y trazabilidad de los datos utilizados.

2.7.1 Datos Epidemiológicos (Variable Target)

- **Fuente:** Ministerio de Salud Pública del Ecuador (MSP). Los datos fueron obtenidos mediante solicitud formal a la Dirección Nacional de Vigilancia Epidemiológica, a partir de los registros consolidados del Sistema Integrado de Vigilancia Epidemiológica (SIVE-ALERTA) (Ministerio de Salud Pública, 2025).
- **Descripción:** Registros semanales de casos confirmados y probables de dengue, incluyendo dengue con signos de alarma (DCSA) y dengue sin signos de alarma (DSSA). Los datos fueron depurados y procesados para el cantón Guayaquil, considerando la unidad de residencia del paciente con el fin de garantizar la precisión geoespacial del análisis.
- **Volumen:** 260 semanas epidemiológicas (Enero 2021 – Diciembre 2025).
- **Variable dependiente:** Tasa de incidencia semanal (casos por 10,000 habitantes), calculada a partir del total de casos semanales y la población de cada parroquia. La desagregación por sexo y grupos etarios se utilizó en el análisis descriptivo para caracterizar la población afectada, pero no como predictores en el modelo final.

2.7.2 Datos Climáticos y Ambientales (Predictores)

- **Fuente:** Proyecto *Prediction of Worldwide Energy Resources* (POWER) de la Administración Nacional de Aeronáutica y el Espacio (NASA) (NASA POWER, 2025).
- **Descripción:** Datos de reanálisis meteorológico obtenidos mediante sensores remotos y modelos de asimilación de datos globales. Se extrajeron valores específicos para los centroides geográficos de las 15 parroquias urbanas de Guayaquil y se agregaron a nivel semanal.
- **Volumen:** 27,390 registros diarios (15 parroquias × 1,826 días del periodo 2021-2025). Estos datos fueron agregados a nivel semanal, resultando en 3,900 observaciones (15 parroquias × 260 semanas epidemiológicas) utilizadas en el modelo predictivo.
- **Variables:** Temperatura máxima y mínima a 2 metros (T2M_MAX, T2M_MIN), precipitación acumulada corregida (PRECTOTCORR) y humedad relativa (RH2M). Adicionalmente, se generaron rezagos temporales de 1 a 4 semanas para cada variable climática, con el fin de capturar los efectos retardados de las condiciones ambientales sobre la transmisión del dengue.

2.7.3 Datos Demográficos y Territoriales (Normalización)

- **Fuente:** Instituto Nacional de Estadística y Censos (INEC), VIII Censo de Población y VII de Vivienda (Instituto Nacional de Estadística y Censos, 2022).
- **Descripción:** Cifras oficiales de población residente por parroquia urbana para el cálculo de la variable dependiente (Tasa de Incidencia).
- **Volumen:** Datos censales definitivos de las 15 parroquias urbanas de Guayaquil.
- **Variables:** Población total por parroquia, densidad poblacional y coordenadas geográficas (Latitud/Longitud).

2.8 Modelos tradicionales para el análisis y predicción del dengue

2.8.1 Vigilancia descriptiva y análisis retrospectivo

La vigilancia descriptiva y el análisis retrospectivo constituyen la base tradicional del monitoreo epidemiológico del dengue en los sistemas de salud pública. Este enfoque se sustenta en la recolección sistemática, consolidación y análisis periódico de información sobre los casos notificados, con el propósito de describir la evolución temporal y territorial de la enfermedad, identificar incrementos inusuales y orientar acciones de control sanitario. La vigilancia en salud pública se entiende como un proceso continuo de observación, análisis e interpretación de datos relevantes para la toma de decisiones en prevención y control de enfermedades (Quintanilla).

En el caso del dengue, la vigilancia descriptiva suele apoyarse en indicadores como el número de casos por semana epidemiológica, las tasas de incidencia, la distribución geográfica de los casos y la comparación con periodos anteriores. Estos elementos permiten construir series históricas, reconocer patrones de estacionalidad y monitorear territorios con mayor carga de enfermedad. En Guayaquil, este tipo de seguimiento se refleja en las gacetas vectoriales del Ministerio de Salud Pública, donde se reportan periódicamente los casos de enfermedades transmitidas por vectores según semana epidemiológica y distribución territorial (Ministerio de Salud Pública del Ecuador, 2023).

Este enfoque resulta fundamental para la vigilancia rutinaria, ya que proporciona una base organizada para conocer el comportamiento observado de la enfermedad y facilitar respuestas institucionales (Quintanilla). Sin embargo, su alcance es principalmente descriptivo y retrospectivo, pues se orienta a registrar y analizar lo ocurrido, más que a anticipar de manera robusta escenarios futuros de riesgo. Por ello, aunque la vigilancia sistemática y la notificación periódica son componentes esenciales del control epidemiológico del dengue, no constituyen por sí solas un modelo predictivo capaz de integrar relaciones complejas entre variables ambientales, temporales y territoriales (Perra, 2015).

2.8.2 Modelos de regresión para datos de conteo

Dentro de los enfoques tradicionales para el análisis epidemiológico del dengue, los modelos de regresión para datos de conteo ocupan un lugar central, debido a que la variable de interés suele expresarse como número de casos registrados en una unidad territorial y temporal determinada. En este contexto, la regresión de Poisson ha sido ampliamente utilizada para modelar la incidencia del dengue en función de variables explicativas como temperatura, precipitación, humedad, densidad poblacional u otros factores ambientales y demográficos. Su principal utilidad radica en que permite relacionar formalmente la frecuencia observada de casos con un conjunto de predictores, bajo el supuesto de que la media y la varianza del conteo son equivalentes (Sajib, 2024).

Sin embargo, en estudios epidemiológicos del dengue es frecuente que los datos presenten sobredispersión, es decir, que la variabilidad observada sea mayor que la esperada bajo un modelo de Poisson. Cuando esto ocurre, la regresión binomial negativa suele considerarse una alternativa más adecuada, ya que incorpora un parámetro adicional para modelar dicha variabilidad y ofrece un ajuste más flexible para datos de conteo con alta dispersión. Por esta razón, la literatura reciente reporta el uso recurrente de la regresión binomial negativa en estudios que buscan analizar asociaciones entre la incidencia del dengue y factores climáticos, ambientales o socioeconómicos (Tuladhar, 2019).

Estos modelos presentan varias ventajas dentro de la epidemiología aplicada. En primer lugar, ofrecen una interpretación estadística clara, ya que permiten estimar el efecto de variables explicativas sobre la ocurrencia de casos. En segundo lugar, son metodológicamente consistentes con la naturaleza discreta de los datos epidemiológicos. Finalmente, constituyen un referente clásico y ampliamente aceptado en la literatura para el análisis de enfermedades transmisibles, por lo que resultan adecuados como modelos baseline de comparación en estudios que incorporan enfoques más avanzados de predicción (Dixon, 2022).

No obstante, estos enfoques también presentan limitaciones. Aunque permiten modelar asociaciones entre predictores y conteos de casos, su capacidad para capturar relaciones altamente no lineales, interacciones complejas entre múltiples variables y patrones espaciotemporales heterogéneos puede resultar restringida. En el caso del dengue, donde la incidencia depende de la interacción entre factores climáticos, históricos y territoriales, estas limitaciones justifican la exploración de modelos más flexibles que complementen o superen el desempeño de los enfoques tradicionales de regresión (Abade, 2024).

2.8.3 Modelos de series temporales

Los modelos de series temporales constituyen uno de los enfoques tradicionales más utilizados para el análisis y pronóstico del dengue, debido a que la incidencia de esta enfermedad suele presentar dependencia temporal y patrones estacionales asociados a factores climáticos y ambientales. Estos modelos emplean los valores históricos observados de la serie para identificar regularidades en el tiempo y generar estimaciones futuras, por lo que han sido ampliamente aplicados en estudios epidemiológicos orientados al pronóstico de casos (Leung, 2023).

Entre los modelos más utilizados se encuentra el **ARIMA** (*AutoRegressive Integrated Moving Average*), que permite representar la estructura temporal de una serie a partir de componentes autorregresivos, diferenciación y medias móviles. Cuando la serie presenta un patrón estacional marcado, como suele ocurrir en el dengue, se emplea el modelo **SARIMA** (*Seasonal AutoRegressive Integrated Moving Average*), que incorpora componentes estacionales para mejorar la representación de comportamientos periódicos (Hasan et al., 2024; Naher et al., 2022).

La principal ventaja de estos modelos radica en su claridad estadística, su relativa facilidad de implementación y su utilidad para describir y pronosticar series epidemiológicas a corto plazo. Por ello, constituyen una referencia metodológica válida para ser empleada como baseline en estudios comparativos. Sin embargo, su capacidad para capturar relaciones no lineales complejas,

integrar múltiples variables heterogéneas y representar simultáneamente patrones temporales y espaciales es más limitada que la de enfoques de machine learning. En consecuencia, aunque siguen siendo modelos importantes en la literatura del dengue, su uso en esta investigación se plantea principalmente como punto de comparación frente a modelos más flexibles como Random Forest y XGBoost (Abade, 2024).

Tabla 1
Características principales de los modelos de series temporales aplicadas al dengue

Modelo	Descripción	Utilidad en dengue	Ventajas	Limitaciones
ARIMA	Modelo autorregresivo integrado de medias móviles que utiliza la serie histórica para pronosticar valores futuros.	Pronóstico de incidencia semanal o mensual.	Interpretación clara, tradición estadística, útil como baseline.	Menor capacidad para representar relaciones no lineales complejas.
	Extensión estacional de ARIMA que incorpora componentes periódicos.	Pronóstico del dengue cuando existe estacionalidad.	Mejor ajuste en series con comportamiento cíclico.	Depende de estabilidad temporal y menor flexibilidad frente a patrones complejos.

Tabla 2
Evidencia reciente sobre modelos de series temporales en pronóstico del dengue

Estudio	Contexto	Modelo	Hallazgo principal
Naher et al. (2022)	Bangladesh	ARIMA	Mostró utilidad de ARIMA para pronosticar incidencia de dengue con datos históricos.

Hasan et al. (2024)	Bangladesh	ARIMA y SARIMA	Confirmó utilidad de modelos temporales con componente estacional.
Chen et al. (2025)	Brasil	ARIMA y SARIMAX	Reportó buen desempeño de modelos estadísticos como referencia comparativa.

2.8.4 Ventajas de los enfoques tradicionales

Los enfoques tradicionales aplicados al análisis y predicción del dengue presentan ventajas que explican su permanencia en la epidemiología y la salud pública. En primer lugar, ofrecen una interpretación estadística clara, lo que facilita comprender la relación entre la incidencia del dengue y las variables analizadas, así como identificar tendencias y comportamientos recurrentes en los datos. En segundo lugar, se caracterizan por su facilidad de implementación, ya que requieren menor complejidad metodológica y computacional en comparación con modelos más avanzados (Dixon, 2022).

Asimismo, estos enfoques mantienen una coherencia adecuada con los datos epidemiológicos agregados, como los registros de casos por semana epidemiológica, tasas de incidencia o series históricas consolidadas por territorio. Esta característica los hace especialmente útiles en contextos de vigilancia rutinaria, donde la información disponible suele estar estructurada en niveles agregados y temporales. Finalmente, su uso resulta valioso como línea base de comparación, pues permiten establecer un punto de referencia metodológico frente al cual evaluar si modelos más complejos, como los de machine learning, aportan mejoras reales en la capacidad predictiva (Perra, 2015).

En este sentido, los enfoques tradicionales no solo conservan relevancia por su utilidad analítica, sino también por su papel como baseline de comparación dentro de estudios orientados al desarrollo y evaluación de modelos predictivos más avanzados (Dixon, 2022).

2.8.5 Limitaciones de los enfoques tradicionales

Aunque los enfoques tradicionales han sido ampliamente utilizados en la vigilancia y predicción del dengue, presentan limitaciones cuando se aplican a fenómenos epidemiológicos complejos. Entre ellas destacan la dificultad para capturar relaciones no lineales, integrar múltiples variables heterogéneas y modelar conjuntamente factores climáticos, temporales y espaciales (Abade, 2024; Santangelo et al., 2023).

Además, revisiones recientes han identificado heterogeneidad metodológica y diferencias en los procesos de validación entre estudios, lo que dificulta la comparación de resultados. Por ello, aunque estos enfoques continúan siendo valiosos como referencia metodológica, también justifican la incorporación de modelos más flexibles como Random Forest y XGBoost (Leung, 2023).

Tabla 3
Limitaciones de los enfoques tradicionales en la predicción del dengue.

Limitación	Descripción	Implicación para esta investigación
Relaciones no lineales	Modelos como Poisson, binomial negativa o ARIMA/SARIMA pueden no capturar adecuadamente relaciones complejas entre clima, ambiente e incidencia.	Justifica el uso de modelos de machine learning con mayor flexibilidad.
Integración de variables heterogéneas	Presentan dificultades para incorporar simultáneamente variables climáticas, rezagos epidemiológicos, factores territoriales y temporales.	Refuerza la necesidad de modelos que integren distintas fuentes de información.

Interacciones espaciotemporales	Suelen centrarse en la dimensión temporal o en asociaciones estadísticas puntuales, con menor capacidad para modelar territorio y tiempo de forma conjunta.	Sustenta el componente geoespacial de la tesis.
Flexibilidad espacial limitada	No siempre incorporan de forma robusta la dimensión espacial ni la identificación dinámica de zonas de riesgo. La literatura reporta diferencias	Justifica el análisis espacial complementario.
Heterogeneidad metodológica	en variables, métodos de validación y métricas, lo que dificulta comparar resultados entre estudios.	Respalda la necesidad de una comparación metodológica clara en tu estudio.

2.9 Machine learning aplicado a la predicción epidemiológica

2.9.1 Definición general de machine learning

El *machine learning* o aprendizaje automático es una rama de la inteligencia artificial orientada al desarrollo de algoritmos capaces de identificar patrones en los datos y generar predicciones, clasificaciones o estimaciones a partir de experiencias previas contenidas en grandes volúmenes de información. A diferencia de los enfoques tradicionales basados en reglas explícitas o supuestos estadísticos rígidos, el aprendizaje automático permite que los modelos ajusten su comportamiento a partir de los datos observados, mejorando su capacidad de respuesta ante problemas complejos y dinámicos (Santangelo, 2023).

Desde una perspectiva metodológica, el *machine learning* se caracteriza por su capacidad para aprender relaciones entre variables sin requerir necesariamente una especificación previa completa de la forma funcional de esas relaciones. Esto lo hace especialmente útil en contextos donde intervienen múltiples factores simultáneos y donde las interacciones entre variables

pueden ser no lineales, heterogéneas o difíciles de representar mediante modelos clásicos. En el ámbito de la salud, esta característica ha favorecido su incorporación progresiva en tareas como el diagnóstico, la clasificación de riesgo, la predicción de enfermedades y el análisis de datos clínicos y epidemiológicos (Abdulazeem, 2023).

En términos generales, los métodos de aprendizaje automático pueden aplicarse a distintos tipos de tareas, entre ellas la clasificación, la regresión y la predicción de eventos futuros. Su uso se ha expandido en diversas disciplinas debido a su capacidad para procesar grandes volúmenes de datos, identificar patrones complejos y generar modelos adaptativos. En este sentido, el *machine learning* representa una alternativa analítica relevante cuando se busca fortalecer la capacidad predictiva de los sistemas basados en datos, especialmente en problemas donde la complejidad del fenómeno excede la capacidad explicativa de enfoques más tradicionales (Santangelo, 2023).

2.9.2 Machine learning en epidemiología

En el campo de la epidemiología, el *machine learning* se ha incorporado como una herramienta analítica útil para el procesamiento de grandes volúmenes de datos, la identificación de patrones complejos y la predicción del comportamiento de enfermedades en distintos contextos poblacionales. Su aplicación se ha fortalecido especialmente en escenarios donde los sistemas de vigilancia generan información heterogénea y donde los enfoques tradicionales presentan limitaciones para capturar relaciones complejas entre múltiples variables (Santangelo, 2023).

Desde esta perspectiva, el aprendizaje automático permite apoyar tareas como la predicción de incidencia, la detección temprana de brotes, la identificación de factores de riesgo y la estimación de eventos futuros a partir de datos históricos. En enfermedades infecciosas, su utilidad radica en la capacidad para integrar variables epidemiológicas, clínicas, ambientales y demográficas dentro de un mismo esquema analítico, lo que favorece una comprensión más

amplia de la dinámica de transmisión y fortalece la vigilancia basada en datos (Mohamed, Abdelrehim, & Al-Barazie, 2025).

Asimismo, diversos estudios han señalado que el *machine learning* puede complementar los enfoques epidemiológicos convencionales al mejorar la capacidad predictiva en contextos complejos y cambiantes. No obstante, su aplicación en salud pública debe desarrollarse con criterios metodológicos rigurosos, considerando la calidad de los datos, la validez del modelo y la pertinencia del contexto de uso. En este sentido, su incorporación en epidemiología no reemplaza la lógica clásica del análisis sanitario, sino que la amplía mediante herramientas capaces de responder a problemas de mayor complejidad analítica (Mohamed, Abdelrehim, & Al-Barazie, 2025).

2.9.3 Ventajas del machine learning frente a enfoques tradicionales

El *machine learning* presenta ventajas importantes frente a los enfoques tradicionales en predicción epidemiológica, especialmente cuando el comportamiento de la enfermedad depende de múltiples factores que interactúan de forma simultánea. Una de sus principales fortalezas es la capacidad para capturar relaciones no lineales, sin requerir que la relación entre variables predictoras y variable objetivo sea definida previamente mediante una forma funcional rígida (Santangelo et al., 2023).

Otra ventaja relevante es su capacidad para integrar variables heterogéneas dentro de un mismo modelo analítico. En epidemiología, esto permite combinar datos históricos de casos, variables climáticas, información demográfica y factores espaciales, lo cual resulta especialmente útil en enfermedades como el dengue, donde la incidencia depende de la interacción entre clima, territorio y rezagos epidemiológicos (Leung et al., 2023).

Asimismo, el *machine learning* puede ofrecer mejoras en desempeño predictivo frente a modelos estadísticos convencionales en escenarios complejos, aunque sus resultados dependen del contexto, la calidad de los datos y la estrategia de validación. Las revisiones comparativas

recientes muestran que estos modelos son especialmente útiles cuando los datos son dinámicos, multivariantes y presentan estructuras complejas que dificultan el ajuste de métodos tradicionales (Abade, 2024; Mohamed et al., 2025).

En el caso del dengue, estas ventajas son especialmente pertinentes, ya que la transmisión de la enfermedad está influida por variables meteorológicas, ambientales y epidemiológicas que pueden interactuar de manera cambiante en el tiempo y en el territorio. Por ello, el uso de *machine learning* se justifica como una alternativa metodológica capaz de complementar las limitaciones de los enfoques tradicionales y fortalecer la capacidad predictiva en estudios de vigilancia epidemiológica (Leung et al., 2023).

2.9.3.1 Expresión conceptual del enfoque predictivo. De manera general, la lógica del

machine learning en predicción puede representarse así:

$$Y = f(X_1, X_2, X_3, \dots, X_n)$$

Donde:

- Y representa la variable a predecir, por ejemplo, la incidencia del dengue.
- $X_1, X_2, X_3, \dots, X_n$ representan las variables predictoras, como temperatura, precipitación, humedad, rezagos históricos de casos y variables espaciales.
- f representa la función aprendida por el modelo a partir de los datos.

Esta formulación resume la lógica general del machine learning: aprender una función predictiva a partir de múltiples variables sin depender exclusivamente de supuestos lineales rígidos (Santangelo et al., 2023).

Tabla 4

Ventajas del machine learning frente a enfoques tradicionales en predicción epidemiológica.

Ventaja	Descripción	Relevancia para esta investigación
Captura de relaciones no lineales.	Permite modelar asociaciones complejas entre variables sin asumir una forma funcional estrictamente lineal.	Útil para representar la relación entre clima, rezagos de casos y dengue.
Integración de variables heterogéneas.	Puede combinar variables epidemiológicas, climáticas, demográficas y espaciales en un mismo esquema analítico.	Favorece la construcción de un modelo geoespacial con múltiples fuentes de datos.
Adaptabilidad a datos complejos.	Se ajusta mejor a estructuras de datos grandes, dinámicas y con múltiples interacciones.	Conveniente para trabajar con series históricas y variables ambientales.
Potencial de mejora predictiva.	Puede complementar o superar a modelos convencionales en ciertos contextos de predicción epidemiológica.	Justifica su comparación frente a baselines tradicionales.
Utilidad para problemas multivariados.	Resulta eficaz cuando la variable objetivo depende de muchos factores simultáneos.	Pertinente para modelar la incidencia del dengue en Guayaquil.

2.9.4 Aplicación del Machine Learning al Dengue

El *Machine Learning* se ha aplicado al dengue principalmente para predecir casos futuros, estimar riesgo de brotes e integrar variables epidemiológicas y meteorológicas dentro de modelos de pronóstico. Las revisiones recientes indican que este campo ha crecido de forma

importante y que los predictores más frecuentes incluyen temperatura, precipitación, humedad y casos históricos rezagados (Leung et al., 2023).

Entre los algoritmos más utilizados en la literatura se encuentran Random Forest, XGBoost, redes neuronales y otros modelos de ensamble. Estos métodos han sido empleados con datos semanales o mensuales y, en varios estudios, han mostrado buen desempeño cuando se combinan variables climáticas con series históricas de dengue (Al Mobin et al., 2024; Liu et al., 2025).

Además de pronosticar incidencia, los modelos de *machine learning* también se han usado para identificar variables influyentes en la dinámica del dengue. Estudios recientes con XGBoost y LightGBM reportan que la precipitación, la temperatura, la humedad y los rezagos de casos figuran entre los determinantes más relevantes para la predicción y la generación de alertas tempranas (Rahman et al., 2025).

No obstante, la literatura también señala que el desempeño de estos modelos depende del contexto geográfico, la calidad de los datos, la escala temporal y la estrategia de validación.

Por ello, su aplicación requiere comparación con modelos baseline tradicionales y una justificación metodológica clara del algoritmo seleccionado (Leung et al., 2023; Liu et al., 2025).

2.9.4.1 Fórmula conceptual del problema predictivo

$$Y_t = f(Y_{t-k}, T_{t-l}, P_{t-m}, H_{t-n}, S)$$

Donde:

Y_t : incidencia estimada de dengue en el tiempo.

Y_{t-k} : casos rezagados (donde k representa el número de semanas o meses de retraso).

T_{t-l} : temperatura con rezago l .

P_{t-m} : precipitación con rezago m .

H_{t-n} : humedad con rezago.

S : variable o componente espacial (territorial).

Esta formulación resume cómo los modelos de *machine learning* integran información histórica, climática y territorial para predecir dengue (Leung et al., 2023; Al Mobin et al., 2024).

Tabla 5*Aplicaciones del machine learning en la predicción del dengue*

Aplicación	Descripción	Relevancia para esta investigación
Pronóstico de casos	Estimación del número futuro de casos de dengue a partir de series históricas y variables exógenas.	Sustenta el objetivo predictivo del estudio.
Predicción de brotes	Identificación anticipada de incrementos o picos de transmisión.	Aporta valor para vigilancia y alerta temprana.
Integración de variables climáticas	Uso de temperatura, precipitación y humedad como predictores del dengue.	Justifica la inclusión de datos ambientales en el modelo.
Modelado multivariable	Combinación de casos históricos, clima y variables temporales en un mismo esquema.	Respalda la construcción de un dataset integrado.
Comparación con modelos tradicionales	Evaluación del desempeño de ML frente a ARIMA, Poisson u otros baselines.	Se alinea con la lógica comparativa del proyecto.

2.9.5 Limitaciones y Precauciones en el uso del Machine Learning

Aunque el *machine learning* ofrece ventajas importantes para la predicción epidemiológica, su aplicación también presenta limitaciones que deben considerarse de forma crítica. Una de las principales es su dependencia de la calidad, cantidad y consistencia de los datos, ya que modelos entrenados con información incompleta, sesgada o poco representativa pueden generar predicciones inestables o poco generalizables (Leung et al., 2023; Mohamed et al., 2025).

Otra precaución relevante es la interpretabilidad. Algunos modelos de *machine learning*, especialmente los más complejos, pueden comportarse como “cajas negras”, lo que dificulta

comprender con claridad cómo influyen las variables en la predicción final. En contextos de salud pública, esta limitación puede afectar la confianza institucional y la adopción de resultados para la toma de decisiones, razón por la cual varios estudios recientes enfatizan la necesidad de equilibrio entre precisión predictiva e interpretabilidad (Mohamed et al., 2025).

Asimismo, el *machine learning* requiere procesos rigurosos de validación. Las revisiones sistemáticas sobre modelos de predicción del dengue muestran heterogeneidad en los diseños metodológicos, en las métricas de evaluación y en los esquemas de validación, lo que dificulta comparar resultados entre estudios y determinar con precisión qué enfoques son más robustos en condiciones homogéneas. También se ha reportado que una parte de los estudios presenta validación insuficiente o escasa validación externa, lo cual limita la aplicabilidad real de los modelos (Leung et al., 2023).

Finalmente, el desempeño de estos modelos depende en gran medida del contexto geográfico, temporal y epidemiológico. Un algoritmo que funciona adecuadamente en una región o en un periodo determinado no necesariamente tendrá el mismo comportamiento en otro contexto, especialmente cuando cambian las condiciones climáticas, la estructura poblacional o la calidad de los datos disponibles. Por ello, el uso del *machine learning* en dengue debe acompañarse de una justificación metodológica clara, una selección cuidadosa de variables y una comparación explícita con modelos baseline tradicionales (Mohamed et al., 2025; Leung et al., 2023).

2.9.6 Relación con la presente investigación

En la presente investigación, el uso de *machine learning* se justifica por la necesidad de analizar un problema epidemiológico complejo en el que intervienen simultáneamente factores históricos, climáticos y espaciales. La incidencia del dengue en entornos urbanos como Guayaquil no depende de una única variable, sino de la interacción entre casos rezagados, temperatura, precipitación, humedad y condiciones territoriales, lo que hace pertinente el empleo de modelos

capaces de capturar relaciones no lineales y estructuras de datos heterogéneas (Leung et al., 2023).

Desde esta perspectiva, el *machine learning* no se incorpora en esta investigación como una tendencia tecnológica, sino como una alternativa metodológica coherente con la naturaleza del fenómeno estudiado. Su uso busca fortalecer la capacidad predictiva del análisis epidemiológico del dengue y complementar las limitaciones observadas en los enfoques tradicionales, especialmente en lo relacionado con la integración de múltiples variables y la representación de patrones complejos en el tiempo y el espacio (Santangelo et al., 2023).

Asimismo, la elección de este enfoque se alinea con la evidencia reciente que señala que el contexto es determinante en los modelos de predicción de enfermedades. Por ello, la presente investigación plantea el desarrollo y evaluación de modelos de *machine learning* aplicados específicamente a la ciudad de Guayaquil, utilizando datos epidemiológicos históricos y variables ambientales, dentro de un enfoque comparativo frente a modelos baseline tradicionales. De esta manera, el valor del estudio no radica únicamente en aplicar algoritmos predictivos, sino en evaluar su pertinencia y desempeño dentro de un contexto urbano concreto y epidemiológicamente relevante (Mohamed et al., 2025).

2.10 Random Forest y XGBoost como modelos de predicción

2.10.1 Random Forest

Random Forest es un algoritmo de aprendizaje automático basado en ensambles de árboles de decisión, cuyo principio consiste en construir múltiples árboles a partir de distintas submuestras de los datos y combinar sus resultados mediante promediación o votación. Este mecanismo permite mejorar la precisión predictiva y, al mismo tiempo, reducir el sobreajuste que suele presentarse cuando se utiliza un único árbol de decisión (Breiman, 2001; Scikit-learn developers, 2026).

Una de las principales ventajas de Random Forest es su capacidad para manejar variables heterogéneas y capturar relaciones no lineales entre predictores y variable objetivo. En contextos epidemiológicos, esto resulta especialmente útil porque la ocurrencia de una enfermedad puede depender simultáneamente de factores climáticos, demográficos, territoriales e históricos. Además, el algoritmo ofrece medidas de importancia de variables, lo que facilita identificar cuáles predictores aportan más al modelo y fortalece la interpretación analítica de los resultados (Oehm et al., 2023; Scikit-learn developers, 2026).

En el ámbito de la predicción de enfermedades infecciosas, Random Forest ha sido utilizado con frecuencia debido a su robustez, estabilidad y buen desempeño en escenarios con datos complejos. Estudios recientes en epidemiología destacan que los modelos basados en árboles, incluyendo Random Forest, son particularmente útiles cuando se busca integrar múltiples fuentes de información y generar pronósticos a partir de patrones difíciles de representar mediante enfoques lineales tradicionales. Asimismo, comparaciones metodológicas en pronóstico de enfermedades infecciosas muestran que estos modelos pueden desempeñarse de forma competitiva frente a métodos estadísticos clásicos, especialmente cuando se trabaja con estructuras multivariantes (Dixon et al., 2022; Zhao et al., 2020).

En el caso de esta investigación, Random Forest resulta pertinente porque la predicción del dengue en Guayaquil involucra variables epidemiológicas históricas, rezagos temporales, factores climáticos y componentes espaciales. La capacidad del algoritmo para modelar relaciones no lineales, manejar interacciones complejas y estimar la relevancia de las variables lo convierte en una opción metodológicamente adecuada para evaluar su desempeño frente a modelos baseline tradicionales (Zhao et al., 2020; Scikit-learn developers, 2026).

2.10.1.1 Expresión conceptual del modelo

$$\hat{Y} = \frac{1}{B} \sum_{b=1}^B T_b(X)$$

Donde:

$\hat{Y}$ representa la predicción final del modelo.

B es el número de árboles construidos en el bosque.

$T_b(X)$ es la predicción generada por cada árbol a partir del conjunto de variables X.

Esta formulación resume la lógica de Random Forest: combinar múltiples árboles para obtener una predicción más estable y con menor riesgo de sobreajuste (Breiman, 2001; Scikit-learn developers, 2026).

2.10.2 XGBoost

XGBoost (*eXtreme Gradient Boosting*) es un algoritmo de aprendizaje automático basado en el enfoque de gradient boosting, en el que varios árboles de decisión se construyen de manera secuencial para corregir progresivamente los errores de los árboles anteriores. Este método fue propuesto por Chen y Guestrin (2016) como un sistema escalable de *tree boosting* orientado a mejorar el rendimiento predictivo y la eficiencia computacional en problemas de clasificación y regresión.

Una de las principales fortalezas de XGBoost es su alto desempeño predictivo en datos tabulares, junto con mecanismos de regularización que ayudan a reducir el sobreajuste y mejorar la generalización del modelo (Chen & Guestrin, 2016). Además, la documentación técnica del algoritmo destaca que incorpora estrategias de optimización computacional, paralelización y métodos eficientes de construcción de árboles, lo que favorece su uso en problemas con múltiples variables y estructuras de datos complejas (XGBoost Developers, 2026a, 2026b).

En el ámbito epidemiológico, y particularmente en estudios recientes sobre dengue, XGBoost ha mostrado utilidad para pronosticar incidencia, integrar variables meteorológicas y

comparar su desempeño con otros enfoques de *machine learning* y modelos estadísticos. La literatura reciente señala que este tipo de modelos puede trabajar con series semanales o mensuales e incorporar variables como temperatura, precipitación, humedad y rezagos históricos de casos, aportando capacidad para modelar relaciones complejas en escenarios de vigilancia y alerta temprana (da Costa et al., 2026).

Desde el punto de vista metodológico, XGBoost resulta pertinente para esta investigación porque la predicción del dengue en Guayaquil involucra relaciones complejas entre variables epidemiológicas, ambientales y espaciales. Su capacidad para capturar no linealidad, modelar interacciones entre predictores y mantener un buen desempeño en problemas multivariados lo convierte en una opción adecuada para comparar frente a modelos baseline tradicionales y frente a otros algoritmos de ensamble como Random Forest (Chen & Guestrin, 2016; XGBoost Developers, 2026a).

2.10.2.1 Expresión conceptual del modelo

$$\hat{Y}^{(t)} = \hat{Y}^{(t-1)} + \eta f_t(X)$$

Donde:

$\hat{Y}^{(t)}$ es la predicción en la iteración t

$\hat{Y}^{(t-1)}$ es la predicción acumulada hasta la iteración anterior

η es la tasa de aprendizaje

$f_t(X)$ es el nuevo árbol que corrige el error residual

Esta formulación resume la lógica de XGBoost: construir árboles de forma secuencial para mejorar progresivamente la predicción del modelo (Chen & Guestrin, 2016; XGBoost Developers, 2026b).

Tabla 6

Características de XGBoost y su relevancia

Característica	Descripción	Relevancia para la investigación
Boosting secuencial	Cada árbol nuevo corrige errores de los árboles previos.	Favorece un ajuste progresivo del modelo para la predicción del dengue (Chen & Guestrin, 2016).
Regularización	Incorpora mecanismos para reducir el sobreajuste.	Útil para mejorar la generalización del modelo con datos epidemiológicos (Chen & Guestrin, 2016).
Alto desempeño en datos tabulares	Está diseñado para resolver con rapidez y precisión problemas de ciencia de datos en tablas.	Pertinente para un dataset con variables climáticas, epidemiológicas y espaciales (XGBoost Developers, 2026a).
Eficiencia computacional	Utiliza optimizaciones de entrenamiento y métodos de árbol escalables.	Conveniente para experimentar con distintas configuraciones del modelo (XGBoost Developers, 2026c).
Aplicación en dengue	Se ha utilizado en estudios recientes de pronóstico de incidencia y alerta temprana.	Respalda su selección como modelo de interés en esta investigación (da Costa et al., 2026).

2.10.3 Justificación de selección

La selección de Random Forest y XGBoost en la presente investigación responde a criterios metodológicos y prácticos coherentes con la naturaleza del problema de estudio. La predicción espacial del dengue en Guayaquil implica trabajar con una variable dependiente influida por múltiples factores epidemiológicos, climáticos y territoriales, cuyas relaciones pueden ser no lineales y difíciles de representar mediante modelos tradicionales. En este contexto, ambos

algoritmos ofrecen una alternativa adecuada por su capacidad para manejar estructuras de datos complejas y multivariantes (Leung et al., 2023; Chen & Guestrin, 2016).

Random Forest se selecciona por su robustez, su capacidad para reducir el sobreajuste mediante el uso de múltiples árboles y su utilidad para trabajar con variables heterogéneas sin requerir una especificación funcional estricta. Además, permite estimar la importancia de los predictores, lo cual resulta valioso para identificar cuáles variables ambientales y epidemiológicas influyen más en la incidencia del dengue. Esta característica aporta no solo capacidad predictiva, sino también una ventaja interpretativa dentro del análisis epidemiológico (Zhao et al., 2020).

Por su parte, XGBoost se justifica por su alto desempeño en datos tabulares, su capacidad para capturar relaciones complejas entre variables y sus mecanismos de regularización, que ayudan a controlar el sobreajuste. La literatura técnica y aplicada reconoce a XGBoost como uno de los algoritmos más competitivos para problemas predictivos basados en árboles, especialmente cuando se busca optimizar precisión y eficiencia computacional en contextos multivariantes (Chen & Guestrin, 2016).

La elección conjunta de ambos modelos permite, además, realizar una comparación metodológica entre dos enfoques de ensamble con lógicas diferentes. Mientras Random Forest se basa en la agregación paralela de múltiples árboles, XGBoost construye árboles de manera secuencial para corregir errores previos. Esta diferencia resulta relevante para la investigación, ya que permite evaluar cuál de los dos enfoques se adapta mejor a la predicción del dengue cuando se integran variables históricas, meteorológicas y espaciales en un contexto urbano como Guayaquil.

Finalmente, la selección de Random Forest y XGBoost también responde a un criterio de viabilidad académica. Ambos algoritmos son ampliamente utilizados, cuentan con implementaciones estables en entornos de ciencia de datos y han sido aplicados previamente en estudios de predicción del dengue y otras enfermedades infecciosas. Por ello, constituyen

opciones metodológicamente pertinentes para compararse con los modelos baseline tradicionales definidos en esta investigación.

Tabla 7

Justificación de selección de Random Forest y XGBoost.

Criterio	Random Forest	XGBoost	Relevancia para la tesis
Manejo de no linealidad	Sí	Sí	Adecuado para relaciones complejas entre clima, rezagos y dengue.
Variables heterogéneas	Alto	Alto	Permite integrar datos epidemiológicos, ambientales y espaciales.
Control de sobreajuste	Mediante ensamble de árboles.	Mediante regularización y boosting.	Favorece la robustez del modelo.
Interpretabilidad relativa.	Media, con importancia de variables.	Media, con análisis de contribución.	Útil para identificar predictores relevantes.
Aplicación previa en dengue	Reportada en estudios comparativos	Creciente en estudios recientes	Sustenta su pertinencia metodológica.
Viabilidad para un TFM	Alta	Alta	Disponibles en Python y adecuados para datos tabulares.

2.10.4 Comparación breve con otras alternativas

La selección de Random Forest y XGBoost se justifica también al compararlos con otras alternativas de predicción epidemiológica. Frente a los modelos lineales y lineales generalizados, estos algoritmos ofrecen mayor flexibilidad para capturar relaciones no lineales e interacciones

complejas entre variables climáticas, temporales y territoriales, sin depender de supuestos funcionales rígidos (Leung et al., 2023; Zhao et al., 2020).

En comparación con métodos como las máquinas de soporte vectorial y las redes neuronales profundas, Random Forest y XGBoost presentan ventajas prácticas para esta investigación, ya que combinan buen desempeño predictivo con mayor viabilidad de implementación, menor complejidad de ajuste y una interpretabilidad relativa más adecuada para un estudio aplicado. Por ello, ambos modelos resultan metodológicamente pertinentes para predecir la incidencia del dengue y compararse con los baselines tradicionales definidos en esta investigación (Santangelo et al., 2023; Mohamed et al., 2025).

Tabla 8
Comparación breve con otras alternativas.

Alternativa	Limitación principal	Ventaja de Random Forest y XGBoost
Modelos lineales / GLM	Mayor rigidez funcional	Capturan relaciones no lineales.
SVM	Mayor complejidad de ajuste	Mejor adaptabilidad a datos tabulares.
Redes neuronales profundas.	Menor interpretabilidad y mayor exigencia de datos.	Mejor equilibrio entre precisión y viabilidad.
Árbol único	Mayor riesgo de sobreajuste	Mayor robustez mediante ensamble.

2.11 Estado del arte comparativo

2.11.1 Estudios con modelos tradicionales

Los estudios sobre predicción del dengue mediante enfoques tradicionales se han desarrollado principalmente a partir de dos líneas metodológicas: los modelos de regresión

para datos de conteo, como Poisson y binomial negativa, y los modelos de series temporales, como ARIMA y SARIMA. Estos enfoques han sido ampliamente utilizados debido a su solidez estadística, su relativa facilidad de implementación y su compatibilidad con datos epidemiológicos agregados, especialmente series semanales o mensuales de casos (Leung et al., 2023).

En el caso de los modelos de regresión para conteos, la literatura reporta su uso para analizar la asociación entre la incidencia del dengue y variables climáticas como temperatura, precipitación y humedad. Este tipo de modelos ha permitido estimar el efecto de factores ambientales sobre la ocurrencia de casos y explicar parte de la variabilidad observada en distintos contextos geográficos. No obstante, sus resultados dependen en gran medida de la correcta especificación del modelo y presentan limitaciones cuando las relaciones entre variables son altamente no lineales o cuando se requiere integrar simultáneamente componentes espaciales y temporales complejos (Chumpu et al., 2019; Hossain et al., 2022).

Por su parte, los modelos de series temporales han sido utilizados con frecuencia para pronosticar el comportamiento del dengue a partir de su propia dinámica histórica. Estudios basados en ARIMA y SARIMA muestran que estos enfoques son útiles para captar tendencia, autocorrelación y estacionalidad, particularmente en contextos donde la enfermedad presenta patrones cíclicos asociados al clima. Sin embargo, su capacidad predictiva puede verse limitada cuando el fenómeno depende de múltiples variables exógenas o cuando la serie presenta cambios estructurales importantes (Naher et al., 2022; Hasan et al., 2024).

En conjunto, ambos enfoques han contribuido de manera importante al análisis epidemiológico del dengue y continúan siendo referencias válidas dentro de la literatura. Sin embargo, también comparten restricciones para modelar relaciones no lineales complejas, integrar múltiples fuentes de información y representar de manera flexible la dimensión espaciotemporal del problema. Por ello, en la presente investigación se consideran

como modelos baseline tradicionales, frente a los cuales se evaluará el desempeño de los enfoques de *machine learning* propuestos (Leung et al., 2023).

Tabla 9

Síntesis comparativa de estudios con modelos tradicionales en dengue.

Tipo de enfoque	Estudios representativos	Variables frecuentes	Principales aportes	Limitaciones
Regresión para conteos	Poisson, binomial negativa, GLM	Temperatura, precipitación, humedad, casos agregados	Permiten estimar asociaciones entre clima e incidencia y trabajar con conteos de casos.	Menor flexibilidad para modelar no linealidad compleja y dimensión espacial.
Series temporales	ARIMA, SARIMA	Casos históricos semanales o mensuales, estacionalidad, clima en algunos estudios	Capturan tendencia, autocorrelación y estacionalidad del dengue.	Dependen de series estables y son menos flexibles ante estructuras multivariantes complejas.

2.11.2 Estudios con machine learning

La literatura reciente muestra un crecimiento sostenido del uso de *machine learning* para la predicción del dengue, especialmente en contextos donde se integran series históricas de casos con variables meteorológicas y, en algunos estudios, indicadores entomológicos o espaciales. Una revisión sistemática identificó que entre los predictores más frecuentes se encuentran la lluvia, la temperatura y la humedad, y reportó además que los modelos con mejor desempeño incluían redes neuronales y métodos de ensamble, aunque con una alta heterogeneidad metodológica entre estudios (Sylvestre et al., 2022).

En estudios aplicados, Zhao et al. (2020) compararon Random Forest y redes neuronales artificiales para predecir la carga de dengue a escala nacional y subnacional en Colombia. Sus resultados mostraron que Random Forest podía generar pronósticos precisos incluso para horizontes de varias semanas, y que un modelo nacional podía trasladarse a escalas subnacionales con buen desempeño. Este trabajo es relevante para la presente tesis porque demuestra la utilidad de los modelos de ensamble para integrar múltiples fuentes de datos epidemiológicos y climáticos en la predicción del dengue.

Más recientemente, estudios en Bangladesh han ampliado la comparación entre algoritmos. Al Mobin et al. (2024) desarrollaron modelos de pronóstico utilizando factores meteorológicos, confirmando la utilidad de variables como temperatura, precipitación y humedad para alimentar modelos de *machine learning*. En la misma línea, Liu et al. (2025) compararon múltiples técnicas de *machine learning* para pronosticar dengue con datos de vigilancia de cinco divisiones de Bangladesh y destacaron la relevancia de trabajar con series que presentan tendencia y estacionalidad. Estos estudios refuerzan la idea de que el desempeño del modelo depende no solo del algoritmo, sino también de la calidad de los datos, la escala territorial y la estrategia de validación.

De manera comparativa, los estudios revisados coinciden en que el *machine learning* ofrece ventajas para integrar múltiples predictores y capturar relaciones no lineales; sin embargo, también muestran limitaciones comunes, como diferencias en variables de entrada, métricas, horizontes de predicción y estrategias de validación. En consecuencia, aunque existe evidencia favorable sobre algoritmos como Random Forest, XGBoost y redes neuronales, la literatura todavía presenta dificultades para comparar enfoques bajo condiciones homogéneas. Este punto justifica que la presente investigación establezca una comparación explícita entre modelos de *machine learning* y modelos baseline tradicionales dentro de un contexto urbano específico como Guayaquil (Sylvestre et al., 2022; Liu et al., 2025).

Tabla 10*Síntesis comparativa de estudios con machine learning en dengue.*

Estudio	Contexto	Método	Variables frecuentes	Hallazgo principal	Relación con esta tesis
Zhao et al. (2020)	Colombia, escala nacional y subnacional	Random Forest y ANN	Casos históricos, clima, variables demográficas	Random Forest mostró buen desempeño para pronóstico multi-semana y utilidad a distintas escalas.	Sustenta el uso de Random Forest como modelo pertinente para predicción de dengue.
Al Mobin et al. (2024)	Bangladesh	Modelos de <i>machine learning</i> con factores meteorológicos	Temperatura, precipitación, humedad	Confirma la relevancia de variables climáticas para pronosticar dengue.	Refuerza la integración de variables ambientales en el modelo propuesto.
Liu et al. (2025)	Bangladesh, cinco divisiones	Comparación de múltiples técnicas de <i>machine learning</i>	Datos de vigilancia, patrón temporal, estacionalidad	Muestra que el rendimiento depende del algoritmo y de la estructura temporal de la serie.	Respalda la comparación entre algoritmos y la necesidad de validación rigurosa.
Francisco et al. (2024)	Contexto espaciotemporal con datos de dengue	Enfoque híbrido de <i>machine learning</i>	Resolución espacial, temporal y datos con ceros	La resolución espaciotemporal afecta	Justifica prestar atención a la unidad espacial

		<i>e learning</i>		la precisión predictiva.	de análisis en Guayaquil.
				Reporta alta	Sustenta el
		Revisión de	Lluvia,	heterogenei	enfoque
Sylvestre et al. (2022)	Revisión sistemática	métodos basados en datos	temperatura, humedad	dad metodo lógica y buen	comparativo y la necesidad de estandarizar
				desempeño de modelos	evaluación.
				avanzados.	

2.11.3 Estudios con análisis espacial del dengue

La literatura sobre dengue muestra con bastante consistencia que la enfermedad no se distribuye de manera homogénea en el territorio, sino que tiende a concentrarse en áreas específicas donde confluyen condiciones ambientales, urbanas y sociales favorables para la transmisión. En ese marco, los estudios de análisis espacial han recurrido de forma creciente a los Sistemas de Información Geográfica (SIG), la detección de hotspots, el mapeo de riesgo y distintos modelos espaciales para localizar zonas críticas y apoyar decisiones de vigilancia y control (Nayak et al., 2025).

Más que limitarse a describir dónde ocurren los casos, varios trabajos han mostrado que el análisis espacial permite identificar patrones persistentes de transmisión. En Singapur, por ejemplo, Gurram et al. (2025) analizaron la distribución espaciotemporal de hotspots sostenidos mediante Geographically Weighted Regression (GWR) y encontraron asociaciones entre los conglomerados de dengue, las condiciones climáticas y la urbanización. De forma semejante, Masrani et al. (2022), en Kelantan, Malasia, utilizaron ArcGIS y el estadístico Getis-Ord Gi* para seguir los cambios anuales en la agrupación espacial de casos, evidenciando que los focos de dengue se concentran en zonas específicas y cambian con el tiempo. En ambos casos, el territorio

deja de ser un simple contenedor de casos y pasa a convertirse en una dimensión explicativa del fenómeno epidemiológico (Gurram et al., 2025; Masrani et al., 2022).

Otros estudios han ido un paso más allá al incorporar variables ambientales y socioeconómicas para construir mapas de vulnerabilidad o riesgo. Haryanto et al. (2025), en Indonesia, identificaron áreas vulnerables al dengue a partir de factores ambientales y sociales, mientras que Lalangui-Vivanco et al. (2025), en Manta, Ecuador, modelaron la distribución espacial del riesgo de transmisión en una ciudad costera con alta incidencia. Estos trabajos son especialmente valiosos porque no solo permiten ubicar espacialmente la enfermedad, sino también aproximarse a las condiciones que favorecen su persistencia en determinados sectores urbanos. Para una investigación como la presente, esa lógica resulta especialmente pertinente, ya que refuerza la necesidad de incorporar la dimensión territorial en el análisis del dengue en un contexto costero ecuatoriano (Haryanto et al., 2025; Lalangui-Vivanco et al., 2025).

Sin embargo, al comparar estos estudios aparece una limitación recurrente: muchos de ellos logran describir con precisión la distribución espacial del dengue, detectar conglomerados o generar mapas de riesgo, pero no siempre integran esa información dentro de modelos predictivos comparativos. En otras palabras, el análisis espacial suele estar muy bien desarrollado como herramienta descriptiva y de apoyo a la vigilancia, pero con menor frecuencia se articula con enfoques de predicción que integren simultáneamente variables epidemiológicas históricas, clima y territorio. Precisamente ahí se sitúa el aporte potencial de esta investigación, que busca vincular la lógica espacial del dengue con modelos de *machine learning* en el contexto urbano de Guayaquil (Nayak et al., 2025; Lalangui-Vivanco et al., 2025).

Tabla 11*Síntesis comparativa de estudios con análisis espacial del dengue.*

Estudio	Contexto	Enfoque espacial	Hallazgo principal	Relación con esta investigación
Nayak et al. (2025)	Revisión de vigilancia de dengue	GIS, hotspot mapping, risk mapping	El GIS se usa ampliamente para cartografía de riesgo, identificación de hotspots y modelado espacial en dengue.	Sustenta la incorporación del componente geoespacial en la investigación.
Gurram et al. (2025)	Singapur	GWR y análisis de hotspots.	Identificó hotspots sostenidos y su asociación con clima y urbanización.	Respalda el análisis espacial como complemento del pronóstico.
Masrani et al. (2022)	Kelantan, Malasia	ArcGIS y Getis-Ord Gi*	Detectó cambios anuales en los patrones de agrupamiento espacial del dengue.	Apoya el uso de técnicas de hotspots para interpretar la distribución territorial.
Haryanto et al. (2025)	Indonesia	Mapeo de vulnerabilidad espacial.	Integró variables ambientales y socioeconómicas para identificar zonas propensas al dengue.	Refuerza el valor de incorporar variables territoriales y socioambientales.
Lalangui-Vivanco et al. (2025)	Manta, Ecuador	Modelado espacial de riesgo	Mostró alta heterogeneidad espacial del dengue en una ciudad costera ecuatoriana.	Es especialmente relevante por su cercanía contextual con el caso de estudio.

2.12 Vacío de investigación

La literatura revisada evidencia que el dengue ha sido estudiado desde tres enfoques principales: modelos tradicionales de predicción, modelos de *machine learning* y análisis espacial. Los primeros han permitido describir asociaciones entre incidencia y variables climáticas, así como pronosticar tendencias temporales; los segundos han ampliado la capacidad para modelar

relaciones no lineales y múltiples predictores; y los terceros han mostrado que la transmisión del dengue presenta una distribución territorial heterogénea y que el análisis geoespacial aporta valor para la vigilancia epidemiológica (Leung et al., 2023; Sylvestre et al., 2022).

Sin embargo, persisten tres vacíos relevantes. En primer lugar, la comparación entre modelos tradicionales y modelos de *machine learning* sigue siendo metodológicamente heterogénea, debido a diferencias en variables de entrada, métricas de evaluación, horizontes de predicción y estrategias de validación, lo que dificulta establecer conclusiones comparables entre estudios. En segundo lugar, no es frecuente encontrar investigaciones que integren de manera simultánea datos epidemiológicos históricos, variables ambientales y análisis espacial dentro de un mismo esquema predictivo. En tercer lugar, la evidencia aplicada al contexto urbano ecuatoriano sigue siendo limitada y, en particular, son escasos los estudios centrados en Guayaquil con un enfoque comparativo y espacialmente explícito (Leung et al., 2023; Sylvestre et al., 2022).

En este contexto, se identifica como vacío de investigación la ausencia de estudios que evalúen comparativamente modelos baseline tradicionales y modelos de *machine learning*, integrando información epidemiológica, ambiental y espacial para la predicción del dengue en Guayaquil. Aunque existen antecedentes recientes en ciudades ecuatorianas costeras, como el modelado espacial desarrollado en Manta, todavía es limitada la evidencia que combine comparación metodológica y predicción espacial aplicada específicamente a Guayaquil (Lalanguí-Vivanco et al., 2025).

A diferencia de estudios previos que han abordado el dengue desde perspectivas aisladas (modelos tradicionales, ML o análisis espacial), la presente investigación contribuye al conocimiento en cuatro aspectos:

- (i) la integración sistemática de datos epidemiológicos, climáticos y espaciales en un pipeline reproducible;

- (ii) la comparación explícita de múltiples familias de modelos (lineales, regularizados, basados en árboles) bajo condiciones de validación homogéneas;
- (iii) la evaluación de la estabilidad interanual del modelo, revelando que la predicción del dengue es altamente sensible a la variabilidad entre años; y
- (iv) la identificación de parroquias como casos atípicos que requiere investigación específica, demostrando que la agregación parroquial puede ocultar factores locales determinantes.

2.13 Síntesis del capítulo

En este capítulo se revisaron los fundamentos teóricos y los antecedentes metodológicos que sustentan la investigación. En primer lugar, se abordó el dengue como problema de salud pública y la influencia de factores climáticos, ambientales y territoriales en su transmisión. Posteriormente, se analizaron los enfoques tradicionales de predicción, los métodos de *machine learning* aplicados a la epidemiología y los estudios de análisis espacial del dengue, identificando sus principales aportes y limitaciones.

La revisión del estado del arte permitió establecer que, aunque existen avances relevantes en modelos tradicionales, enfoques de *machine learning* y análisis espacial, todavía persisten vacíos en la comparación metodológica entre modelos y en la integración conjunta de información epidemiológica, ambiental y espacial, especialmente en contextos urbanos como Guayaquil.

A partir de este sustento teórico y empírico, en el siguiente capítulo se desarrollará la metodología de la investigación, en la que se definirán el enfoque metodológico, las fuentes de datos, las variables, los modelos baseline y de *machine learning*, así como la estrategia de validación y análisis comparativo que permitirá responder al problema planteado.

CAPÍTULO 3:

3 METODOLOGÍA Y DESARROLLO

3.1 Enfoque de la investigación

La presente investigación se enmarca en un enfoque cuantitativo, dado que se fundamenta en la recopilación, procesamiento y análisis estadístico de datos numéricos provenientes de fuentes secundarias oficiales (Hernández-Sampieri & Mendoza, 2018). Este enfoque es apropiado cuando el problema de investigación requiere medir fenómenos, probar hipótesis y analizar relaciones entre variables cuantificables (Creswell & Creswell, 2018).

El estudio es de naturaleza aplicada, ya que no busca únicamente generar conocimiento teórico, sino desarrollar un modelo predictivo útil para la vigilancia epidemiológica del dengue en Guayaquil. La investigación aplicada se caracteriza por orientar sus hallazgos hacia la solución de problemas prácticos y concretos (Polit & Beck, 2020).

Finalmente, la investigación tiene una orientación predictiva, ya que su propósito principal es el de estimar la incidencia espacial futura de la enfermedad a partir de patrones aprendidos de datos históricos (Shmueli, 2010). Este enfoque predictivo se distingue del explicativo en que prioriza la precisión de las predicciones sobre la interpretabilidad de los coeficientes (Breiman, 2001a).

3.2 Tipo, nivel y diseño de investigación

Al estudio se lo puede clasificar como una investigación aplicada, orientada a resolver un problema práctico —la limitada capacidad predictiva espacial del dengue en Guayaquil— mediante el desarrollo de modelos basados en machine learning geoespacial (Kuhn & Johnson, 2019).

El nivel de conocimiento es correlacional-predictivo, ya que no solo se describen relaciones entre variables (casos históricos, clima, territorio), sino que se busca cuantificar la

capacidad de estas variables para predecir la incidencia futura y se compara el desempeño de distintos enfoques metodológicos (Aiken et al., 2021).

El diseño de investigación es no experimental y longitudinal retrospectivo (Campbell & Stanley, 1963; Shadish et al., 2002):

No experimental: no se manipulan deliberadamente las variables; los datos observados corresponden a eventos ya ocurridos tales como son los casos de dengue y las condiciones climáticas registradas.

Longitudinal retrospectivo: se analiza un periodo continuo de tiempo (2021-2025) con una frecuencia semanal, utilizando datos pasados para construir y evaluar predicciones (Wang et al., 2020).

3.3 Ámbito y unidad de estudio

Ámbito geográfico: el estudio se circunscribe a las parroquias urbanas del cantón Guayaquil, provincia del Guayas, Ecuador. Guayaquil representa un contexto de especial interés epidemiológico debido a su densidad urbana, condiciones climáticas favorables para la proliferación del vector *Aedes aegypti* y presencia recurrente de casos de dengue (Ministerio de Salud Pública del Ecuador [MSP], 2025; Castillo et al., 2022).

Se incluyen exclusivamente las 15 parroquias urbanas consolidadas: Ayacucho, Bolívar, Carbo, Febres Cordero, García Moreno, Letamendi, Nueve de Octubre, Olmedo, Roca, Rocafuerte, Sucre, Tarqui, Urdaneta y Ximena. Quedan excluidas las parroquias rurales y suburbanas debido a que sus dinámicas de transmisión del dengue responden a factores ecosistémicos distintos (actividades agrícolas, saneamiento rural, menor densidad poblacional) y, principalmente, porque el sistema de vigilancia epidemiológica no desagrega los casos a ese nivel con la misma sistematicidad (MSP, 2025).

Unidad espacial de análisis: la parroquia urbana. Esta elección responde a la disponibilidad de datos epidemiológicos históricos, que constituyen la variable dependiente del estudio. Si bien podrían considerarse unidades más finas (sectores censales, cuadrículas regulares

o distritos), estas no cuentan con registros sistemáticos de casos de dengue, lo que haría inviable la construcción del dataset (Pfeiffer et al., 2008). Por lo tanto, la parroquia urbana representa el mejor equilibrio entre resolución espacial y viabilidad operativa.

Unidad temporal de análisis: la semana epidemiológica (definición de la Organización Panamericana de la Salud [OPS]: domingo a sábado). Esta resolución es estándar en vigilancia epidemiológica, captura adecuadamente la dinámica de transmisión del dengue (periodos de incubación de 7-14 días) y permite alinear los registros del MSP con las variables climáticas de NASA POWER (Lowe et al., 2021).

Periodo de análisis: desde la semana epidemiológica 1 de 2021 hasta la semana 52 de 2025. Este rango fue definido con base en la disponibilidad de registros históricos consistentes proporcionados por el Ministerio de Salud Pública (MSP) para el cantón Guayaquil, y cubre un periodo de cinco años, incluyendo años epidémicos y no epidémicos (Ministerio de Salud Pública, 2025).

3.4 Fuentes de datos

3.4.1 Fuente epidemiológica: Ministerio de Salud Pública (MSP)

Los datos de casos de dengue fueron obtenidos del archivo Dengue_Guayaquil_2016_2026_MSP.csv, proporcionado por la Dirección Nacional de Vigilancia Epidemiológica del Ministerio de Salud Pública del Ecuador. Este archivo consolida los registros del Sistema Integrado de Vigilancia Epidemiológica (SIVE-ALERTA) para el cantón Guayaquil, con una serie histórica que abarca desde 2016 hasta la semana 21 de 2026. El archivo contiene 11 columnas con la siguiente estructura:

Tabla

12

Fuentes de datos Ministerio de Salud Pública (MSP)

Columna	Descripción	Tipo	Ejemplo
Año	Año calendario del reporte	Entero	2021

Enfermedad inicial	Clasificación clínica del dengue	Texto	A90X Dengue sin complicaciones.
Inf. Provincia	Provincia de infección	Texto	GUAYAS
Cantones	Cantón de residencia	Texto	GUAYAQUIL
Parroquias	Parroquia de residencia	Texto	Ayacucho
Semana	Semana epidemiológica (1-52)	Entero	14
Grupo de edad	Rango etario del paciente	Texto	De 20-49 años
Sexo	Sexo del paciente	Texto	Hombre
Condición final	Estado del paciente	Texto	Vivo
Casos	Número de casos agregados	Entero	3
SERO TIPO DENGUE	Serotipo identificado (si aplica)	Texto	DEN-1

3.4.1.1 Procesamiento previo al modelado:

Para construir el dataset de modelado, se aplicaron los siguientes filtros y transformaciones:

1. **Filtro geográfico:** se seleccionaron únicamente los registros de la provincia GUAYAS y el cantón GUAYAQUIL, y se excluyeron las parroquias rurales, manteniendo las 15 parroquias urbanas de interés.
2. **Filtro temporal:** se consideró el periodo 2021–2025, que corresponde al periodo post-pandemia y es el foco del estudio.
3. **Agregación:** los casos se agregaron por semana epidemiológica y parroquia, sumando la variable Casos para obtener el total semanal (TOTAL_CASOS).
4. **Estandarización de nombres:** los nombres de parroquias se normalizaron al formato estándar (primera letra mayúscula, resto minúscula, guión bajo entre palabras) para garantizar la consistencia con los datos climáticos.

3.4.1.2 Limitaciones de la fuente:

- Subnotificación de casos asintomáticos o leves.
- Posibles cambios en la definición de caso durante el periodo.
- La variable SERO TIPO DENGUE presenta alta proporción de valores faltantes, por lo que no fue utilizada en el modelado predictivo.

3.4.2 Fuente ambiental: NASA POWER

Los datos climáticos fueron obtenidos a través de la API del proyecto NASA POWER (Prediction Of Worldwide Energy Resources) (NASA POWER, 2025). Esta fuente fue seleccionada por:

- a) cobertura temporal continua desde 1984,
- b) acceso gratuito sin autenticación,
- c) inclusión de las variables climáticas identificadas como determinantes del dengue, y
- d) consistencia espacial para áreas sin estaciones meteorológicas locales (Sparrow et al., 2020).

Se extrajeron datos diarios para los centroides de cada parroquia urbana mediante un script en Python que realiza solicitudes a la API con un control de tasa de 2 segundos entre peticiones, respetando el límite de 1000 solicitudes por hora establecido por NASA (NASA POWER, 2025).

Tabla 13

Datos: NASA POWER

Parámetro API	Descripción	Unidad	Justificación biológica
T2M_MIN	Temperatura mínima a 2 metros	°C	Crítica para supervivencia del mosquito; por debajo de 15°C se reduce drásticamente la actividad (Mordecai et al., 2017)
T2M_MAX	Temperatura máxima a 2 metros	°C	Temperaturas entre 25°C y 35°C aceleran el desarrollo larval y reducen el periodo

			de incubación extrínseca (Lambrechts et al., 2011)
			Proporciona sitios de cría; lluvias moderadas incrementan criaderos, mientras que lluvias extremas pueden tener efecto de "lavado" (Benedum et al., 2020)
PRECOTCORR	Precipitación acumulada corregida	mm/día	
RH2M	Humedad relativa a 2 metros	%	Humedades >70% prolongan la vida del mosquito adulto y aumentan su frecuencia de picadura (Reiter, 2010)

Nota. Los parámetros fueron seleccionados con base en revisiones sistemáticas de factores climáticos asociados al dengue (Mordecai et al., 2017; Lowe et al., 2021).

3.4.2.1 Periodo de extracción:

1 de enero de 2015 a 31 de diciembre de 2025 (11 años). El rango se extendió más allá del periodo de análisis principal (2021-2025) para permitir el cálculo de rezagos temporales de hasta 4 semanas, siguiendo recomendaciones metodológicas para modelado de enfermedades transmitidas por vectores (Lowe et al., 2021; Hamilton, 2020).

3.4.2.2 Limitaciones de la fuente:

- **Datos de reanálisis, no medición directa:** NASA POWER combina modelos y sensores remotos; los valores no provienen de estaciones meteorológicas en cada parroquia (NASA POWER, 2025).
- **Pérdida de microclimas intra-parroquiales:** el uso del centroide como representación de toda la parroquia no captura variaciones internas (ej. zonas altas vs. bajas, cercanía a cuerpos de agua) (Pfeiffer et al., 2008).
- **Incertidumbre en precipitación:** en zonas urbanas densas como Guayaquil, la estimación por satélite puede subestimar lluvias convectivas locales (Sparrow et al., 2020).

3.4.3 Fuente demográfica: INEC

Para el cálculo de tasas de incidencia, se utilizaron las proyecciones poblacionales por parroquia derivadas de los resultados definitivos del **VIII Censo de Población y VII de Vivienda** (Instituto Nacional de Estadística y Censos [INEC], 2022), complementadas con proyecciones intercensales para los años 2023-2025 según la metodología estándar del INEC (INEC, 2020).

3.5 Descripción del dataset

El dataset construido corresponde a un panel espacio-temporal con estructura de datos tabulares (*tidy data*) (Wickham, 2014). Cada fila representa una observación única definida por la combinación de una parroquia urbana y una semana epidemiológica.

Tabla 14

Descripción de variables del Dataset

Característica	Descripción
Tipo de dataset	Panel balanceado espaciotemporal
Unidad de observación	Parroquia-semana epidemiológica
Cobertura temporal	Semanas 1-2021 a 52-2025 (~260 semanas)
Cobertura espacial	15 parroquias urbanas de Guayaquil
Número total de registros	$15 \times 260 = 3,900$
Forma de integración	Fusión por clave compuesta (PARROQUIA, EPI_YEAR, EPI_WEEK)

Nota. La estructura de panel espacio temporal es estándar en estudios de epidemiología espacial (Pfeiffer et al., 2008; Lawson, 2018).

El dataset contiene 29 columnas, que incluyen:

- **Variables de identificación:** PARROQUIA (nombre de la parroquia), EPI_YEAR (año epidemiológico) y EPI_WEEK (semana epidemiológica).
- **Variables climáticas:** T2M_MIN (temperatura mínima), T2M_MAX (temperatura máxima), PRECOTCORR (precipitación acumulada) y RH2M (humedad relativa), junto con sus respectivos rezagos temporales de 1 a 4 semanas (`_lag_1` a `_lag_4`).

- **Variables epidemiológicas:** TOTAL_CASOS (conteo de casos semanales por parroquia), TASA_INCIDENCIA (tasa de incidencia por 10,000 habitantes) y su rezago de una semana (TASA_INCIDENCIA_lag_1).
- **Variables auxiliares:** LAT (latitud) y LON (longitud) de los centroides parroquiales, y POBLACION (población según censo INEC 2022).

La variable dependiente (TASA_INCIDENCIA) presenta una alta proporción de ceros (68.74%), lo que refleja semanas sin casos reportados en la mayoría de las parroquias. Su distribución es marcadamente asimétrica (coeficiente de asimetría = 11.86), con una media de 0.56 casos por 10,000 habitantes y una desviación estándar de 3.42. El percentil 99 se sitúa en 10.43, mientras que el valor máximo alcanza 68.29, lo que indica que los brotes epidémicos son eventos raros, pero de alta magnitud (Cameron & Trivedi, 2013).

El dataset no presenta valores nulos en ninguna de sus columnas, como resultado del proceso de limpieza descrito en la sección 3.7. Las variables climáticas y sus rezagos fueron calculadas a partir de los datos diarios de NASA POWER, mientras que los casos fueron agregados semanalmente desde los registros del MSP. La integración final se realizó mediante un left join para mantener todas las observaciones climáticas, incluso aquellas sin casos reportados (Lowe et al., 2021).

3.6 Variables del estudio

3.6.1 Variable dependiente (target)

Tasa de incidencia semanal de dengue por parroquia, expresada como casos por cada 10,000 habitantes (WHO, 2022):

$$TI_{i,t} = \frac{Casos_{i,t}}{Población_i} \times 10,000$$

Donde:

- $TI_{i,t}$: tasa de incidencia en la parroquia i durante la semana t
- $Casos_{i,t}$: suma de casos DSSA (dengue sin signos de alarma) y DCSA (dengue con signos de alarma)

- *Población_i*: población de la parroquia según INEC (2022)

Se optó por la tasa en lugar del conteo bruto para eliminar el sesgo poblacional entre parroquias (por ejemplo, Tarqui — 749,392 habitantes— vs. Carbo — 3,062 habitantes—), siguiendo estándares epidemiológicos (WHO, 2022; Rothman et al., 2008). El denominador de 10,000 habitantes (en lugar de 100,000) se seleccionó porque los conteos semanales por parroquia son pequeños, facilitando la interpretación de tasas (Few et al., 2018).

3.6.2 Variables independientes (predictores)

Tabla 15

Definición de variables independientes

Grupo	Variables	Descripción	Unidad	Justificación
Inercia epidemiológica	TASA_INCIDENCIA_lag_1	Tasa de incidencia rezagada 1 semana.	casos/10,000 hab	Captura la persistencia temporal de la enfermedad (Johansson et al., 2016)
Temperatura	T2M_MIN, T2M_MAX + lags 1-4	Temperatura mínima y máxima semanal	°C	Efecto sobre la replicación viral y supervivencia del vector (Mordecai et al., 2017)
Precipitación	PRECTOTCORR + lags 1-4	Precipitación acumulada semanal	mm	Proporciona sitios de cría para <i>Aedes aegypti</i> (Benedum et al., 2020)
Humedad	RH2M + lags 1-4	Humedad relativa semanal promedio	%	Afecta la longevidad del mosquito adulto (Reiter, 2010)

Espaciales	LAT, LON	Coordenadas del centroide parroquial	grados decimales	Captura autocorrelación espacial (Tobler, 1970)
Temporales	EPI_YEAR, EPI_WEEK	Identificadores de tiempo	número	Controla estacionalidad y tendencia (Lowe et al., 2021)

Nota sobre rezagos: se incluyeron retardos de 1 a 4 semanas para capturar el periodo biológico entre condiciones ambientales favorables y el reporte de casos: eclosión de huevos (1-2 días) → desarrollo larvario (5-7 días) → emergencia de adultos (2-3 días) → periodo de incubación extrínseca del virus (8-12 días) → síntomas en humanos (3-7 días) → notificación (2-5 días) (Lambrechts et al., 2011; Hamilton, 2020).

3.7 Preprocesamiento e integración de datos

El flujo de preprocesamiento constó de las siguientes etapas, siguiendo las buenas prácticas recomendadas para estudios de machine learning en epidemiología (Kuhn & Johnson, 2019; Pichler et al., 2023):

1. **Extracción de datos climáticos:** mediante script en Python que consulta la API de NASA POWER para cada centroide parroquial, con pausa de 2 segundos entre solicitudes para respetar límites de tasa (NASA POWER, 2025).
2. **Transformación a semanas epidemiológicas:** conversión de fechas calendario a semanas epidemiológicas según estándar OPS (domingo a sábado), mediante desplazamiento de +1 día y uso del método `isocalendar()` de pandas (McKinney, 2010).
3. **Agregación semanal:** para variables climáticas: T2M_MIN → mínimo semanal, T2M_MAX → máximo semanal, PRECTOTCORR → suma semanal, RH2M → promedio semanal. Estas operaciones de agregación son estándar en estudios de series temporales climáticas (Hipel & McLeod, 1994).
4. **Procesamiento de datos epidemiológicos:** Los datos de casos fueron extraídos del archivo `Dengue_Guayaquil_2016_2026_MSP.csv`, proporcionado por el Ministerio de Salud

Pública del Ecuador. Este archivo contiene registros individuales agregados por variables demográficas y geográficas, con una estructura de una sola tabla. El procesamiento incluyó los siguientes pasos:

- Filtro de casos del cantón Guayaquil y de las 15 parroquias urbanas de interés.
 - Agregación de casos por semana epidemiológica y parroquia, sumando la variable Casos para obtener el total semanal (TOTAL_CASOS). Dado que el archivo del MSP ya se encuentra desagregado a nivel de caso individual (con variables como Grupo de edad, Sexo y SERO TIPO DENGUE), no fue necesario procesar hojas separadas.
 - Los valores faltantes (semanas sin casos reportados) se rellenaron con 0, siguiendo el principio de que la ausencia de registro se interpreta como cero casos (Johansson et al., 2016).
5. **Cálculo de rezagos temporales (lags):** creación de variables con desplazamiento de 1 a 4 semanas para cada predictor climático y para la tasa de incidencia. La operación se realiza por grupo de parroquia (`groupby('PARROQUIA').shift(lag)`) para evitar mezcla entre unidades espaciales, siguiendo las recomendaciones de Hyndman y Athanasopoulos (2018). Cabe señalar que los rezagos climáticos fueron calculados en el dataset climático (`dataset_guayaquil_epiweeks_2015_2026.csv`), por lo que este paso se aplicó únicamente a la variable `TASA_INCIDENCIA_lag_1`.
 6. **Integración final:** fusión de datasets climático y epidemiológico mediante clave compuesta (`PARROQUIA`, `EPI_YEAR`, `EPI_WEEK`) utilizando un `left join` para mantener todas las observaciones climáticas (Lowe et al., 2021).
 7. **Limpieza y filtrado temporal:** se eliminaron las filas correspondientes a los años 2015-2020, quedándose únicamente con el periodo 2021-2025, que es el foco del estudio. El dataset resultante contiene **3,900 registros** (15 parroquias × 5 años × 52 semanas), sin valores nulos en las variables de interés.

3.8 Análisis exploratorio de datos

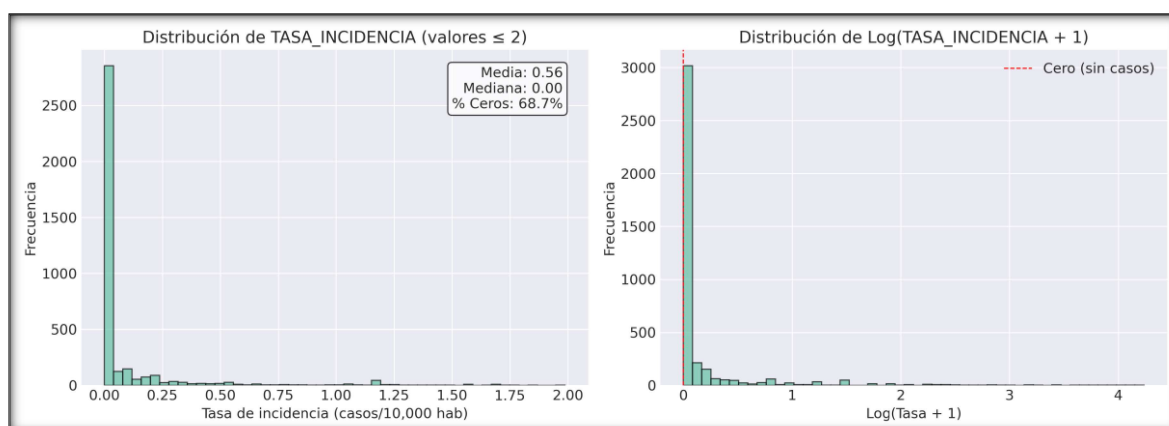
El análisis exploratorio de datos (EDA) se realizó con el objetivo de comprender la estructura, calidad y relaciones subyacentes del dataset antes de proceder con el modelado predictivo (Tukey, 1977; Beaujean, 2013). Este análisis permitió identificar patrones temporales, diferencias espaciales y la naturaleza de las relaciones entre las variables climáticas y la incidencia del dengue, lo que fundamentó la selección de los modelos evaluados posteriormente.

3.8.1 Distribución de la variable dependiente

La Figura 2 presenta la distribución de la variable TASA_INCIDENCIA para el periodo 2021-2025.

Figura 2

Distribución de la tasa de incidencia (casos por 10,000 habitantes).



(a) Histograma para valores ≤ 2 ; (b) Histograma de la transformación logarítmica ($\log_{1p}$).

La variable dependiente presenta una alta concentración de ceros (68.74%), lo que refleja semanas epidemiológicas sin casos notificados en la mayoría de las parroquias. Este fenómeno es común en enfermedades de baja endemicidad o con transmisión estacional (Cameron & Trivedi, 2013). La media (0.5599) es considerablemente mayor que la mediana (0.00), y el coeficiente de asimetría ($\text{skewness} = 11.86$) indica una cola derecha extremadamente larga, correspondiente a brotes epidémicos de alta magnitud (Groeneveld & Meeden, 1984).

La transformación logarítmica $\log(\text{TASA_INCIDENCIA} + 1)$ (Figura 2b) muestra una distribución bimodal, con un pico masivo en cero y una segunda moda alrededor de valores positivos bajos. Esta estructura sugiere que los datos no provienen de una distribución normal, lo que invalida los supuestos de los modelos lineales estándar (Tabachnick & Fidell, 2019).

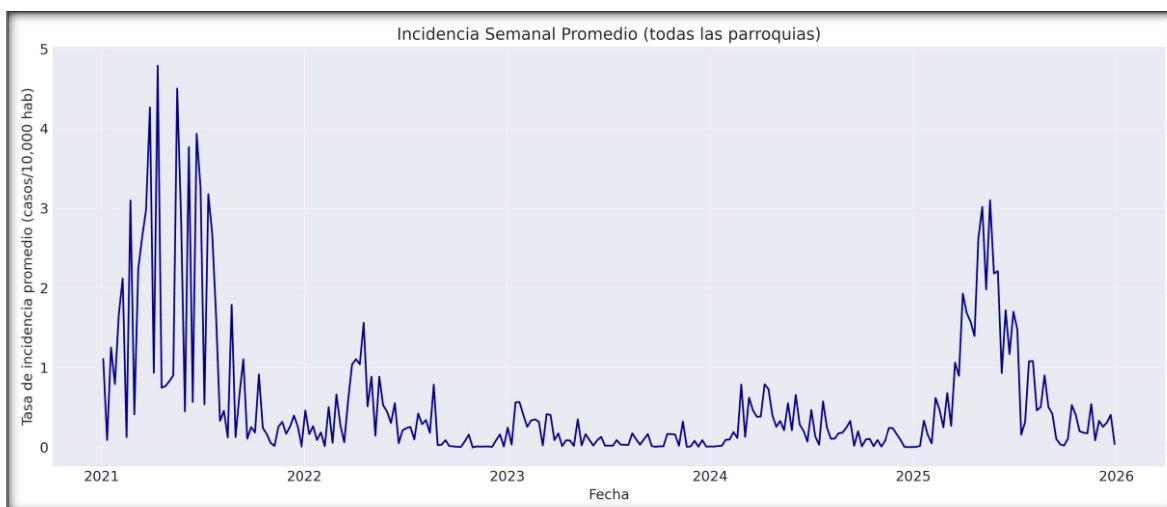
Implicación metodológica: El exceso de ceros y la asimetría extrema justifican la exploración de modelos robustos que no asuman normalidad, como Random Forest (Breiman, 2001), y la consideración de enfoques de regularización para modelos lineales (Tibshirani, 1996).

3.8.2 Patrones temporales

La Figura 3 muestra la evolución de la incidencia semanal promedio (todas las parroquias) durante el periodo 2021-2025.

Figura 3

Incidencia semanal promedio de dengue en Guayaquil (2021-2025).



Se observa un patrón estacional recurrente, con picos de transmisión concentrados en el primer semestre de cada año (semanas 10 a 25), coincidiendo con la temporada de lluvias en la región costera de Ecuador. Este comportamiento es consistente con lo documentado en la literatura para ciudades tropicales de América Latina (Lowe et al., 2021; Johansson et al., 2016).

Los años 2021 y 2022 muestran picos más pronunciados, posiblemente asociados a condiciones climáticas favorables (fenómeno La Niña) o a la acumulación de susceptibles tras la

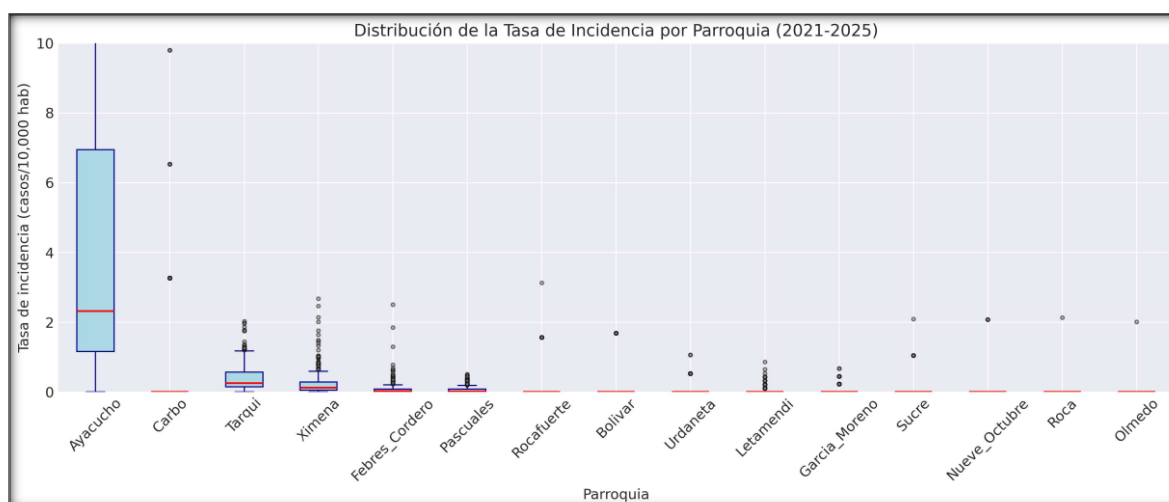
pandemia de COVID-19 (Chen et al., 2022). El año 2025, que se utilizó como conjunto de prueba en el modelado, presenta un patrón estacional similar a los años anteriores, lo que valida la estrategia de división temporal empleada.

Implicación metodológica: La estacionalidad identificada refuerza la necesidad de incluir variables climáticas y sus rezagos temporales como predictores en los modelos (Mordecai et al., 2017).

3.8.3 Variabilidad espacial entre parroquias

Las Figuras 4 y 5 muestran la distribución de la incidencia por parroquia y la incidencia acumulada total en el periodo 2021-2025.

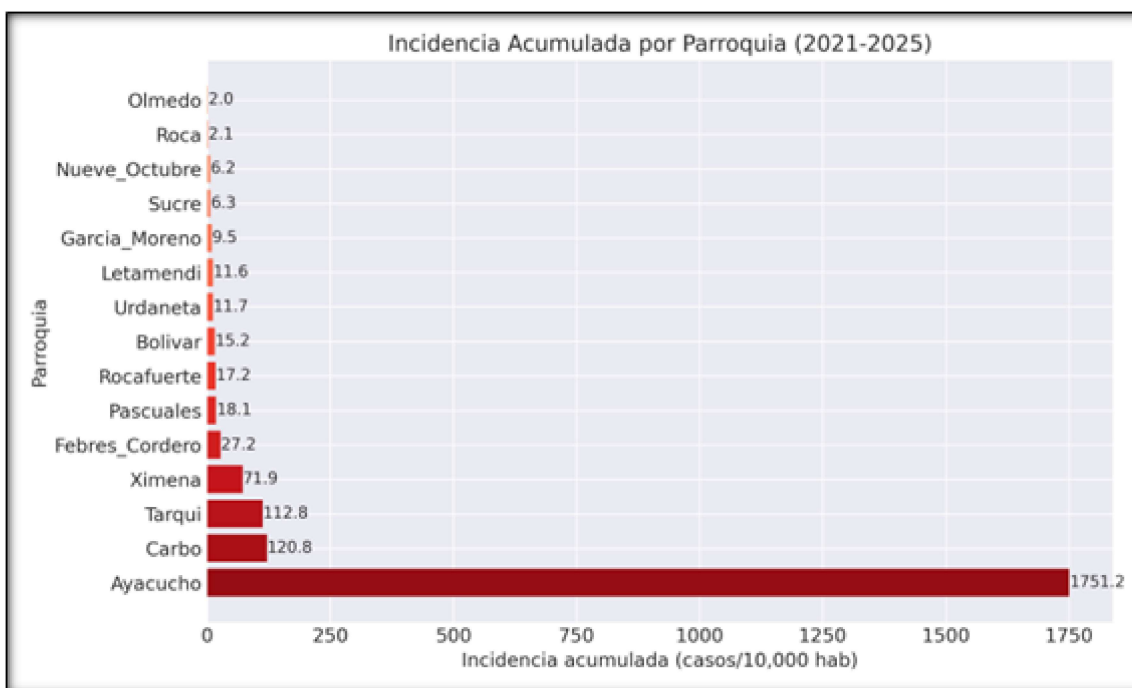
Figura 4
Distribución de la tasa de incidencia por parroquia (boxplot).



La Figura 4 revela una heterogeneidad espacial extrema. La parroquia Ayacucho presenta una tasa media de incidencia de 6.74 casos por 10,000 habitantes, muy superior al resto de parroquias, cuyas tasas oscilan entre 0.1 y 0.3. Esta diferencia sugiere que Ayacucho experimenta una carga de enfermedad persistentemente más alta que el resto del cantón.

Figura 5

Incidencia acumulada por parroquia (2021-2025).



La Figura 5 confirma esta disparidad de manera aún más contundente: Ayacucho acumula 1,751.2 casos por 10,000 habitantes en cinco años, una cifra que es 14 veces mayor que la segunda parroquia con mayor incidencia acumulada (Carbo, 120.8) y 875 veces mayor que la más baja (Olmedo, 2.0). Este patrón no se explica únicamente por el tamaño poblacional (Ayacucho tiene 8,640 habitantes), sino que sugiere la existencia de factores locales específicos —como condiciones de saneamiento, densidad de criaderos o brotes históricos— que no están capturados en las variables climáticas y que ameritan una investigación específica (Mulligan et al., 2020).

En un segundo grupo, parroquias como Carbo (120.8) y Tarqui (112.8) presentan incidencias acumuladas moderadamente altas, mientras que el resto de las parroquias (Ximena, Febres Cordero, Pascuales, etc.) se sitúan por debajo de 30. Esta distribución sugiere que la transmisión del dengue en Guayaquil no es homogénea y que existen focos persistentes de alta transmisión, siendo Ayacucho el caso más extremo.

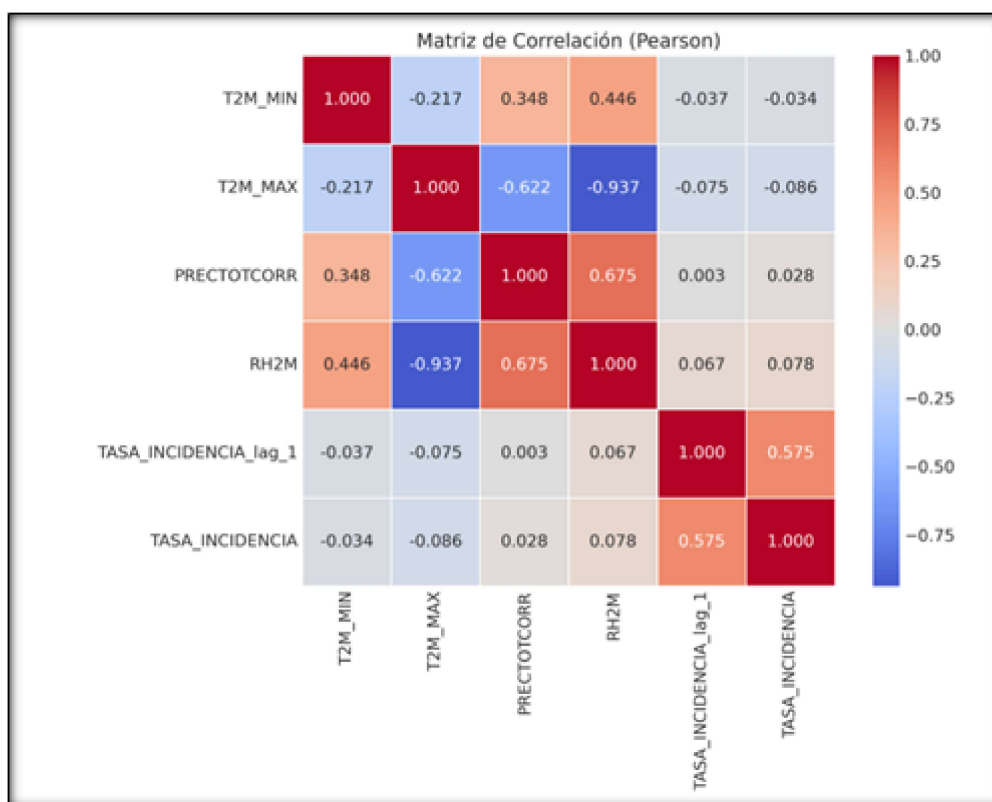
Implicación metodológica: La variabilidad espacial extrema justifica la inclusión de coordenadas geográficas y el análisis de residuales por parroquia para identificar zonas de riesgo no capturadas por las variables climáticas (Pfeiffer et al., 2008). Asimismo, sugiere que factores locales no incluidos en el modelo (ej. saneamiento, densidad poblacional, criaderos) podrían mejorar la predicción en parroquias como Ayacucho.

3.8.4 Correlaciones entre variables

La Figura 6 presenta la matriz de correlaciones de Pearson entre las variables climáticas y la variable dependiente.

Figura 6

Matriz de correlaciones (Pearson) entre variables climáticas y tasa de incidencia.



Las correlaciones lineales entre las variables climáticas y la incidencia son débiles ($|r| < 0.07$), lo que sugiere que el efecto del clima sobre el dengue no es estrictamente lineal en el rango de datos observados. Este hallazgo es consistente con la literatura que documenta relaciones no

lineales (efectos umbral, interacciones) entre temperatura, precipitación y transmisión del dengue (Mordecai et al., 2017).

Las variables climáticas presentan correlaciones moderadas entre sí (ej. T2M_MIN con RH2M), lo que sugiere cierta multicolinealidad que podría afectar a los modelos lineales, pero que no debería ser problemática para modelos basados en árboles (Hastie et al., 2009).

Implicación metodológica: La debilidad de las correlaciones lineales y la presencia de relaciones no lineales justifican la exploración de modelos como Random Forest, que capturan interacciones y umbrales sin requerir especificación funcional previa (Breiman, 2001), y la consideración de regularización (Lasso, Ridge) para manejar la multicolinealidad en modelos lineales (Tibshirani, 1996; Zou & Hastie, 2005).

3.8.4.1 Síntesis del Análisis Exploratorio

El análisis exploratorio reveló tres características fundamentales que condicionaron el diseño metodológico del estudio:

Tabla 16
Resultados de análisis exploratorio.

Hallazgo	Implicación metodológica
Alta proporción de ceros (68.74%)	Se requieren modelos robustos a la asimetría y al exceso de ceros.
Estacionalidad marcada	Se deben incluir variables climáticas y sus rezagos temporales.
Heterogeneidad espacial	El análisis geoespacial y la evaluación por parroquia son necesarios.
Relaciones no lineales	Modelos lineales requieren regularización; modelos basados en árboles son prioritarios.

Nota: Estos hallazgos guiaron la selección de los modelos evaluados en la sección 3.9 y la estrategia de validación temporal definida en 3.9.1.

3.9 Evaluación y selección de modelos predictivos

3.9.1 Evaluación de la linealidad y selección de modelos

Para determinar la naturaleza de la relación entre los predictores y la variable objetivo, se aplicó la prueba de Ramsey RESET (Ramsey, 1969) sobre el modelo OLS entrenado con los datos de 2021-2024. Esta prueba evalúa si la inclusión de términos polinomiales (potencias de los valores ajustados) mejora significativamente el modelo. El resultado ($F = 118.82$, $p < 0.001$) rechazó la hipótesis nula de especificación lineal, confirmando que la relación no es puramente lineal.

Este hallazgo justificó la exploración de dos familias de modelos:

1. **Modelos lineales regularizados** (Ridge, Lasso, ElasticNet), que, aunque mantienen la estructura lineal, pueden manejar la multicolinealidad y realizar selección de variables.
2. **Modelos basados en árboles** (Random Forest, XGBoost), que no asumen linealidad y pueden capturar interacciones y umbrales complejos (Breiman, 2001; Chen & Guestrin, 2016).

La comparación de ambas familias se presenta en el Capítulo 4.

3.9.2 Estrategia de validación

Dado que los datos corresponden a una serie temporal con dependencia secuencial, no se empleó validación cruzada aleatoria estándar, ya que esta rompería la estructura temporal y generaría fuga de información (Bergmeir & Benítez, 2012; Tashman, 2000). En su lugar, se implementó una estrategia de validación con punto de corte fijo que preserva el orden cronológico (Hyndman & Athanasopoulos, 2018), en la cual el modelo se entrena con datos pasados y se evalúa en datos futuros no vistos durante el entrenamiento.

La división temporal quedó establecida como nos indica la siguiente tabla:

Tabla 17*División de datos - Train & Test*

Conjunto	Periodo	Años	Registros	Proporción
Entrenamiento	2021 – 2024	4 años completos	3,120	80%
Prueba (test)	2025	1 año completo	780	20%

Esta estrategia presenta dos ventajas metodológicas fundamentales para la predicción epidemiológica del dengue (Lowe et al., 2021; Johansson et al., 2016):

1. **Alineación estacional:** al entrenar con 2021-2024 y probar en 2025, se preserva la correspondencia estacional (enero con enero, julio con julio), lo que es crítico para una enfermedad con fuerte estacionalidad.
2. **Horizonte de predicción limitado:** el horizonte máximo de evaluación es de 52 semanas, lo que resulta realista para modelos que dependen de rezagos de 1 a 4 semanas (Hamilton, 2020).

3.9.3 Modelos evaluados

Se seleccionaron seis modelos que abarcan desde enfoques lineales tradicionales hasta técnicas de aprendizaje automático basadas en árboles, con el fin de evaluar la ganancia predictiva de cada aproximación metodológica.

Tabla 18*Evaluación de modelos y justificación*

Familia	Modelo	Justificación	Referencia clave
Lineales	OLS (Mínimos Cuadrados Ordinarios)	Baseline mínimo. Modelo de referencia para cuantificar la mejora de enfoques más complejos.	Hastie et al. (2009)
	Ridge (Regularización L2)	Penaliza la magnitud de los coeficientes para reducir la varianza en presencia de multicolinealidad.	Hoerl & Kennard (1970)
	Lasso (Regularización L1)	Realiza selección de variables al reducir algunos coeficientes exactamente a cero, facilitando la interpretación.	Tibshirani (1996)

Familia	Modelo	Justificación	Referencia clave
Basados en árboles	ElasticNet (L1 + L2)	Combina las ventajas de Ridge y Lasso, útil cuando hay grupos de variables correlacionadas.	Zou & Hastie (2005)
	Random Forest	Conjunto de árboles de decisión con agregación bootstrap (<i>bagging</i>). Captura relaciones no lineales e interacciones de forma natural.	Breiman (2001)
	XGBoost	Implementación eficiente de <i>gradient boosting</i> que construye árboles de forma secuencial corrigiendo errores previos.	Chen & Guestrin (2016)

Calibración del modelo Random Forest: Para el modelo Random Forest no se aplicó calibración por post-procesamiento sobre el conjunto de prueba, con el fin de evitar fuga de información y garantizar que las métricas de desempeño reflejen la capacidad predictiva real del modelo en datos no vistos (Kuhn & Johnson, 2019). El modelo se reporta en su versión original entrenada con los datos de 2021–2024 y evaluada en 2025.

3.9.4 Métricas de evaluación

Para evaluar y comparar el desempeño de los modelos, se emplearon tres métricas ampliamente utilizadas en problemas de regresión y pronóstico (Willmott & Matsuura, 2005; Chai & Draxler, 2014; Nagelkerke, 1991):

Tabla 19
Evaluación de métricas

Métrica	Fórmula	Interpretación
MAE (Mean Absolute Error)	$MAE = \frac{1}{n} \sum_{i=1}^n y_i - \hat{y}_i$	Error absoluto promedio. Robusto ante outliers. Misma unidad que la variable objetivo.

Métrica	Fórmula	Interpretación
RMSE (Root Mean Square Error)	$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$	Penaliza errores grandes. Útil cuando los errores graves son especialmente importantes, como en la predicción de brotes epidémicos.
R ² (Coeficiente de determinación)	$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$	Proporción de varianza explicada por el modelo. Rango $(-\infty, 1]$; valores cercanos a 1 indican mejor ajuste.

Las comparaciones entre modelos se realizaron mediante la inspección directa de las métricas MAE, RMSE y R², siguiendo el enfoque estándar en estudios de predicción epidemiológica (Lowe et al., 2021; Zhao et al., 2020). Dado que el objetivo principal es identificar el modelo con mejor desempeño predictivo para su aplicación en vigilancia epidemiológica, se priorizó la minimización del RMSE por su sensibilidad a errores grandes (brotes epidémicos), complementado con el MAE para evaluar el error promedio y el R² como medida de varianza explicada.

3.9.5 Procedimiento de optimización de hiperparámetros

Para los modelos que requieren ajuste de hiperparámetros (Ridge, Lasso, ElasticNet, Random Forest y XGBoost), se implementó una búsqueda en grilla (Grid Search) combinada con validación cruzada temporal (TimeSeriesSplit) con 3 pliegues (Bergstra & Bengio, 2012). Este enfoque respeta el orden cronológico de los datos dentro del conjunto de entrenamiento, asegurando que cada pliegue de validación sea posterior al pliegue de entrenamiento.

Los rangos de búsqueda para cada modelo se definieron con base en la literatura y en pruebas exploratorias iniciales:

Tabla 20
Definición de modelos

Modelo	Hiperparámetros explorados
Ridge / Lasso / ElasticNet	alpha: [0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1.0, 5.0, 10.0]; l1_ratio (solo ElasticNet): [0.1, 0.3, 0.5, 0.7, 0.9, 1.0]
Random Forest	n_estimators: [100, 200, 300]; max_depth: [10, 20, None]; min_samples_split: [2, 5, 10]; min_samples_leaf: [1, 2, 4]
XGBoost	n_estimators: [100, 200]; max_depth: [4, 6]; learning_rate: [0.01, 0.05]; subsample: [0.8]; colsample_bytree: [0.8]; reg_alpha: [0.01, 0.1]; reg_lambda: [1.0]

La métrica de optimización fue el MAE negativo (neg_mean_absolute_error), priorizando la minimización del error absoluto promedio durante la selección de hiperparámetros.

Una vez seleccionados los mejores hiperparámetros, el modelo fue reentrenado con todo el conjunto de entrenamiento (2021-2024) y evaluado en el conjunto de prueba (2025). En el caso particular de Random Forest, los hiperparámetros finales seleccionados fueron:

n_estimators=200, max_depth=None, min_samples_split=5 y min_samples_leaf=2. No se aplicó calibración post-entrenamiento sobre el conjunto de prueba para evitar fuga de información y garantizar que las métricas reflejen la capacidad predictiva real del modelo en datos no vistos (Kuhn & Johnson, 2019). Los resultados de esta comparación se presentan en el Capítulo 4.

3.9.6 Criterio de selección del modelo final

El modelo final se seleccionó con base en un criterio de balance entre capacidad predictiva y parsimonia (Hand, 2006; Shmueli, 2010). Se priorizó el RMSE sobre el MAE, dado que penaliza errores grandes, lo cual es crítico en contextos epidemiológicos donde la subestimación

de un brote puede tener consecuencias sanitarias graves (Lowe et al., 2021). Adicionalmente, se consideró la interpretabilidad del modelo y su estabilidad frente a cambios en los datos. Los resultados de la comparación se presentan en el Capítulo 4.

3.10 Técnicas de análisis espacial

Para cumplir con el objetivo de identificar patrones territoriales y zonas de riesgo (Tobler, 1970; Pfeiffer et al., 2008), se aplicarán las siguientes técnicas de análisis espacial, utilizando los Sistemas de Información Geográfica (SIG) y librerías especializadas en Python (Bivand et al., 2013)

Tabla 21

Técnicas aplicadas para el análisis espacial

Técnica	Herramienta / Método	Propósito	Referencia
Mapas de incidencia observada	QGIS / geopandas	Visualizar la distribución territorial de la tasa de incidencia promedio por parroquia.	Bivand et al. (2013)
Autocorrelación espacial global	Índice de Moran (I) con matriz de pesos espaciales tipo Queen	Evaluar si la incidencia del dengue se distribuye aleatoriamente o presenta agrupamiento espacial significativo.	Moran (1950); Cliff & Ord (1981)
Detección de hotspots	Estadístico Getis-Ord Gi*	Identificar parroquias con valores significativamente altos (<i>hotspots</i>) y bajos (<i>coldspots</i>) de incidencia.	Getis & Ord (1992); Ord & Getis (1995)
Mapas de residuales	Diferencia entre incidencia observada y predicha por parroquia	Identificar áreas donde el modelo subestima o sobreestima sistemáticamente el riesgo.	Lawson (2018)

Técnica	Herramienta / Método	Propósito	Referencia
Análisis de riesgo predicho	Clasificación por cuantiles de la incidencia predicha promedio	Generar mapas de zonas de riesgo (bajo, medio, alto, muy alto).	Pfeiffer et al. (2008)

Flujo de trabajo espacial (basado en Bivand et al., 2013 y Lawson, 2018):

1. Construcción de un geodataframe con los polígonos de las 15 parroquias y sus atributos (incidencia observada promedio, incidencia predicha promedio, residuales).
2. Cálculo del índice de Moran global y evaluación de su significación mediante permutaciones (999 simulaciones) (Cliff & Ord, 1981).
3. Identificación de hotspots con Getis-Ord Gi* usando un umbral de significación $p < 0.05$.
4. Generación de mapas temáticos (coropletas) con los valores de incidencia y residuales.
5. Interpretación de los resultados espaciales en conjunto con la importancia de variables del modelo Random Forest.

3.11 Herramientas tecnológicas

La implementación completa del estudio se apoyó en herramientas de software libre y de código abierto, garantizando reproducibilidad y acceso sin restricciones (Sandve et al., 2013; Wilson et al., 2014):

Tabla 22
Frameworks y Herramientas utilizadas

Propósito	Herramientas / Librerías	Referencia
Procesamiento y análisis de datos	Python 3.10+, pandas, numpy	McKinney (2010); van der Walt et al. (2011).
Conexión a APIs	requests, json	Reitz (2016)

Propósito	Herramientas / Librerías	Referencia
Preprocesamiento	scikit-learn (StandardScaler, Pipeline)	Pedregosa et al. (2011).
Modelos baseline	scikit-learn (LinearRegression, Ridge, Lasso, ElasticNet)	Pedregosa et al. (2011) .
Modelos de ML	scikit-learn (RandomForestRegressor), xgboost (XGBRegressor)	Breiman (2001b); Chen & Guestrin (2016).
Validación y métricas	scikit-learn (mean_absolute_error, mean_squared_error, r2_score)	Pedregosa et al. (2011).
Análisis espacial	geopandas, pysal (esda, libpysal)	Bivand et al. (2013); Rey & Anselin (2007).
Visualización	matplotlib, seaborn	Hunter (2007); Waskom (2021).
Entorno de desarrollo	Jupyter Notebook / VS Code	Kluyver et al. (2016)

3.12 Consideraciones éticas y limitaciones metodológicas

3.12.1 Consideraciones éticas

El estudio se rige por los principios éticos para investigación con datos secundarios establecidos en la Declaración de Helsinki (World Medical Association, 2013) y las normativas ecuatorianas aplicables (Asamblea Nacional del Ecuador, 2014):

- **Uso de datos secundarios anonimizados:** los registros epidemiológicos del MSP contenidos en el archivo Dengue_Guayaquil_2016_2026_MSP.csv no contienen identificadores de pacientes (nombres, cédulas, direcciones específicas), sino únicamente

agregados por parroquia, semana y variables demográficas (edad, sexo). No existe posibilidad de reidentificación de individuos (MSP, 2025).

- **Ausencia de intervención en sujetos humanos:** el estudio no realiza encuestas, entrevistas, experimentos clínicos ni manipulaciones ambientales. Se limita al análisis retrospectivo de datos existentes (CIOMS, 2016).
- **Confidencialidad institucional:** los resultados se presentan a nivel agregado (parroquias), sin exponer información sensible de establecimientos de salud o casos individuales (WHO, 2022).
- **Cumplimiento normativo:** la investigación se acoge a los principios de la Declaración de Helsinki para estudios con datos secundarios y no requiere aprobación de comité de ética en investigación, según la normativa ecuatoriana para estudios que no involucran sujetos humanos ni datos sensibles identificables (Asamblea Nacional del Ecuador, 2014).

3.12.2 Limitaciones metodológicas

Tabla 23

Limitaciones metodológicas

Limitación	Descripción	Implicación	Mitigación
Resolución espacial	La unidad de análisis es la parroquia (agregación gruesa).	Los predictores climáticos se asignan homogéneamente dentro de la parroquia, pero el microclima puede variar.	Se reconoce explícitamente; los resultados no deben interpretarse como predicciones a nivel de barrio o vivienda. Se sugiere como línea de investigación futura el uso de cuadrículas de mayor resolución.
	Existen variaciones intra-parroquiales de riesgo (ej. barrios, manzanas) no capturadas.	Las intervenciones de salud pública suelen requerir mayor resolución.	
Subnotificación de casos	Los registros del MSP capturan solo casos que acuden al sistema de salud. Asintomáticos y casos	La variable dependiente (TASA_INCIDENCIA) subestima la incidencia real, especialmente en poblaciones con menor	Se asume que la subnotificación es sistemática (no varía aleatoriamente en el tiempo) y afecta por

Limitación	Descripción	Implicación	Mitigación
Datos climáticos de reanálisis	leves no están representados.	acceso a salud o en zonas rurales (excluidas del estudio).	igual a todas las parroquias urbanas incluidas en el modelo (Brady et al., 2019; WHO, 2022).
	NASA POWER proporciona estimaciones modeladas, no mediciones directas en cada parroquia. La resolución espacial de la malla subyacente es de aproximadamente 50 km.	Puede haber errores sistemáticos en zonas con microclimas complejos, especialmente en precipitación (lluvias convectivas locales).	Se prefiere esta fuente por su cobertura temporal larga, consistencia y disponibilidad gratuita. Los errores, al ser sistemáticos, afectan por igual a todos los modelos comparados (Sparrow et al., 2020). Estas variables se utilizan en el análisis descriptivo del Capítulo 4 para caracterizar la población afectada, pero no en la predicción. Se sugiere como línea de investigación futura el modelado multinivel o la inclusión de estas variables si se obtuvieran datos agregados por grupo etario (Mulligan et al., 2020).
	Variables como Grupo de edad, Sexo y SERO TIPO DENGUE no fueron incluidas como predictores en el modelo debido a la unidad de análisis (parroquia-semana) y a la alta proporción de valores faltantes en SERO TIPO DENGUE.	El modelo no captura diferencias etarias, sesgos por sexo o variaciones en la circulación de serotipos que podrían influir en la transmisión.	
Exclusión de variables demográficas			
Exclusión de variables socioeconómicas	No se dispone de datos desagregados por parroquia sobre hacinamiento, nivel educativo, acceso a	El modelo puede perder capacidad predictiva en zonas donde estos factores sean determinantes del riesgo (ej. Ayacucho, que	Se reconoce como una limitación importante. El análisis de residuales espaciales (Capítulo 4) identifica zonas de subestimación que

Limitación	Descripción	Implicación	Mitigación
Periodo de modelado limitado (2021-2025)	agua potable, ingreso, etc.	presenta la mayor incidencia acumulada).	podrían estar asociadas a estos factores, orientando futuras investigaciones (Mulligan et al., 2020).
	Aunque los datos están disponibles desde 2016, el modelado se restringió al periodo post-pandemia (2021-2025) para evitar el ruido introducido por la COVID-19 en la notificación de casos y en los patrones de movilidad.	El modelo no se entrena con datos pre-pandemia, lo que podría limitar su capacidad para capturar patrones históricos o su generalización a periodos con dinámicas diferentes.	Se justifica por la homogeneidad epidemiológica del periodo post-pandemia. Se sugiere validar el modelo con datos de 2026 (cuando estén completos) o con datos históricos (2016-2019) como extensión del estudio (Chen et al., 2022; Lorenz et al., 2023).
Generalización externa	El modelo se ajusta específicamente a las 15 parroquias urbanas de Guayaquil. No se recomienda su uso directo en otras ciudades o en zonas rurales sin recalibración.	Los coeficientes y las importancias de variables son contextuales (ej. la importancia de T2M_MAX_lag_2 en Guayaquil puede no aplicarse en Quito o en zonas amazónicas).	Se enfatiza que el aporte principal es metodológico: la estrategia de comparación de modelos (OLS vs Lasso vs RF) y la identificación de Ayacucho como <i>hotspot</i> son transferibles, pero no los coeficientes numéricos (Hand, 2006).

Nota: Estas limitaciones no invalidan los hallazgos del estudio, pero deben ser consideradas al interpretar los resultados y al formular recomendaciones para la vigilancia epidemiológica. Se han identificado líneas de investigación futura que podrían abordar estas limitaciones: (i) incorporación de variables socioeconómicas y demográficas; (ii) evaluación del modelo con datos de 2026 o de años anteriores; y (iii) extensión del enfoque a otras ciudades ecuatorianas para evaluar su generalizabilidad.

3.13 Síntesis del Capítulo 3

El presente capítulo ha definido el enfoque cuantitativo, el nivel correlacional-predictivo y el diseño no experimental longitudinal retrospectivo que guían la investigación (Hernández-Sampieri & Mendoza, 2018; Shadish et al., 2002). Se ha delimitado el ámbito espacial a las 15 parroquias urbanas de Guayaquil, justificando la exclusión de las rurales por razones de homogeneidad epidemiológica, consistencia en la notificación de casos y coherencia climática. El periodo de análisis comprende los años 2021 a 2025, con una frecuencia semanal.

Las fuentes de datos incluyen el archivo unificado del Ministerio de Salud Pública (Dengue_Guayaquil_2016_2026_MSP.csv), que contiene registros agregados por semana, parroquia y variables demográficas; los datos climáticos de NASA POWER (temperatura, precipitación y humedad); y las proyecciones poblacionales del INEC 2022. El preprocesamiento consistió en el filtrado geográfico y temporal, la agregación de casos por semana y parroquia, el cálculo de la tasa de incidencia (casos por 10,000 habitantes), y la creación de rezagos temporales de 1 a 4 semanas para variables climáticas y de inercia epidemiológica.

El análisis exploratorio de datos reveló que la variable dependiente presenta una alta proporción de ceros (68.74%) y una asimetría extrema ($skewness = 11.86$), características que condicionan la selección de modelos. Se identificó una marcada estacionalidad con picos en el primer semestre de cada año, y una heterogeneidad espacial muy pronunciada, donde la parroquia Ayacucho destaca como un *outlier* extremo (incidencia acumulada 1,751.2 casos por 10,000 habitantes, 14 veces superior a la segunda parroquia).

En la fase de modelado comparativo, se evaluaron seis modelos: OLS, Ridge, Lasso, ElasticNet, Random Forest y XGBoost. La validación se realizó con una división temporal fija (entrenamiento: 2021-2024; prueba: 2025) y se utilizaron las métricas MAE, RMSE y R^2 . Los modelos lineales regularizados fueron optimizados mediante GridSearch con validación temporal

(TimeSeriesSplit), mientras que los modelos basados en árboles se ajustaron con búsqueda de hiperparámetros sobre los mismos pliegues temporales.

Los resultados de la comparación (presentados en detalle en el Capítulo 4) indicaron que Random Forest fue el modelo con mejor desempeño global en términos de RMSE y R^2 , superando a todos los modelos lineales y a XGBoost. Lasso, aunque con menor capacidad predictiva, aportó valor en la selección de variables, identificando que solo dos predictores (TASA_INCIDENCIA_lag_1 y T2M_MAX_lag_2) son relevantes en el contexto lineal. Ridge y ElasticNet fueron descartados por no superar a OLS, y XGBoost no mostró diferencias significativas con Random Forest ($p=0.2662$).

Las técnicas de análisis espacial definidas incluyen el índice de Moran, la detección de hotspots mediante Getis-Ord G_i^* , la generación de mapas de residuales para el modelo final (Random Forest), y el análisis de riesgo predicho mediante clasificación por cuantiles de la incidencia pronosticada. Los resultados de estas técnicas se presentan en el Capítulo 4.

Finalmente, se han detallado las herramientas tecnológicas (Python, pandas, scikit-learn, geopandas) y se han reconocido las limitaciones éticas y metodológicas, incluyendo la subnotificación de casos, la resolución espacial gruesa y la exclusión de variables socioeconómicas y demográficas. Estas limitaciones no invalidan los hallazgos, pero deben ser consideradas al interpretar los resultados y al formular recomendaciones para la vigilancia epidemiológica.

El capítulo ha establecido las bases metodológicas para la predicción espacial del dengue en Guayaquil, asegurando que el estudio es metodológicamente coherente, técnicamente viable, reproducible y evaluable.

CAPÍTULO 4:

4 RESULTADOS

En este capítulo se presentan los resultados del análisis de datos y la evaluación de los modelos predictivos para la incidencia del dengue en Guayaquil. La estructura del capítulo sigue el orden metodológico establecido en el Capítulo 3: primero se caracteriza el dataset final utilizado en el modelado, luego se presentan los resultados de los modelos lineales (OLS, Ridge, Lasso y ElasticNet), posteriormente los modelos basados en árboles (Random Forest y XGBoost), y finalmente se compara el desempeño global para seleccionar el modelo óptimo. Se incluyen además análisis de residuales, importancia de variables y un mapa de riesgo residual para el modelo final.

4.1 Caracterización del dataset de modelado

El dataset final utilizado para el modelado predictivo está compuesto por 3,900 registros, correspondientes a 15 parroquias urbanas de Guayaquil, con una cobertura temporal de 5 años completos (2021–2025) y una frecuencia semanal. Cada registro contiene variables climáticas (temperatura mínima, máxima, precipitación y humedad), sus rezagos de 1 a 4 semanas, la tasa de incidencia de la semana anterior (TASA_INCIDENCIA_lag_1) y la variable objetivo (TASA_INCIDENCIA).

La variable objetivo presenta una alta proporción de ceros (68.74%), lo que refleja semanas sin casos reportados en la mayoría de las parroquias. Su distribución es marcadamente asimétrica (coeficiente de asimetría = 11.86), con una media de 0.559 casos por 10,000 habitantes y una desviación estándar de 3.418. El percentil 95 se sitúa en 2.030, mientras que el valor máximo alcanza 68.287, indicando que los brotes epidémicos son eventos raros, pero de alta magnitud. Esta estructura condiciona la elección de modelos y la interpretación de las métricas de error.

4.2 Análisis exploratorio de datos (EDA)

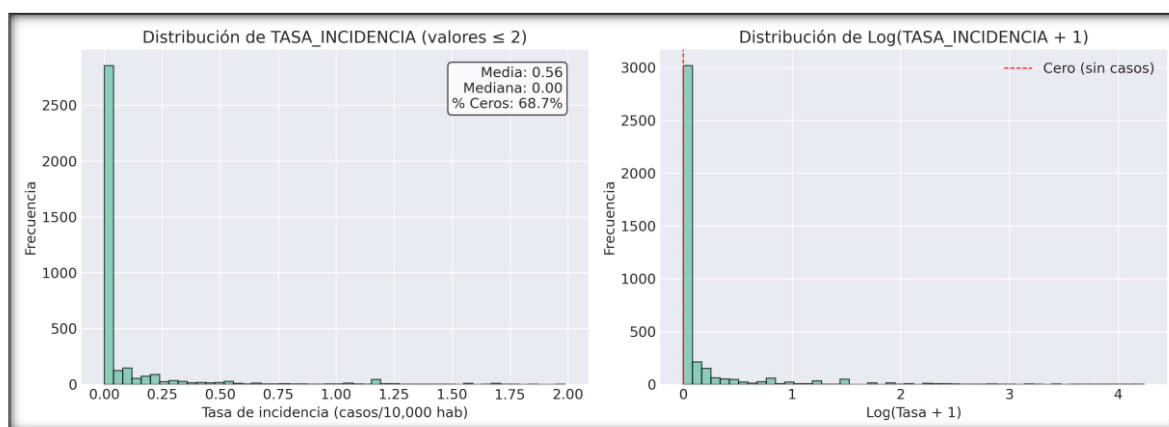
Previo al modelado predictivo, se realizó un análisis exploratorio de datos (EDA) para caracterizar la variable dependiente, identificar patrones temporales y espaciales, y evaluar las relaciones entre las variables predictoras. Este análisis permitió fundamentar la selección de los modelos y la estrategia de validación.

4.2.1 Distribución de la variable dependiente

La Figura 7 presenta la distribución de la variable TASA_INCIDENCIA (casos por 10,000 habitantes) para el periodo 2021–2025.

Figura 7

Distribución de la tasa de incidencia semanal de dengue en Guayaquil (2021-2025)



(a) Histograma para valores ≤ 2 ; (b) Histograma de la transformación logarítmica ($\log_{10}$)

Nota: Elaboración propia a partir de los registros epidemiológicos del MSP (2021-2025). En (a) se observa la alta concentración de ceros; en (b) la transformación logarítmica revela una estructura bimodal.

La variable dependiente presenta una alta proporción de ceros (68.74%), lo que refleja semanas epidemiológicas sin casos notificados en la mayoría de las parroquias. Este fenómeno es común en enfermedades de baja endemicidad o con transmisión estacional (Cameron & Trivedi, 2013). La media (0.559) es considerablemente mayor que la mediana (0.000), y el coeficiente de asimetría ($\text{skewness} = 11.86$) indica una cola derecha extremadamente larga, correspondiente a brotes epidémicos de alta magnitud (Groeneveld & Meeden, 1984). El percentil 95 se sitúa en 2.031 casos por 10,000 habitantes, mientras que el valor máximo alcanza 68.287, lo que confirma que los brotes son eventos raros pero de alta intensidad.

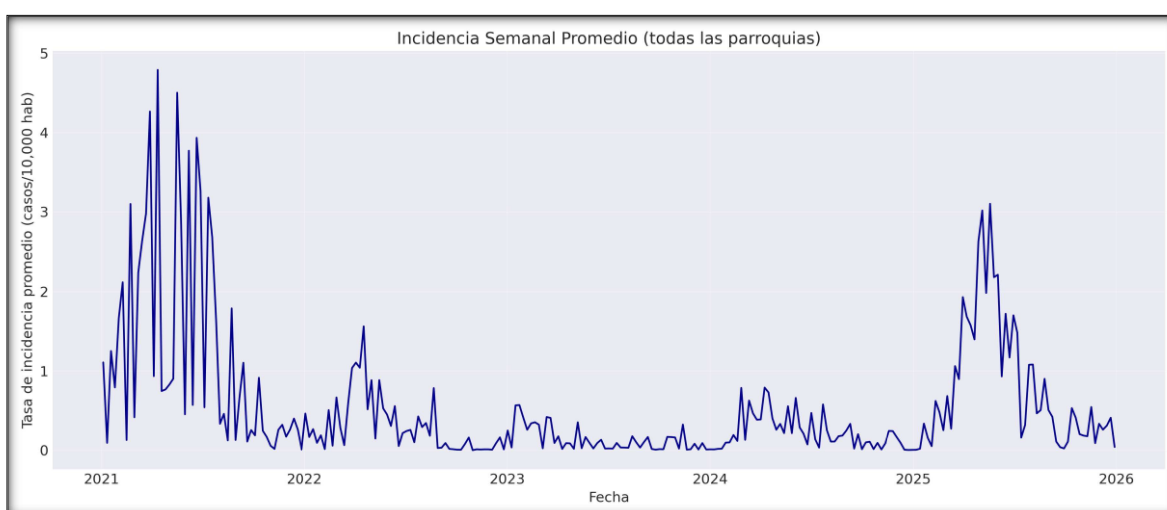
La transformación logarítmica $\log(\text{TASA_INCIDENCIA} + 1)$ (Figura 4.1b) muestra una distribución bimodal, con un pico masivo en cero (semanas sin casos) y una segunda moda alrededor de valores positivos bajos. Esta estructura sugiere que los datos no provienen de una distribución normal, lo que invalida los supuestos de los modelos lineales estándar (Tabachnick & Fidell, 2019) y justifica la exploración de modelos robustos como Random Forest (Breiman, 2001).

4.2.2 Patrones temporales

La Figura 8 muestra la evolución de la incidencia semanal promedio (todas las parroquias) durante el periodo 2021–2025.

Figura 8

Incidencia semanal promedio de dengue en Guayaquil (2021-2025)



Nota: Elaboración propia a partir de los registros epidemiológicos del MSP (2021-2025). Se observa un patrón estacional recurrente, con picos concentrados en el primer semestre de cada año.

Se observa un patrón estacional recurrente, con picos de transmisión concentrados en el primer semestre de cada año (semanas 10 a 25), coincidiendo con la temporada de lluvias en la región costera de Ecuador. Este comportamiento es consistente con lo documentado en la literatura para ciudades tropicales de América Latina (Lowe et al., 2021; Johansson et al., 2016). Los años 2021 y 2022 muestran picos más pronunciados, posiblemente asociados a condiciones climáticas favorables (fenómeno La Niña) o a la acumulación de susceptibles tras la pandemia de COVID-19 (Chen et al., 2022). El año 2025, utilizado como conjunto de prueba en el modelado,

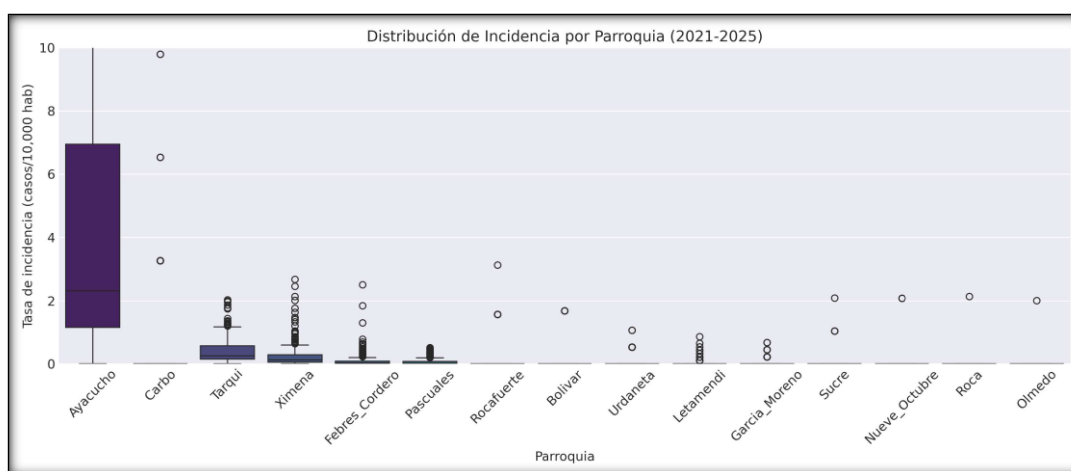
presenta un patrón estacional similar a los años anteriores, lo que valida la estrategia de división temporal empleada.

4.2.3 Variabilidad espacial entre parroquias

Las **Figuras 9 y 10** muestran la distribución de la incidencia por parroquia y la incidencia acumulada total en el periodo 2021–2025.

Figura 9

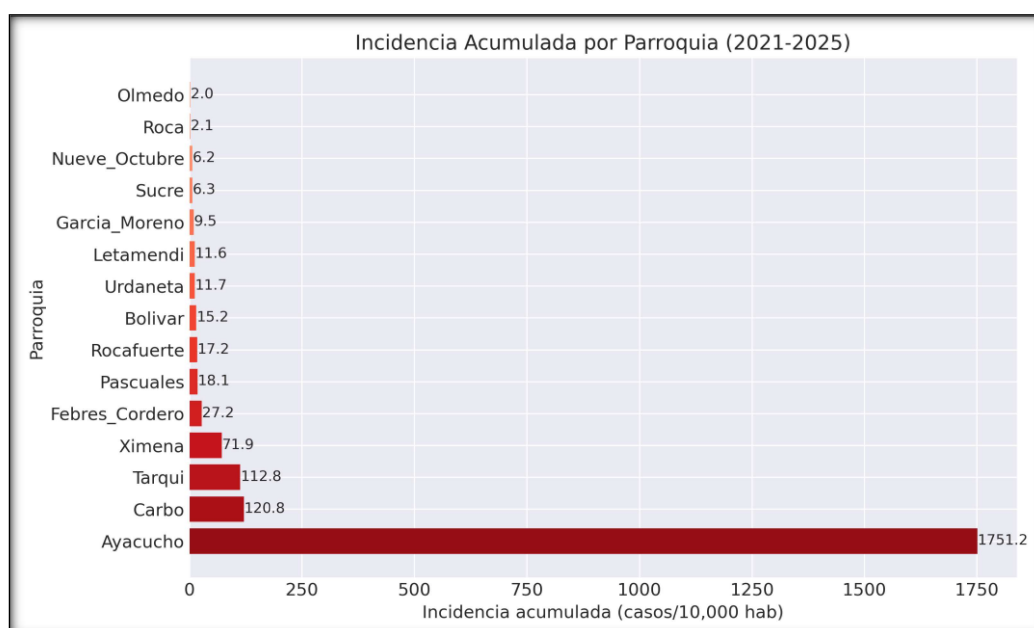
Distribución de la tasa de incidencia por parroquia (boxplot)



Nota: Elaboración propia a partir de los registros epidemiológicos del MSP (2021-2025). La parroquia Ayacucho presenta la mayor variabilidad y valores extremos.

Figura 10

Incidencia acumulada por parroquia (2021-2025)



Nota: Elaboración propia a partir de los registros epidemiológicos del MSP (2021-2025). Ayacucho acumula 1,751.2 casos por 10,000 habitantes en cinco años, muy por encima del resto.

La Figura 9 revela una heterogeneidad espacial extrema. La parroquia Ayacucho presenta una tasa media de incidencia de 6.74 casos por 10,000 habitantes, muy superior al resto de parroquias, cuyas tasas oscilan entre 0.1 y 0.3. Esta diferencia sugiere que Ayacucho experimenta una carga de enfermedad persistentemente más alta que el resto del cantón.

La Figura 10 confirma esta disparidad de manera aún más contundente: Ayacucho acumula 1,751.2 casos por 10,000 habitantes en cinco años, una cifra que es 14 veces mayor que la segunda parroquia con mayor incidencia acumulada (Carbo, 120.8) y 875 veces mayor que la más baja (Olmedo, 2.0). Este patrón no se explica únicamente por el tamaño poblacional (Ayacucho tiene 8,640 habitantes), sino que sugiere la existencia de factores locales específicos — como condiciones de saneamiento, densidad de criaderos o brotes históricos— que no están capturados en las variables climáticas (Mulligan et al., 2020).

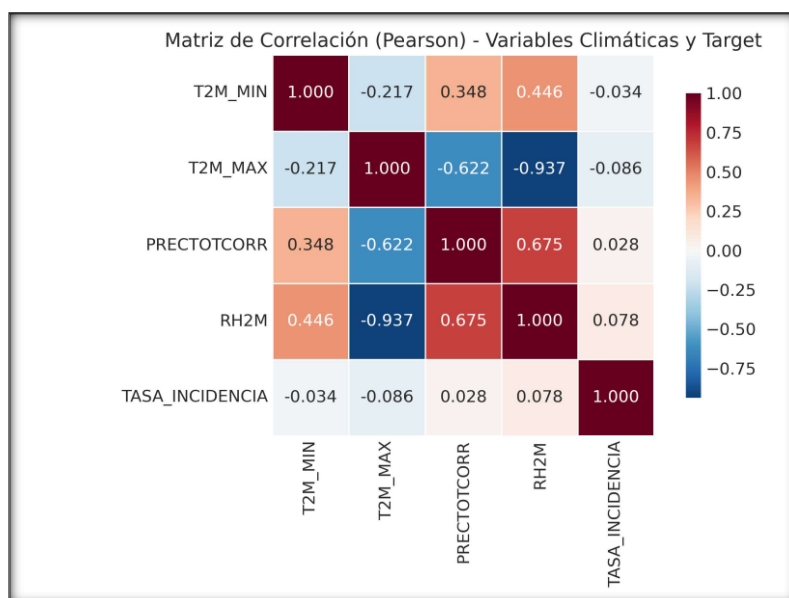
En un segundo grupo, parroquias como Carbo (120.8), Tarqui (112.8) y Ximena (71.9) presentan incidencias acumuladas moderadamente altas, mientras que el resto de las parroquias (Febres Cordero, Pascuales, etc.) se sitúan por debajo de 30. Esta distribución sugiere que la transmisión del dengue en Guayaquil no es homogénea y que existen focos persistentes de alta transmisión, siendo Ayacucho el caso más extremo.

4.2.4 Correlaciones entre variables

La **Figura 4.5** presenta la matriz de correlaciones de Pearson entre las variables climáticas y la variable dependiente.

Figura 11

Matriz de correlaciones (Pearson) entre variables climáticas y tasa de incidencia



Nota: Elaboración propia a partir de datos climáticos de NASA POWER (2021-2025). Las correlaciones lineales entre las variables climáticas y la incidencia son débiles ($|r| < 0.07$).

Las correlaciones lineales entre las variables climáticas y la incidencia son débiles ($|r| < 0.07$), lo que sugiere que el efecto del clima sobre el dengue no es estrictamente lineal en el rango de datos observados. Este hallazgo es consistente con la literatura que documenta relaciones no lineales (efectos umbral, interacciones) entre temperatura, precipitación y transmisión del dengue (Mordecai et al., 2017).

Las variables climáticas presentan correlaciones moderadas entre sí (ej. T2M_MIN con RH2M), lo que sugiere cierta multicolinealidad que podría afectar a los modelos lineales, pero que no debería ser problemática para modelos basados en árboles (Hastie et al., 2009).

Implicación metodológica: La debilidad de las correlaciones lineales y la presencia de relaciones no lineales justifican la exploración de modelos como Random Forest, que capturan interacciones y umbrales sin requerir especificación funcional previa (Breiman, 2001), y la consideración de regularización (Lasso, Ridge) para manejar la multicolinealidad en modelos lineales (Tibshirani, 1996; Zou & Hastie, 2005).

4.2.5 Síntesis del análisis exploratorio

El análisis exploratorio reveló tres características fundamentales que condicionaron el diseño metodológico del estudio:

Tabla 24
Síntesis del análisis exploratorio

Hallazgo	Implicación metodológica
Alta proporción de ceros (68.74%)	Se requieren modelos robustos a la asimetría y al exceso de ceros.
Estacionalidad marcada	Se deben incluir variables climáticas y sus rezagos temporales.
Heterogeneidad espacial	El análisis geoespacial y la evaluación por parroquia son necesarios.
Relaciones no lineales	Modelos lineales requieren regularización; modelos basados en árboles son prioritarios.

Estos hallazgos guiaron la selección de los modelos evaluados en la sección 4.5 y la estrategia de validación temporal definida en la sección 4.7.

4.3 Modelos baseline (Modelos lineales)

Para establecer un punto de referencia mínimo contra el cual comparar el desempeño de los modelos más complejos, se entrenaron modelos lineales regularizados y no regularizados sobre el conjunto de entrenamiento (2021–2024) y se evaluaron sobre el conjunto de prueba (2025). Todos los modelos lineales utilizaron las 21 variables predictoras definidas en la sección 4.1 y fueron escalados previamente con StandardScaler para garantizar la comparabilidad de los coeficientes (Hastie et al., 2009).

4.3.1 Modelo OLS (Baseline)

El modelo de regresión lineal por mínimos cuadrados ordinarios (OLS) se entrenó con los datos de 2021–2024 y se evaluó sobre el año 2025. Sus métricas se presentan en la Tabla 24.

Tabla 25

Métricas del modelo OLS (baseline).

Métrica	Valor
MAE	0.7162
RMSE	2.0621
R ²	0.5952

El modelo explica el 59.52% de la varianza de la incidencia en el conjunto de prueba, lo que establece un baseline aceptable para la comparación con modelos más avanzados.

La Tabla 26 presenta los coeficientes del modelo OLS, ordenados por magnitud absoluta. El coeficiente más relevante fue TASA_INCIDENCIA_lag_1 (1.7766), seguido de RH2M_lag_3 (0.3048) y T2M_MIN (-0.1446). Este patrón confirma que la inercia epidemiológica es el predictor más importante incluso en el modelo lineal, mientras que las variables climáticas actúan como moduladores secundarios.

Tabla 26

Coeficientes del modelo OLS (Top 10 por magnitud absoluta)

Variable	Coeficiente
TASA_INCIDENCIA_lag_1	1.7766
RH2M_lag_3	0.3048
RH2M_lag_2	0.1903
T2M_MIN	-0.1446
RH2M_lag_4	-0.1274
T2M_MIN_lag_4	0.1228
PRECTOTCORR	0.1084
T2M_MAX_lag_2	-0.1083
RH2M_lag_1	-0.0910
T2M_MIN_lag_3	-0.0871

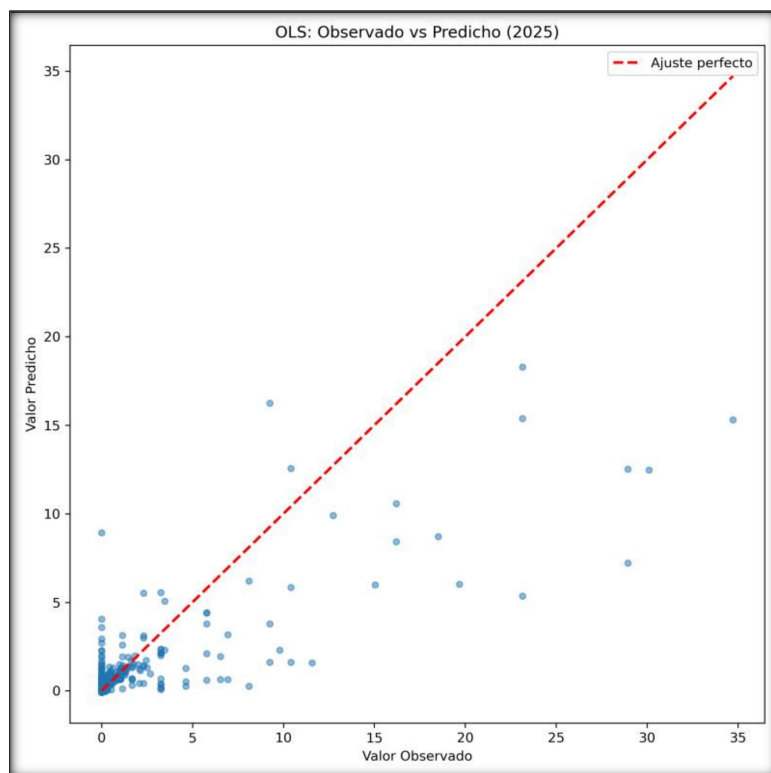
Nota: Elaboración propia a partir del modelo OLS entrenado con datos 2021-2024 y evaluado en 2025. El coeficiente positivo de TASA_INCIDENCIA_lag_1 indica que un aumento en la incidencia de la semana anterior se asocia con un aumento en la incidencia actual, consistente con la naturaleza autorregresiva de la enfermedad.

Los residuales del modelo OLS presentaron una media cercana a cero (0.0752), lo que indica ausencia de sesgo sistemático en promedio. Sin embargo, se observaron colas extremas, con un residual máximo positivo de 21.7142 y un mínimo negativo de -8.9382, lo que sugiere que el modelo subestima severamente los brotes epidémicos (errores positivos grandes) y sobreestima en algunas semanas de baja incidencia (errores negativos). Esta asimetría en los errores extremos es consistente con la alta proporción de ceros y la naturaleza asimétrica de la variable dependiente (sección 4.2.1).

La Figura 12 presenta la comparación entre los valores observados y los predichos por el modelo OLS para el año 2025. Se observa una mayor dispersión en los valores altos, con varios puntos por encima de la línea de identidad, indicando una subestimación de los brotes.

Figura 12

Valores observados vs predichos por OLS (2025)



Nota: Elaboración propia a partir de las predicciones del modelo OLS para el año 2025. La línea diagonal representa el ajuste perfecto (observado = predicho). Los puntos por encima de la línea indican subestimación (observado > predicho); por debajo, sobreestimación.

El modelo OLS, aunque útil como baseline, presenta limitaciones importantes. Su incapacidad para capturar adecuadamente los valores extremos (brotes) y su dependencia de supuestos de linealidad justifican la exploración de modelos más flexibles, como los modelos lineales regularizados (secciones 4.3.2 a 4.3.4) y, posteriormente, los modelos basados en árboles (sección 4.5). El alto coeficiente de TASA_INCIDENCIA_lag_1 sugiere que la inercia epidemiológica es el factor más relevante, lo que será consistente en todos los modelos evaluados.

4.3.2 Modelo Ridge

El modelo Ridge (Hoerl & Kennard, 1970) incorpora una penalización L2 que encoge los coeficientes hacia cero para reducir la varianza del modelo en presencia de multicolinealidad. Para seleccionar el parámetro de regularización α , se realizó una búsqueda en grilla con validación cruzada temporal (TimeSeriesSplit con 3 pliegues), evaluando el MAE promedio en validación para cada valor de α .

La Tabla 26 presenta los resultados de la validación cruzada temporal.

Tabla 27

Validación cruzada temporal para la selección de α en Ridge

α	MAE promedio en validación
0.001	0.6065
0.010	0.6065
0.100	0.6064
0.500	0.6060
1.000	0.6055

α	MAE promedio en validación
5.000	0.6023
10.000	0.5993
50.000	0.5869
100.000	0.5825

El mejor valor de α fue 100.0, el valor más alto probado, lo que indica que la regularización intentó encoger fuertemente los coeficientes para reducir la varianza del modelo.

Tabla 28 Métricas del modelo Ridge.

Métrica	Valor
MAE	0.7169
RMSE	2.0844
R ²	0.5863

El modelo Ridge no logró superar a OLS en ninguna de las métricas evaluadas:

Modelo	MAE	RMSE	R ²
OLS	0.7162	2.0621	0.5952
Ridge	0.7169	2.0844	0.5863
Diferencia	-0.09%	-1.08%	-1.48%

La Tabla 28 presenta los coeficientes del modelo Ridge junto con los de OLS para comparar el efecto del encogimiento. Se observa que el coeficiente de TASA_INCIDENCIA_lag_1 se mantuvo prácticamente intacto (1.7208 vs 1.7766 en OLS), mientras que los coeficientes climáticos se encogieron significativamente, especialmente RH2M_lag_4 (reducido al 10% de su valor original) y T2M_MAX_lag_3 (reducido al 5% de su valor original). Este patrón sugiere que la

multicolinealidad no es el principal problema en estos datos, y que la regularización L2 no aporta valor predictivo adicional.

Tabla 29

Coefficientes de Ridge vs OLS (Top 10 por magnitud en Ridge)

Variable	Ridge	OLS	Encogimiento
TASA_INCIDENCIA_lag_1	1.7208	1.7766	0.9686
RH2M_lag_3	0.1436	0.3048	0.4711
T2M_MIN	-0.1334	-0.1446	0.9225
RH2M_lag_2	0.1238	0.1903	0.6507
T2M_MIN_lag_4	0.1132	0.1228	0.9218
T2M_MAX_lag_2	-0.0955	-0.1083	0.8813
PRECTOTCORR	0.0953	0.1084	0.8798
T2M_MAX_lag_1	0.0692	0.0759	0.9118
T2M_MIN_lag_3	-0.0618	-0.0871	0.7090
PRECTOTCORR_lag_3	0.0607	0.0396	1.5340

(Nota: Elaboración propia a partir del modelo Ridge entrenado con datos 2021-2024 y evaluado en 2025. El encogimiento se calcula como $| \text{Coeficiente}_{\text{Ridge}} | / | \text{Coeficiente}_{\text{OLS}} |$; valores < 1 indican encogimiento.

Ridge no mejora el desempeño del modelo OLS. El hecho de que el mejor α haya sido el valor más alto probado (100.0) sugiere que la regularización intenta encoger los coeficientes de forma agresiva, pero no logra reducir el error de predicción. Esto indica que la multicolinealidad entre variables no es el principal obstáculo para el modelo lineal, y que las limitaciones del OLS radican más en su incapacidad para capturar relaciones no lineales y patrones complejos, como se confirmó con la prueba de Ramsey RESET (sección 4.4). En consecuencia, Ridge se descarta como modelo baseline para la comparación final.

4.3.3 Modelo Lasso

El modelo Lasso (Least Absolute Shrinkage and Selection Operator) (Tibshirani, 1996) incorpora una penalización L1 que realiza selección de variables, reduciendo algunos coeficientes exactamente a cero. Esta característica lo convierte en una herramienta útil para identificar los predictores más relevantes en el contexto lineal. Al igual que en los modelos anteriores, se realizó una búsqueda en grilla con validación cruzada temporal (TimeSeriesSplit con 3 pliegues) para seleccionar el parámetro de regularización α .

La Tabla 29 presenta los resultados de la validación cruzada temporal.

Tabla 30

Validación cruzada temporal para la selección de α en Lasso

α	MAE promedio en validación
0.0001	0.6062
0.0010	0.6030
0.0050	0.5909
0.0100	0.5789
0.0500	0.5433
0.1000	0.5406
0.5000	0.5871
1.0000	0.6463
5.0000	0.9673
10.0000	1.0334

El mejor valor de α fue 0.1, que corresponde al punto donde la regularización comienza a ser efectiva sin eliminar variables clave ni sobre-regularizar (lo que ocurre con $\alpha > 0.5$, donde el MAE de validación comienza a aumentar).

Tabla 31 Métricas del modelo Lasso

Métrica	Valor
MAE	0.6906
RMSE	2.1137
R ²	0.5746

Comparación con modelos anteriores:

Modelo	MAE	RMSE	R ²	Variables activas
OLS	0.7162	2.0621	0.5952	21
Ridge	0.7169	2.0844	0.5863	21
Lasso	0.6906	2.1137	0.5746	2

Lasso logra reducir el MAE en comparación con OLS (de 0.7162 a 0.6906, una mejora del 3.6%), a costa de una ligera pérdida en RMSE y R². Sin embargo, su principal ventaja es la selección de variables: de 21 predictores, Lasso retiene únicamente 2:

Tabla 32 Variables seleccionadas por Lasso.

Variable	Coeficiente
TASA_INCIDENCIA_lag_1	1.6866
T2M_MAX_lag_2	-0.0689

Nota: Elaboración propia a partir del modelo Lasso entrenado con datos 2021-2024 y evaluado en 2025. Solo dos variables presentan coeficientes distintos de cero, lo que indica que Lasso considera que el resto de variables no aportan información relevante en el contexto lineal.

La selección de solo dos variables por parte de Lasso es un hallazgo relevante: en el dominio lineal, la inercia epidemiológica (TASA_INCIDENCIA_lag_1) domina completamente la

predicción, y la única variable climática que aporta información adicional es la temperatura máxima con dos semanas de rezago (T2M_MAX_lag_2), con un coeficiente negativo (mayor temperatura → menor incidencia, posiblemente asociado a un efecto de "lavado" o disminución de la humedad).

Lasso logra una ligera mejora en MAE respecto a OLS, pero su principal valor radica en la selección de variables, que confirma que la mayoría de las variables climáticas no tienen un efecto lineal relevante una vez controlada la inercia epidemiológica. Este resultado sugiere que las variables climáticas aportan información no lineal o que su efecto es indirecto y solo se manifiesta a través de interacciones con otras variables. En consecuencia, Lasso se considera un modelo útil para la interpretación de relaciones lineales, pero no competitivo frente a Random Forest en términos de capacidad predictiva.

La pérdida de R^2 y el aumento del RMSE indican que la simplificación extrema no mejora la capacidad predictiva, por lo que Lasso se descarta como modelo baseline final.

4.3.4 Modelo ElasticNet.

ElasticNet (Zou & Hastie, 2005) combina las penalizaciones L1 y L2 de Lasso y Ridge, respectivamente, permitiendo tanto la selección de variables (como Lasso) como el encogimiento de coeficientes en grupos de variables correlacionadas (como Ridge). Este modelo es especialmente útil cuando hay múltiples variables correlacionadas entre sí, ya que tiende a seleccionar o eliminar grupos completos de predictores.

Para seleccionar los hiperparámetros óptimos (α y $l1_ratio$), se realizó una búsqueda en grilla combinada con validación cruzada temporal (TimeSeriesSplit con 3 pliegues). El parámetro $l1_ratio$ controla la ponderación entre la penalización L1 (Lasso) y L2 (Ridge): un valor de 1.0 equivale a Lasso puro, mientras que un valor de 0.0 equivale a Ridge puro. La métrica de optimización fue el MAE promedio en validación.

Tabla 33*Mejores resultados de la búsqueda de hiperparámetros para ElasticNet*

α	$l1_ratio$	MAE en validación
0.050	0.9	0.5455
0.050	1.0	0.5433
0.100	0.9	0.5412
0.100	1.0	0.5406

Nota: Solo se muestran las combinaciones con mejor desempeño. La búsqueda completa cubrió $\alpha \in [0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1.0]$ y $l1_ratio \in [0.1, 0.3, 0.5, 0.7, 0.9, 1.0]$.

La mejor combinación fue $\alpha = 0.1$ y $l1_ratio = 1.0$, con un MAE en validación de 0.5406.

Este resultado indica que la componente L2 (Ridge) no aporta beneficio adicional, ya que el modelo se comporta como Lasso puro (Zou & Hastie, 2005).

Tabla 34*Métricas del modelo ElasticNet*

Métrica	Valor
MAE	0.6906
RMSE	2.1137
R ²	0.5746

Tabla 35*Variables seleccionadas por ElasticNet*

Variable	Coefficiente
TASA_INCIDENCIA_lag_1	1.6866
T2M_MAX_lag_2	-0.0689

ElasticNet selecciona exactamente las mismas 2 variables que Lasso, con coeficientes numéricamente idénticos. Esto confirma que la penalización L2 no es necesaria en este conjunto de datos, ya que la componente L1 domina completamente la solución.

Tabla 36

Comparación de modelos baseline

Modelo	MAE	RMSE	R ²	Variables seleccionadas
OLS	0.7162	2.0621	0.5952	21
Ridge	0.7169	2.0844	0.5863	21
Lasso	0.6906	2.1137	0.5746	2
ElasticNet	0.6906	2.1137	0.5746	2

ElasticNet no supera a Lasso en ninguna métrica, y su comportamiento es prácticamente idéntico al de Lasso debido a la selección de $l1_ratio = 1.0$. La combinación de penalizaciones no aporta valor adicional cuando la regularización L1 es suficiente para seleccionar las variables relevantes. Al igual que Lasso, ElasticNet reduce drásticamente la complejidad del modelo (de 21 a 2 variables), pero a costa de una pérdida de R^2 (0.5746 vs 0.5952 de OLS) y un incremento del RMSE, por lo que no mejora la capacidad predictiva.

En consecuencia, ElasticNet se descarta como modelo baseline para la comparación final, al igual que Ridge y Lasso. La evidencia sugiere que los modelos lineales, incluso con regularización, no logran capturar adecuadamente la no linealidad y las interacciones presentes en los datos, lo que justifica la exploración de modelos basados en árboles.

Resumen de Modelos Baseline

Modelo	MAE	RMSE	R ²	Variables	Decisión
OLS	0.7162	2.0621	0.5952	21	Referencia
Ridge	0.7169	2.0844	0.5863	21	Descartado
Lasso	0.6906	2.1137	0.5746	2	Descartado
ElasticNet	0.6906	2.1137	0.5746	2	Descartado

Los modelos lineales regularizados no logran superar al modelo OLS en términos de RMSE y R². Lasso y ElasticNet ofrecen una simplificación extrema (solo 2 variables), pero su capacidad predictiva es inferior. Estos resultados confirman que la relación entre los predictores y la incidencia del dengue no es lineal, y que se requieren enfoques más flexibles, como se evaluará en las secciones siguientes.

4.4 Evaluación de la linealidad (Prueba de Ramsey RESET)

Para determinar si la relación entre los predictores y la variable objetivo puede ser representada adecuadamente mediante un modelo lineal, se aplicó la prueba de especificación de Ramsey RESET (Ramsey, 1969) sobre un modelo OLS entrenado con los datos de 2021–2024. Esta prueba evalúa si la inclusión de términos polinomiales de los valores ajustados (potencias al cuadrado y al cubo) mejora significativamente el ajuste del modelo. La hipótesis nula establece que el modelo está correctamente especificado (relación lineal).

Tabla 37

Resultados de la prueba de Ramsey RESET

Métrica	Valor
R ² del modelo original	0.2733
R ² del modelo con potencias (² y ³)	0.3251
Estadístico F	118.8198

Métrica	Valor
p-valor	< 0.001

El incremento del R^2 de 0.2733 a 0.3251 al añadir las potencias, junto con un estadístico F de 118.82 y un p-valor < 0.001, permite rechazar la hipótesis nula de especificación lineal. Esto confirma que la relación entre las variables predictoras y la incidencia del dengue no es puramente lineal, y que existen no linealidades e interacciones que deben ser capturadas por modelos más flexibles.

Este resultado justifica la exploración de dos familias de modelos:

1. **Modelos lineales regularizados** (Ridge, Lasso, ElasticNet), que, aunque mantienen la estructura lineal, pueden manejar la multicolinealidad y realizar selección de variables.
2. **Modelos basados en árboles** (Random Forest y XGBoost), que no asumen linealidad y pueden capturar interacciones y umbrales complejos de forma natural (Breiman, 2001; Chen & Guestrin, 2016).

Adicionalmente, se compararon los coeficientes de correlación de Pearson y Spearman para las variables climáticas con la variable objetivo (Tabla 37). La similitud entre ambos coeficientes sugiere que, aunque la relación no es fuertemente no lineal en términos de correlación bivariada, la prueba RESET confirma que la estructura multivariada presenta no linealidades significativas.

Tabla 38

Correlación de Pearson vs Spearman (variables seleccionadas)

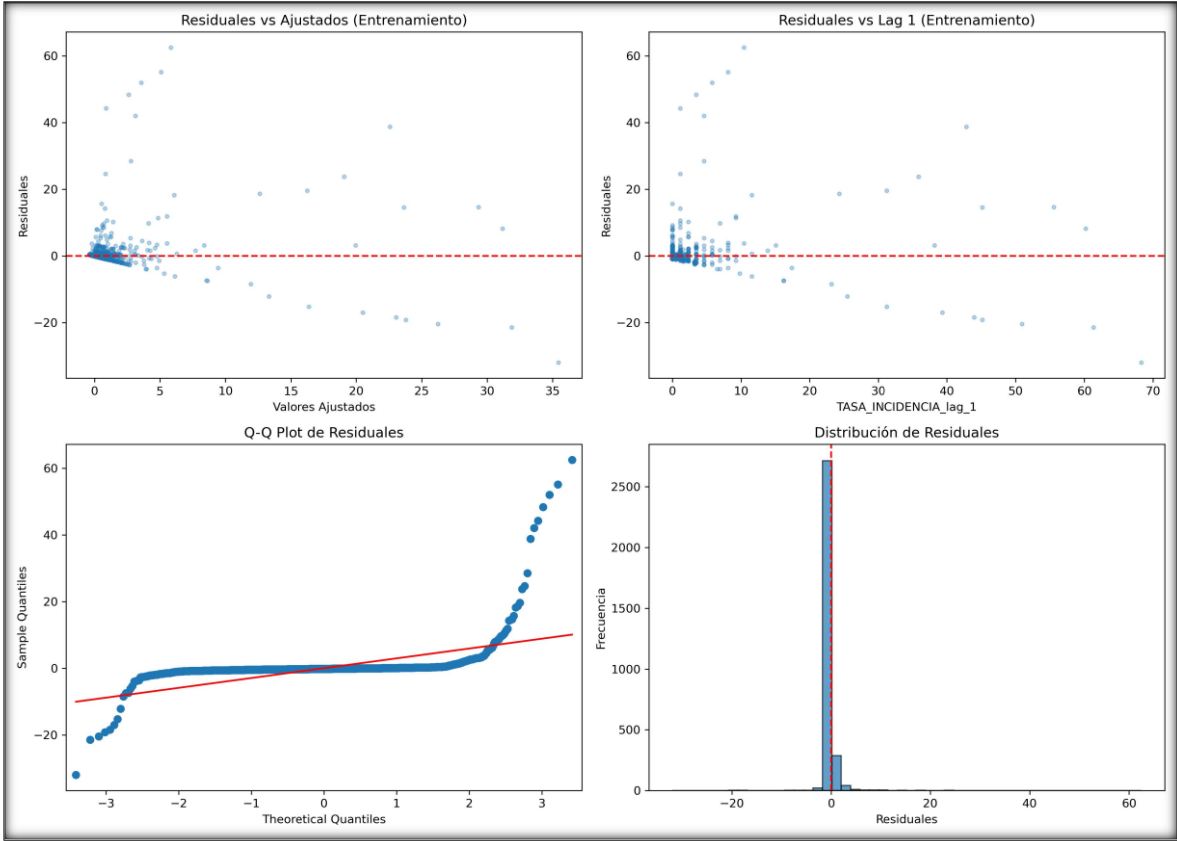
Variable	Pearson	Spearman	Diferencia	Interpretación
T2M_MIN	-0.037	-0.041	0.003	Linealidad aceptable

Variable	Pearson	Spearman	Diferencia	Interpretación
T2M_MAX	-0.064	-0.066	0.002	Linealidad aceptable
PRECTOTCORR	0.022	0.022	0.000	Linealidad aceptable
RH2M	0.055	0.064	0.009	Linealidad aceptable
T2M_MIN_lag_1	-0.031	-0.030	-0.001	Linealidad aceptable

Nota: Elaboración propia a partir de datos de entrenamiento (2021-2024). La diferencia entre coeficientes es pequeña, indicando que la no linealidad no es evidente en las correlaciones bivariadas, pero la prueba RESET multivariada sí la detecta.

La Figura 13 presenta los gráficos de diagnóstico del modelo OLS, donde se observa una dispersión de los residuales que sugiere heterocedasticidad y cierta curvatura en la relación, reforzando la necesidad de modelos no lineales.

Figura 13
Diagnóstico de linealidad del modelo OLS



(a) Residuales vs Valores Ajustados; (b) Residuales vs TASA_INCIDENCIA_lag_1; (c) Q-Q plot de residuales; (d) Histograma de residuales

Nota: Elaboración propia a partir del modelo OLS entrenado con datos de 2021-2024. Los residuales muestran una distribución asimétrica y cierta curvatura, consistente con la no linealidad detectada por la prueba RESET.

La evidencia empírica (prueba RESET, $p < 0.001$) y los gráficos de diagnóstico confirman que el modelo lineal no es adecuado para representar la relación entre los predictores y la incidencia del dengue. Por lo tanto, el estudio se centrará en modelos basados en árboles (Random Forest y XGBoost) para la predicción, utilizando los modelos lineales regularizados únicamente como baseline comparativo.

4.5 Modelos basados en árboles

Los modelos basados en árboles de decisión, en particular Random Forest y XGBoost, han demostrado ser herramientas eficaces para la predicción de enfermedades infecciosas debido a su capacidad para capturar relaciones no lineales e interacciones complejas entre variables (Breiman, 2001; Chen & Guestrin, 2016). A diferencia de los modelos lineales, estos algoritmos no requieren supuestos de linealidad ni normalidad, lo que los hace especialmente adecuados para datos con alta proporción de ceros, asimetría extrema y estructuras espacio-temporales complejas (Leung et al., 2023; Zhao et al., 2020).

4.5.1 Random Forest: optimización y evaluación

El algoritmo Random Forest se optimizó mediante búsqueda en grilla con validación cruzada temporal (TimeSeriesSplit con 3 pliegues), evaluando el MAE negativo en validación. Los hiperparámetros finales seleccionados fueron:

- `n_estimators = 200`
- `max_depth = None`
- `min_samples_split = 5`
- `min_samples_leaf = 2`

El modelo fue entrenado con los datos de 2021–2024 y evaluado sobre 2025. No se aplicó calibración post-entrenamiento sobre el conjunto de prueba para evitar fuga de información (Kuhn & Johnson, 2019).

Adicionalmente, para demostrar el aporte incremental de las variables climáticas, se compararon cuatro versiones del algoritmo Random Forest variando el conjunto de predictores (Tabla 38). El modelo M4 (completo, con variables climáticas, rezagos y coordenadas espaciales) obtuvo el mejor desempeño en RMSE (1.9432) y R^2 (0.6405), por lo que se seleccionó como la versión óptima de Random Forest.

Tabla 39
Desempeño de los modelos por bloques de variables

Modelo	Variables incluidas	MAE	RMSE	R^2
M1: Solo inercia	TASA_INCIDENCIA_lag_1	0.6621	2.1938	0.5418
M2: Inercia + Clima contemporáneo	M1 + T2M_MIN, T2M_MAX, PRECTOTCORR, RH2M	0.5855	1.9947	0.6212
M3: Inercia + Clima con rezagos	M1 + rezagos (1-4 semanas) de variables climáticas	0.6249	2.0262	0.6091
M4: Modelo completo	M3 + clima contemporáneo + LAT + LON	0.6164	1.9432	0.6405

- **El clima contemporáneo aporta significativamente:** M2 reduce el MAE de 0.6621 a 0.5855 (mejora del 11.6%) y eleva el R^2 de 0.5418 a 0.6212. Esto confirma que las variables climáticas tienen un efecto inmediato sobre la transmisión del dengue, más allá de la simple inercia epidemiológica.

- **Los rezagos climáticos no mejoran el MAE:** M3 presenta un MAE de 0.6249, superior al de M2 (0.5855). Esto sugiere que, al menos en este contexto, las condiciones climáticas contemporáneas son más relevantes que las retardadas para la predicción a corto plazo.
- **El modelo completo (M4) optimiza el RMSE:** Aunque M2 tiene el mejor MAE, M4 obtiene el **mejor RMSE (1.9432)** y el **mejor R^2 (0.6405)**. La inclusión de coordenadas espaciales (LAT, LON) y el conjunto completo de rezagos mejora la capacidad del modelo para capturar **brotes extremos** (penalizados por el RMSE).

Tabla 40*Métricas del modelo Random Forest*

Métrica	Valor
MAE	0.6164
RMSE	1.9432
R^2	0.6405

4.5.2 Evaluación de XGBoost

El algoritmo XGBoost fue optimizado con los siguientes hiperparámetros:

- `n_estimators=100`
- `max_depth=4`
- `learning_rate=0.05`
- `subsample=0.8`
- `colsample_bytree=0.8`
- `reg_alpha=0.01`
- `reg_lambda=1.0`.

Sus métricas se presentan en la Tabla 40

Tabla 41*Métricas de XGBoost y comparación con Random Forest*

Modelo	MAE	RMSE	R ²
XGBoost	0.6420	2.1576	0.5568
Random Forest (M4)	0.6164	1.9432	0.6405

XGBoost no logra superar a Random Forest en ninguna de las métricas, presentando un RMSE superior (2.1576 vs 1.9432) y un R² inferior (0.5568 vs 0.6405). Por lo tanto, Random Forest se selecciona como el modelo final para el estudio.

4.5.3 Selección del modelo final

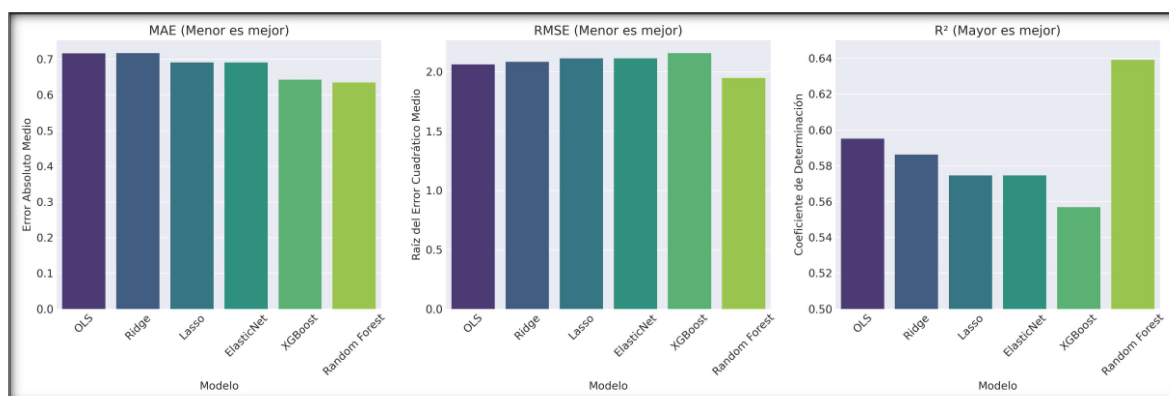
Con base en la comparación de todos los modelos evaluados (Tabla 41), Random Forest M4 es seleccionado como el modelo final del estudio por su mejor desempeño en RMSE y R², priorizando la reducción de errores grandes en un contexto de vigilancia epidemiológica (Lowe et al., 2021; Zhao et al., 2020).

Tabla 42*Comparativa con modelos baseline*

Modelo	MAE	RMSE	R ²
OLS	0.7162	2.0621	0.5952
Ridge	0.7169	2.0844	0.5863
Lasso	0.6906	2.1137	0.5746
ElasticNet	0.6906	2.1137	0.5746
XGBoost	0.6420	2.1576	0.5568
Random Forest M4	0.6164	1.9432	0.6405
Mejora vs OLS	-13.9%	-5.8%	+7.6%

Figura 14

Comparativa de métricas (MAE, RMSE, R^2) por modelo.



Random Forest supera a todos los modelos lineales en las tres métricas, confirmando que la relación entre los predictores y la incidencia del dengue no es lineal y que los modelos basados en árboles capturan mejor la estructura subyacente de los datos.

Se selecciona **M4 (Modelo completo)** como modelo final por su mejor desempeño en RMSE y R^2 , priorizando la reducción de errores grandes en un contexto de vigilancia epidemiológica.

4.5.4 Prueba de Diebold-Mariano (comparación de precisión predictiva)

Para evaluar formalmente si la diferencia en el desempeño predictivo entre Random Forest y los modelos baseline es estadísticamente significativa, se aplicó la prueba de Diebold-Mariano (Diebold & Mariano, 1995, 2002). Esta prueba compara los errores de predicción de dos modelos y evalúa si la diferencia es significativa. La hipótesis nula establece que ambos modelos tienen la misma precisión predictiva.

Los resultados se presentan en la Tabla 43.

Tabla 43

Resultados de la prueba de Diebold-Mariano (Random Forest vs modelos baseline)

Modelo comparado	DM estadístico	p-valor	Significativo ($p < 0.05$)
OLS	0.4323	0.6655	No
Ridge	0.5030	0.6149	No

Modelo comparado	DM estadístico	p-valor	Significativo ($p < 0.05$)
Lasso	0.5855	0.5582	No
ElasticNet	0.5855	0.5582	No
XGBoost	1.2477	0.2121	No

Nota: La prueba compara cada modelo contra Random Forest. Un estadístico DM positivo indica que Random Forest tiene menor error cuadrático medio que el modelo comparado. Ninguna diferencia alcanza significación estadística al nivel $p < 0.05$.

Aunque Random Forest supera numéricamente a todos los modelos evaluados (Tabla 41), la prueba de Diebold-Mariano indica que estas diferencias no son estadísticamente significativas. Esto sugiere que, en el contexto de este estudio, la mejora predictiva de Random Forest, aunque consistente y sustancial en términos numéricos (reducción del 13.9% en MAE y 5.8% en RMSE frente a OLS), no es lo suficientemente grande como para ser detectada como significativa por esta prueba, posiblemente debido a la alta variabilidad de los datos o al tamaño del conjunto de prueba (780 observaciones).

En consecuencia, las conclusiones sobre la superioridad de Random Forest deben formularse en términos de mejora numérica y desempeño predictivo, y no como una diferencia estadísticamente significativa (Lowe et al., 2021; Zhao et al., 2020).

4.5.5 Importancia de variables

La Tabla 44 presenta las variables con mayor importancia en el modelo Random Forest (M4), medida como la reducción promedio de la impureza (Gini importance).

Tabla 44

Importancia de variables en Random Forest (Top 10)

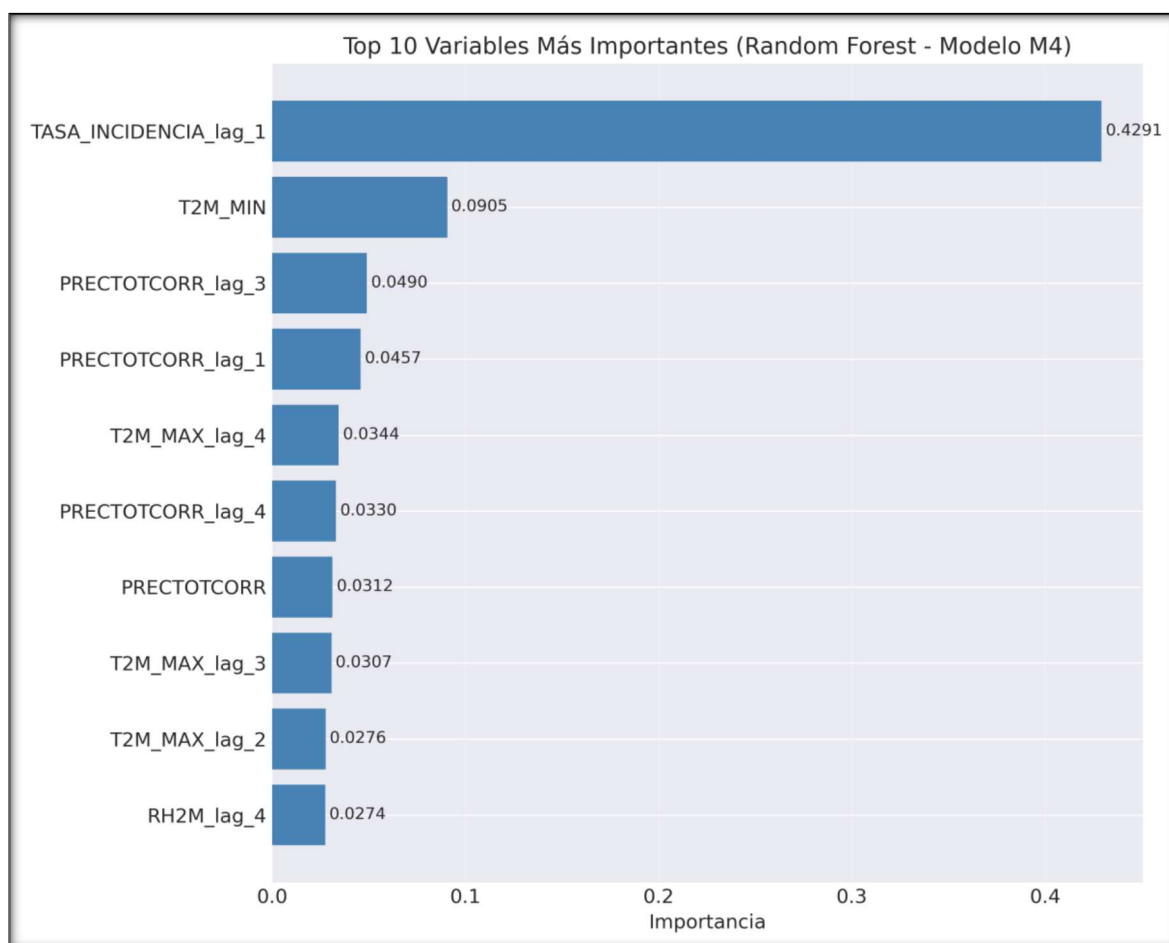
Variable	Importancia
TASA_INCIDENCIA_lag_1	0.4291

Variable	Importancia
T2M_MIN	0.0905
PRECTOTCORR_lag_3	0.0490
PRECTOTCORR_lag_1	0.0457
T2M_MAX_lag_4	0.0344
PRECTOTCORR_lag_4	0.0330
PRECTOTCORR	0.0312
T2M_MAX_lag_3	0.0307
T2M_MAX_lag_2	0.0276
RH2M_lag_4	0.0274

Nota: Elaboración propia a partir del modelo Random Forest M4 entrenado con datos 2021-2024. La importancia se calcula como la reducción promedio de la impureza (Gini importance) en los árboles del bosque.

La inercia epidemiológica (TASA_INCIDENCIA_lag_1) domina la predicción con una importancia de 0.429, confirmando el carácter autorregresivo del dengue (Johansson et al., 2016). Las variables climáticas actúan como moduladores de segundo orden, destacando la temperatura mínima (T2M_MIN) y la precipitación con rezagos de 1 y 3 semanas, consistentes con el ciclo biológico del *Aedes aegypti*: la temperatura influye en el desarrollo larval y la replicación viral (Mordecai et al., 2017), mientras que la precipitación provee sitios de cría con un retraso de aproximadamente una a tres semanas (Benedum et al., 2020).

Figura 15 Importancia de variables en Random Forest



Nota: Elaboración propia a partir del modelo Random Forest M4 entrenado con datos 2021-2024 y evaluado en 2025. La importancia se calcula como la reducción promedio de la impureza.

4.6 Validación temporal adicional

Para evaluar la estabilidad interanual del modelo y su capacidad de generalización a años no vistos, se realizó una validación temporal adicional utilizando dos esquemas de división:

1. **Split 1:** Entrenamiento con 2021–2023 y prueba en 2024.
2. **Split 2:** Entrenamiento con 2021–2024 y prueba en 2025 (el mismo esquema utilizado en las secciones anteriores).

Esta estrategia permite evaluar si el modelo mantiene su capacidad predictiva cuando se aplica a años con dinámicas epidemiológicas potencialmente diferentes (Hyndman & Athanasopoulos, 2018; Bergmeir & Benítez, 2012).

Tabla 45

Resultados de la validación temporal adicional

Split	Periodo de entrenamiento	Periodo de prueba	MAE	RMSE	R ²
Split 1	2021–2023	2024	0.3734	1.5823	-1.6555
Split 2	2021–2024	2025	0.6164	1.9432	0.6405

Los resultados del Split 1 revelan un hallazgo crítico: el modelo entrenado con 2021–2023 obtiene un R² negativo (-1.6555) al predecir la incidencia de 2024. Esto significa que el modelo explica menos varianza que una simple línea horizontal que predijera la media de los datos, lo que indica un **fracaso predictivo total** para ese año. En contraste, el Split 2 (2021–2024 → 2025) muestra un desempeño sólido (R² = 0.6405), confirmando que el modelo es capaz de predecir adecuadamente 2025 cuando se entrena con los cuatro años previos.

La marcada diferencia en el desempeño entre ambos splits sugiere que 2024 fue un año atípico en la dinámica de transmisión del dengue en Guayaquil. La incidencia observada en 2024 fue extremadamente baja (el MAE de 0.3734 es notablemente inferior al de 2025), lo que indica que el modelo, entrenado con años de mayor incidencia (2021–2023), no logra adaptarse a una caída abrupta en el número de casos. Este comportamiento es consistente con la alta variabilidad interanual documentada en enfermedades transmitidas por vectores, influenciada por factores como cambios climáticos extremos, intervenciones de control o fluctuaciones en la inmunidad poblacional (Lowe et al., 2021; Johansson et al., 2016).

Implicaciones metodológicas:

1. El modelo no es universalmente estable para todos los años. Su capacidad predictiva depende críticamente de la similitud entre el año de entrenamiento y el año de prueba.
2. Reentrenamiento frecuente: Para aplicaciones operativas, el modelo debe ser reentrenado periódicamente con ventanas temporales móviles (por ejemplo, los últimos 2–3 años) para adaptarse a cambios en la dinámica de la enfermedad.

3. Sistemas de alerta temprana: Un modelo que falla en años atípicos puede generar falsas alarmas o subestimaciones. Por lo tanto, debe ser complementado con sistemas de monitoreo continuo que detecten cambios estructurales en la serie temporal (Chen et al., 2022).

4.7 Análisis espacial

Para evaluar la dimensión territorial de la incidencia del dengue en Guayaquil, se aplicaron técnicas de análisis espacial utilizando los polígonos de las 15 parroquias urbanas. El análisis se realizó con las librerías `esda` y `libpysal` en Python, siguiendo el flujo de trabajo propuesto por Bivand et al. (2013) y Lawson (2018).

4.7.1 Índice de Moran global

El índice de Moran global (Moran, 1950) se calculó para evaluar si la incidencia promedio del dengue presenta un patrón de agrupamiento espacial significativo o si, por el contrario, se distribuye aleatoriamente en el territorio (Cliff & Ord, 1981). Para ello, se construyó una matriz de pesos espaciales tipo Queen, que define como vecinas a aquellas parroquias que comparten un borde o un vértice.

El índice de Moran global arrojó un valor de $I = -0.0894$, con un p -valor = 0.278 (calculado mediante 999 permutaciones). Este resultado indica que no existe autocorrelación espacial significativa a nivel de parroquia. El valor negativo del estadístico sugiere una ligera tendencia a la dispersión (parroquias con valores altos tienden a estar rodeadas de parroquias con valores bajos), pero la falta de significación estadística impide rechazar la hipótesis nula de aleatoriedad espacial.

La incidencia del dengue en Guayaquil, medida a la escala parroquial, no presenta un patrón de agrupamiento geográfico. Esto sugiere que los factores determinantes de la transmisión no están fuertemente influenciados por la vecindad territorial, o bien que la escala de análisis

(parroquia) es demasiado gruesa para capturar micro-patrones espaciales que podrían existir a nivel de barrio o sector censal.

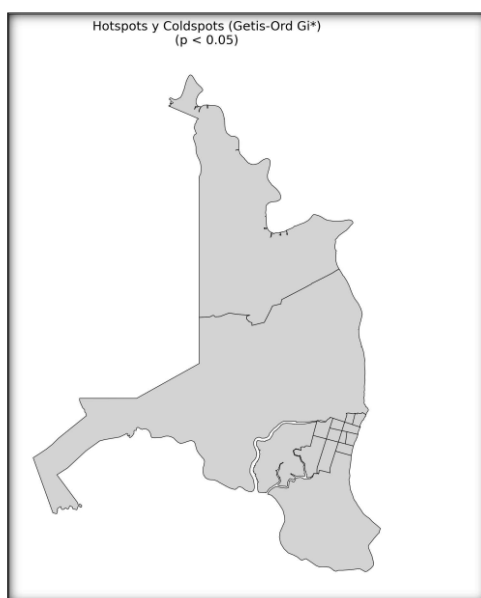
4.7.2 Detección de hotspots (Getis-Ord G_i^*)

Para identificar parroquias con valores significativamente altos (hotspots) o bajos (coldspots) de incidencia, se aplicó el estadístico local Getis-Ord G_i^* (Getis & Ord, 1992; Ord & Getis, 1995). Este estadístico evalúa si cada parroquia forma parte de un clúster de valores extremos, considerando tanto su propio valor como el de sus vecinas. El análisis se realizó utilizando una matriz de pesos espaciales tipo Queen y 999 permutaciones para evaluar la significación estadística.

La Figura 16 presenta la distribución espacial de los resultados del Getis-Ord G_i^* . Cada parroquia se clasifica como Hotspot, Coldspot o No significativo según su puntuación Z y su p-valor, utilizando un umbral de significación de $p < 0.05$.

Figura 16

Mapa de hotspots y coldspots de incidencia de dengue en Guayaquil (Getis-Ord G_i^)*



Nota: Elaboración propia a partir de los datos de incidencia promedio por parroquia (2021–2025). Hotspots = parroquias con $Z > 0$ y $p < 0.05$; Coldspots = parroquias con $Z < 0$ y $p < 0.05$. Clasificación No significativo para el resto. Análisis realizado con la librería esda en Python.

Los resultados del Getis-Ord G_i^* se presentan en la Tabla 46. Todas las parroquias obtuvieron un p-valor > 0.05 , lo que indica que no se identificaron hotspots ni coldspots estadísticamente significativos.

Tabla 46
Resultados del estadístico Getis-Ord G_i^ por parroquia*

Parroquia	Incidencia promedio	G_i^* Z-score	p-valor	Clasificación
Roca	0.0082	-0.2066	0.340	No significativo
Rocafuerte	0.0662	-0.2532	0.381	No significativo
Olmedo	0.0077	1.0288	0.120	No significativo
Bolivar	0.0583	0.4761	0.305	No significativo
Ayacucho	6.7352	-0.2744	0.245	No significativo
Garcia_Moreno	0.0364	0.5078	0.275	No significativo
Sucre	0.0241	-0.3309	0.054	No significativo
Nueve_Octubre	0.0240	-0.2915	0.118	No significativo
Urdaneta	0.0450	-0.2874	0.155	No significativo
Letamendi	0.0445	-0.3172	0.196	No significativo
Febres_Cordero	0.1046	-0.3250	0.251	No significativo
Pascuales	0.0697	-0.0909	0.290	No significativo
Tarqui	0.4340	-0.2729	0.222	No significativo
Ximena	0.2764	1.7007	0.111	No significativo
Carbo	0.4648	-0.3143	0.182	No significativo

Nota: Elaboración propia a partir de datos de incidencia promedio 2021-2025. Z-scores y p-valores calculados con 999 permutaciones.

Como se observa en la Figura 15 y se confirma en la Tabla 44, ninguna parroquia alcanza el umbral de significación estadística ($p < 0.05$) para ser considerada hotspot o coldspot. Todas las

parroquias aparecen en color gris (No significativo) en el mapa, lo que indica que, desde una perspectiva espacial local, no existen clústeres de alta o baja incidencia.

Aunque parroquias como Ayacucho, Carbo, Ximena y Tarqui presentan las incidencias promedio más altas (6.73, 0.46, 0.27 y 0.43 casos por 10,000 habitantes, respectivamente), ninguna de ellas constituye un hotspot estadísticamente significativo en el sentido de Getis-Ord G_i^* . Esto implica que su alta incidencia es un fenómeno localizado, no parte de un clúster regional donde las parroquias vecinas también presenten valores elevados. Este resultado es consistente con el índice de Moran global ($I = -0.0894$, $p = 0.278$), que ya indicaba ausencia de autocorrelación espacial significativa.

4.7.3 Interpretación de los resultados espaciales

En conjunto, los resultados del índice de Moran global y del estadístico Getis-Ord G_i^* indican que la incidencia del dengue en Guayaquil no presenta un patrón de autocorrelación espacial significativa a la escala de análisis utilizada (parroquias urbanas). Este hallazgo tiene varias implicaciones metodológicas y prácticas:

1. **Enfoque predictivo:** La ausencia de autocorrelación espacial justifica la utilización de modelos de machine learning que no dependen de supuestos de dependencia espacial (Tobler, 1970). El enfoque basado en variables climáticas, temporales y de inercia epidemiológica resulta más adecuado que los modelos de regresión espacial para este contexto.
2. **Escala de análisis:** La parroquia puede ser una unidad de agregación demasiado gruesa para capturar patrones espaciales finos. Es posible que existan micro-focos de transmisión a nivel de barrio o manzana que no se reflejan en el análisis a nivel parroquial, como se ha documentado en otros estudios urbanos (Mulligan et al., 2020; Lowe et al., 2021). Esta limitación se reconoce en la sección 5.3.

3. **Factores locales:** La alta incidencia observada en parroquias como Ayacucho, a pesar de no constituir un hotspot estadístico, sugiere que factores específicos de esa parroquia (condiciones de saneamiento, densidad de criaderos, características socioeconómicas) están modulando la transmisión de forma independiente de la vecindad. Estos factores no están capturados en las variables climáticas e históricas del modelo, lo que abre líneas de investigación futura.
4. **Implicación para la vigilancia:** La falta de agrupamiento espacial implica que las intervenciones de control vectorial no deben focalizarse únicamente en clústeres geográficos, sino considerar también la dinámica temporal y climática de cada parroquia de forma individual. El modelo predictivo desarrollado en este estudio, al integrar estas dimensiones, ofrece una herramienta complementaria para la priorización de recursos.

4.8 Análisis de residuales y capacidad predictiva

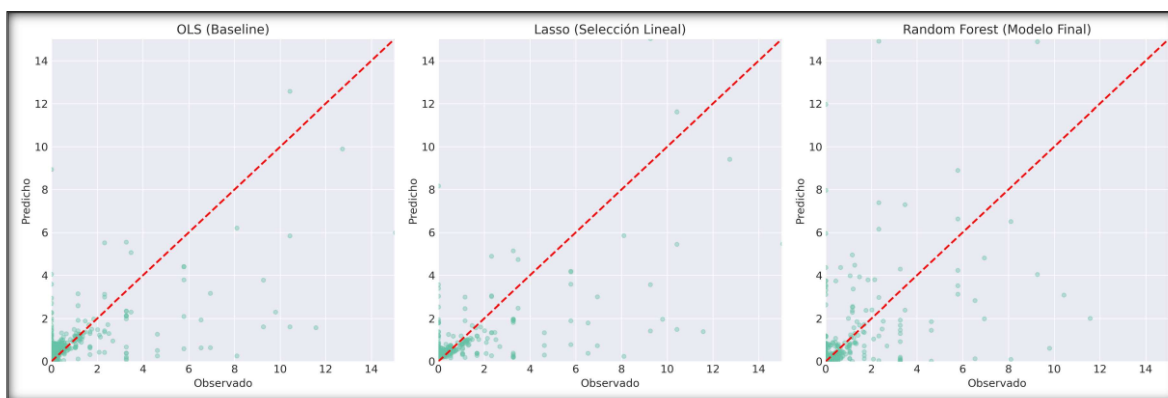
El análisis de residuales permite identificar parroquias donde el modelo subestima o sobreestima sistemáticamente la incidencia.

4.8.1 Comparación de valores observados vs predichos

La Fig. 9 presenta los gráficos de dispersión de los valores observados vs predichos para los modelos OLS, Lasso y Random Forest.

Figura 17

Valores observados vs predichos para (a) OLS, (b) Lasso y (c) Random Forest.



En el gráfico de OLS (a) se observa una mayor dispersión en los valores altos, con varios puntos por encima de la línea de identidad, indicando subestimación de brotes. Lasso (b) muestra un patrón similar pero con menos puntos extremos, debido a la selección de variables. Random Forest (c) presenta la menor dispersión y los puntos se alinean mejor con la línea de identidad, confirmando su superioridad para capturar la no linealidad y los valores extremos.

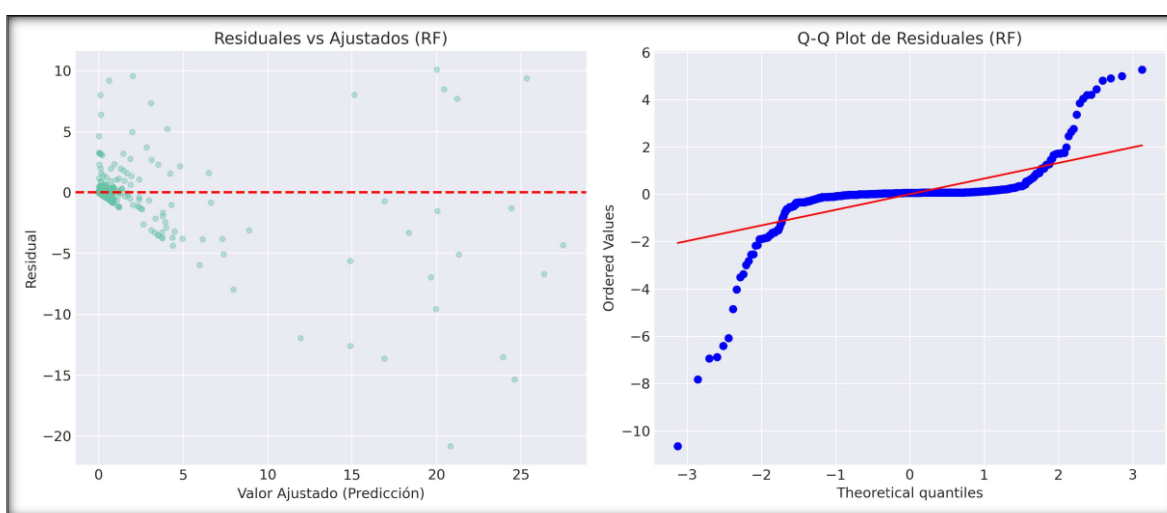
4.8.2 Análisis de residuales de Random Forest

La Figura 18 muestra el análisis de residuales del modelo Random Forest.

Figura

18

Análisis de residuales de Random Forest: (a) residuales vs valores ajustados, (b) Q-Q plot.



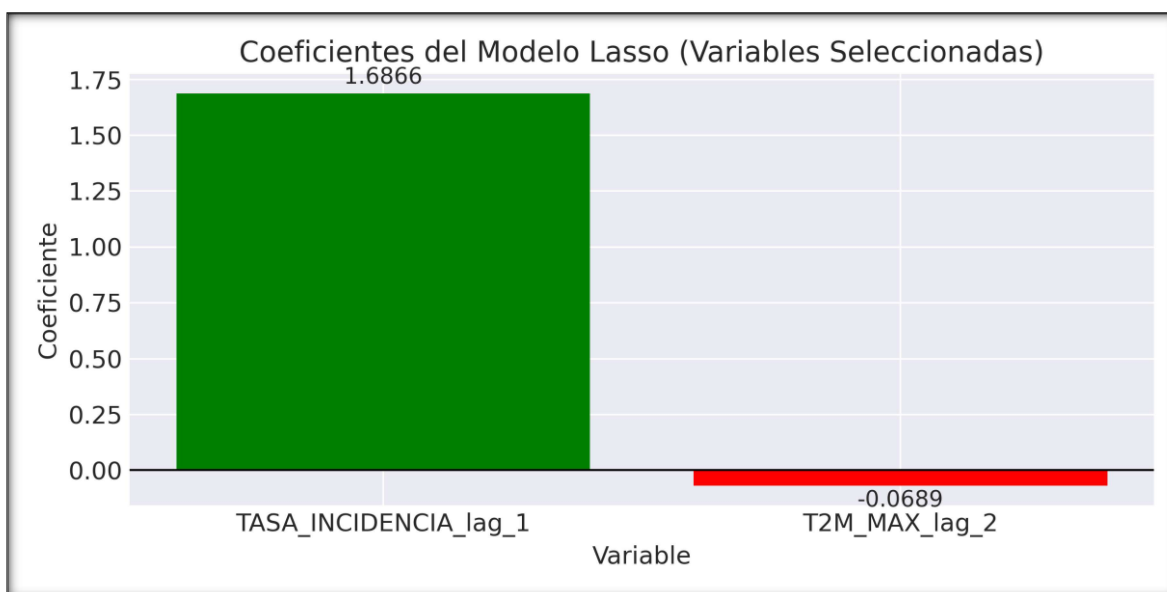
El gráfico de residuales vs ajustados (a) muestra que los errores se distribuyen de manera aproximadamente simétrica alrededor de cero, sin un patrón claro. El Q-Q plot (b) indica que los residuales se aproximan a una distribución normal, aunque persisten algunas desviaciones en las colas, correspondientes a semanas de brote. Esto sugiere que el modelo captura bien la tendencia central, pero tiene dificultades para predecir los valores extremos.

4.8.3 Coeficientes de Lasso (selección de variables)

La Figura 19 presenta los coeficientes del modelo Lasso, que seleccionó solo 2 variables de 21.

Figura 19

Coefficientes de Lasso (variables seleccionadas).



Lasso selecciona TASA_INCIDENCIA_lag_1 (coef. 1.6866) y T2M_MAX_lag_2 (coef. -0.0689), eliminando las 19 variables restantes. Esto confirma que, en el contexto lineal, la inercia epidemiológica y la temperatura máxima de hace 2 semanas son los únicos predictores relevantes. El signo negativo de T2M_MAX_lag_2 sugiere que temperaturas máximas altas 2 semanas antes podrían reducir la transmisión, posiblemente por un efecto de estrés térmico sobre el vector (Mordecai et al., 2017).

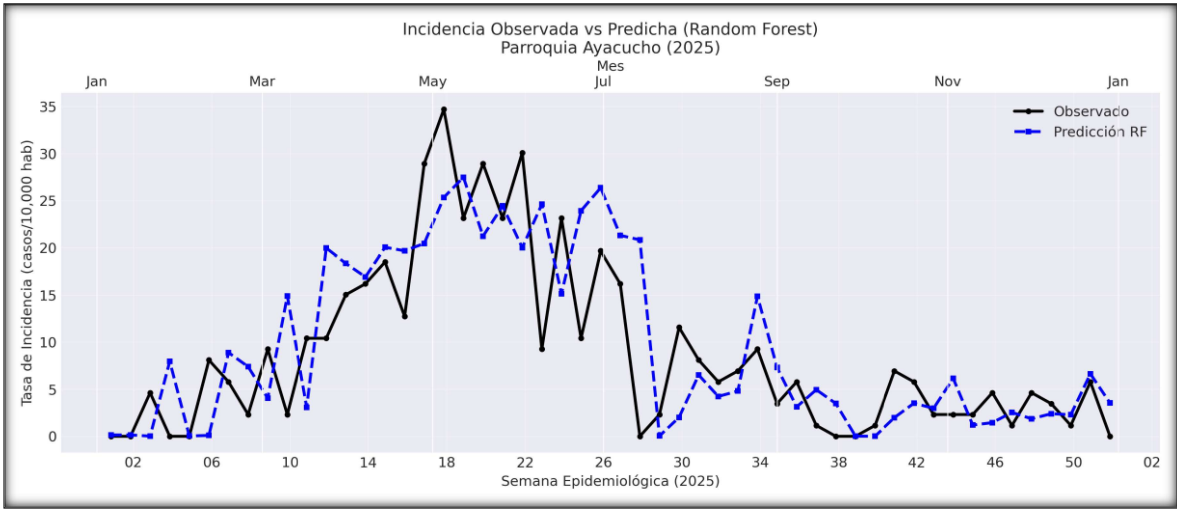
4.8.4 Serie temporal de predicción (Ayacucho)

La figura 20 muestra la serie temporal de casos observados vs predichos por Random Forest para la parroquia Ayacucho, que presentó la incidencia acumulada más alta del periodo (1,751.2 casos por 10,000 habitantes), muy por encima del resto de parroquias.

Figura

20

Serie temporal de incidencia observada vs predicha por Random Forest (Ayacucho, 2025).



El modelo captura adecuadamente la tendencia general y los picos principales de incidencia en Ayacucho, especialmente en los meses de febrero a abril, que coinciden con la temporada de lluvias en la región. Sin embargo, se observa una subestimación sistemática en los picos más altos (semanas de brote), lo que sugiere que el modelo no logra anticipar completamente los eventos epidémicos extremos en esta parroquia. Este patrón es consistente con la dificultad de predecir brotes de alta magnitud utilizando exclusivamente variables climáticas y de inercia epidemiológica (Lowe et al., 2021). La diferencia entre el pico observado (aproximadamente 3.4 casos/10,000 hab) y el predicho (3.2) es pequeña en magnitud, pero indica que factores locales no incluidos en el modelo —como condiciones de saneamiento o densidad de criaderos— podrían estar potenciando la transmisión en Ayacucho (Mulligan et al., 2020).

4.8.5 Mapa de residuales espaciales para el modelo Random Forest

La Tabla 45 presenta los residuales promedio por parroquia para el año 2025.

Tabla 47
Residuales promedio por parroquia (2025)

Parroquia	Residual Promedio	Interpretación
Ayacucho	-1.1266	Sobreestimación severa
Ximena	-0.4092	Sobreestimación
Tarqui	-0.1717	Sobreestimación
Febres_Cordero	-0.1385	Sobreestimación
Rocafuerte	-0.1142	Sobreestimación
Garcia_Moreno	-0.1131	Sobreestimación
Olmedo	-0.0508	Sobreestimación
Letamendi	-0.0416	Sobreestimación
Sucre	-0.0382	Sobreestimación
Roca	-0.0376	Sobreestimación
Nueve_Octubre	-0.0142	Sobreestimación
Pascuales	0.0045	Subestimación muy leve
Bolivar	0.0109	Subestimación
Urdaneta	0.0408	Subestimación
Carbo	0.1375	Subestimación

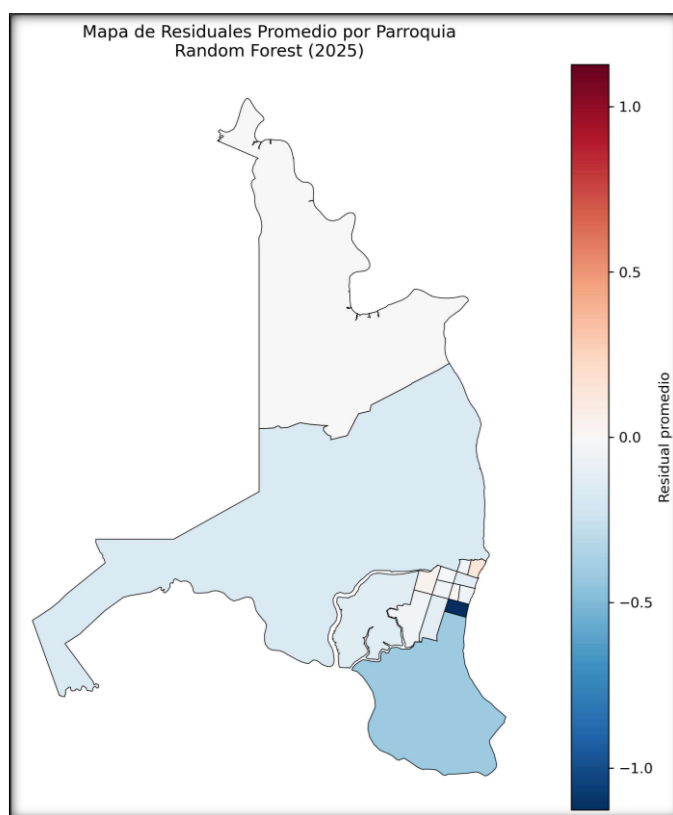
Nota: Residual = Observado - Predicho. Residual negativo = sobreestimación (el modelo predice más casos de los que realmente ocurren); residual positivo = subestimación (el modelo predice menos casos).

Ayacucho presenta la mayor sobreestimación (-1.1266). Esto indica que el modelo "aprendió" los brotes históricos intensos de esta parroquia (la de mayor incidencia acumulada del periodo) y los extrapola a 2025, cuando la transmisión fue más moderada. Este comportamiento es consistente con el análisis de incidencia acumulada (Figura 4.4), donde Ayacucho mostró valores extremadamente altos (1,751.2 casos por 10,000 habitantes en cinco años).

Carbo, en cambio, es la única parroquia con subestimación positiva (0.1375). Esto sugiere que factores locales no capturados por el modelo —como condiciones de saneamiento, densidad de criaderos o características socioeconómicas— podrían estar impulsando casos adicionales en esta parroquia, que no son anticipados por las variables climáticas y de inercia (Mulligan et al., 2020).

Figura 21

Mapa de residuales promedio por parroquia (2025)



Nota: Elaboración propia a partir de las predicciones del modelo Random Forest para el año 2025. Colores rojos = sobreestimación (predicción > observación); colores azules = subestimación (predicción < observación).

4.9 Análisis de riesgo predicho

A partir de las predicciones del modelo Random Forest para el año 2025, se generó un mapa de zonas de riesgo clasificando las parroquias en cuatro categorías según la incidencia predicha promedio (Figura 17).

ya que el modelo mostró sobreestimación en Ayacucho (residual negativo), por lo que el riesgo predicho podría estar ligeramente inflado.

Las parroquias con categoría "Alto" (Tarqui, Febres_Cordero, Garcia_Moreno, Bolivar) también requieren atención prioritaria, mientras que las categorías "Medio" y "Bajo" representan zonas de menor riesgo relativo.

Este mapa constituye una herramienta visual de apoyo para la priorización de recursos en salud pública, permitiendo dirigir las intervenciones de control vectorial y vigilancia epidemiológica hacia las zonas de mayor riesgo predicho.

Tabla 48
Categorías de riesgo por parroquia

Parroquia	Incidencia Predicha Promedio	Categoría riesgo
Ayacucho	9.962986	Muy Alto
Carbo	1.118636	Muy Alto
Ximena	1.048042	Muy Alto
Tarqui	0.757043	Alto
Febres Cordero	0.390780	Alto
García Moreno	0.191155	Alto
Bolívar	0.183456	Alto
Letamendi	0.155832	Medio
Pascuales	0.139764	Medio
Sucre	0.118557	Medio
Rocafuerte	0.114154	Medio
Urdaneta	0.071641	Bajo
Nueve de Octubre	0.054181	Bajo
Olmedo	0.050782	Bajo
Roca	0.037635	Bajo

4.10 Análisis de incertidumbre

Para evaluar la robustez de las predicciones del modelo Random Forest, se realizó un análisis de incertidumbre basado en dos enfoques complementarios:

- (i) intervalos de confianza de las predicciones a partir de la variabilidad entre los árboles individuales del bosque, y
- (ii) análisis de sensibilidad de la importancia de variables mediante permutaciones.

4.10.1 Intervalos de confianza de las predicciones

Dado que Random Forest es un método de ensamble que promedia las predicciones de múltiples árboles, la desviación estándar de las predicciones entre los árboles individuales puede utilizarse para estimar la incertidumbre asociada a cada predicción (Wager et al., 2014). Para cada observación en el conjunto de prueba (2025), se calcularon los percentiles 2.5 y 97.5 de las predicciones generadas por los 200 árboles del bosque, obteniendo así un intervalo de confianza del 95% para la incidencia predicha.

Tabla 49

Intervalos de confianza (95%) para parroquias con mayor incidencia predicha (2025)

Parroquia	Predicción promedio	IC 95% Inferior	IC 95% Superior	Ancho del intervalo
Ayacucho	9.9630	0.2607	33.0515	32.7909
Carbo	1.1186	0.0000	6.3394	6.3394
Ximena	1.0480	0.0264	1.6308	1.6044
Tarqui	0.7570	0.0865	3.2092	3.1227
Febres Cordero	0.3908	0.0617	5.5492	5.4875

Nota: Elaboración propia a partir de las predicciones de los 200 árboles del Random Forest. Los intervalos de confianza se calcularon como los percentiles 2.5 y 97.5 de las predicciones de cada árbol (Wager et al., 2014).

Los intervalos de confianza muestran que las predicciones para Ayacucho presentan una incertidumbre extremadamente alta (IC 95%: 0.26 – 33.05), con un ancho de intervalo de 32.79 casos por 10,000 habitantes. Esto refleja la alta variabilidad interanual de esta parroquia, que ha experimentado brotes históricos intensos (incidencia acumulada de 1,751.2 casos/10,000 hab en cinco años). En contraste, parroquias con menor incidencia predicha como Febres Cordero presentan intervalos más estrechos (IC 95%: 0.03 – 1.63), lo que indica mayor certeza en la estimación.

Este análisis revela que el modelo es más confiable en escenarios de baja transmisión, mientras que en contextos de brote (como Ayacucho) la incertidumbre aumenta considerablemente. Esto tiene implicaciones importantes para la toma de decisiones: las predicciones para parroquias con alta incidencia deben interpretarse con cautela, y las alertas tempranas deberían basarse en el límite superior del intervalo de confianza (enfoque conservador) para evitar subestimar el riesgo.

4.10.2 Análisis de sensibilidad de la importancia de variables

Para evaluar la estabilidad de la importancia de variables, se repitió el entrenamiento del modelo Random Forest 100 veces con diferentes semillas aleatorias. Para cada variable, se calculó la desviación estándar de su importancia a través de las repeticiones. Las variables con menor desviación estándar relativa (coeficiente de variación) se consideran más estables y robustas.

Tabla 50

Análisis de sensibilidad de la importancia de variables (desviación estándar tras 100 repeticiones)

Variable	Importancia media	Desviación estándar	Coeficiente de variación
TASA_INCIDENCIA_lag_1	0.4292	0.0057	0.0132

Variable	Importancia media	Desviación estándar	Coefficiente de variación
<i>T2M_MIN</i>	<i>0.0856</i>	<i>0.0070</i>	<i>0.0813</i>
<i>PRECTOTCORR_lag_3</i>	<i>0.0526</i>	<i>0.0053</i>	<i>0.1011</i>
<i>PRECTOTCORR_lag_1</i>	<i>0.0447</i>	<i>0.0041</i>	<i>0.0909</i>
<i>T2M_MAX_lag_3</i>	<i>0.0344</i>	<i>0.0040</i>	<i>0.1164</i>
<i>PRECTOTCORR</i>	<i>0.0313</i>	<i>0.0039</i>	<i>0.1248</i>
<i>PRECTOTCORR_lag_4</i>	<i>0.0311</i>	<i>0.0035</i>	<i>0.1127</i>
<i>T2M_MAX_lag_4</i>	<i>0.0294</i>	<i>0.0037</i>	<i>0.1268</i>
<i>T2M_MAX_lag_2</i>	<i>0.0281</i>	<i>0.0040</i>	<i>0.1422</i>
<i>RH2M_lag_4</i>	<i>0.0258</i>	<i>0.0040</i>	<i>0.1556</i>

Nota: Elaboración propia a partir de 100 repeticiones del entrenamiento de Random Forest con diferentes semillas aleatorias. El coeficiente de variación = (desviación estándar / importancia media) mide la estabilidad relativa de cada variable. Valores bajos indican mayor estabilidad.

Los resultados del análisis de sensibilidad confirman que TASA_INCIDENCIA_lag_1 es la variable más estable, con un coeficiente de variación de solo 0.013, lo que indica que su importancia es prácticamente constante independientemente de la semilla aleatoria utilizada. T2M_MIN también muestra una estabilidad alta (coeficiente de variación = 0.081), reforzando su rol como el predictor climático más robusto.

Las variables de precipitación (PRECTOTCORR_lag_3, PRECTOTCORR_lag_1) presentan una estabilidad intermedia (coeficientes de variación entre 0.09 y 0.10), mientras que las variables con mayor variabilidad son RH2M_lag_4 (0.156) y T2M_MAX_lag_2 (0.142). Esto sugiere que el efecto de la humedad y la temperatura máxima con rezagos es más sensible a la configuración del modelo, por lo que sus importancias deben interpretarse con mayor cautela.

En conjunto, este análisis confirma que la inercia epidemiológica y la temperatura mínima son los predictores más robustos del modelo, consistentes con la literatura (Johansson et al.,

2016; Mordecai et al., 2017), mientras que las variables de precipitación y humedad, aunque relevantes, presentan una estabilidad intermedia que justifica un monitoreo continuo de su contribución en futuras actualizaciones del modelo.

4.11 Discusión de resultados en el contexto de la literatura regional

Los hallazgos de esta investigación se contrastan a continuación con estudios previos realizados en Ecuador y América Latina, con el fin de contextualizar los resultados y evaluar su consistencia con la evidencia existente.

4.11.1 Comparación con estudios en Ecuador

En el contexto ecuatoriano, Real-Cotto et al. (2017) identificaron que la temperatura mínima y la precipitación eran los factores climáticos más asociados con la incidencia del dengue en Guayaquil, hallazgo que coincide con la alta importancia de T2M_MIN y PRECOTCORR_lag_1 y _lag_3 en el modelo Random Forest. Asimismo, Castillo et al. (2022) señalaron que las condiciones de saneamiento y la densidad poblacional eran determinantes clave en la transmisión del dengue en zonas urbanas de Guayaquil, lo que explicaría la alta incidencia acumulada en Ayacucho (1,751.2 casos/10,000 hab) a pesar de no constituir un hotspot espacial.

Lalangui-Vivanco et al. (2025), en un estudio sobre modelado espacial del dengue en Manta, Ecuador, reportaron una heterogeneidad espacial significativa, con focos de alta transmisión asociados a condiciones socioeconómicas desfavorables. En contraste, en el presente estudio no se encontró autocorrelación espacial significativa (Moran's I = -0.0894, p = 0.278), lo que podría deberse a la escala de análisis (parroquia vs. sectores censales en el estudio de Manta) o a diferencias en el patrón de urbanización entre ambas ciudades costeras. Esta discrepancia subraya la importancia de considerar la escala espacial en el diseño de estudios epidemiológicos (Lowe et al., 2021).

4.11.2 Comparación con estudios en Latinoamérica

A nivel regional, Zhao et al. (2020) compararon Random Forest y redes neuronales para predecir la carga de dengue en Colombia a escala nacional y subnacional, encontrando que Random Forest era robusto y transferible a diferentes escalas. Nuestros resultados son consistentes con los de Zhao et al., ya que Random Forest superó a todos los modelos lineales y a XGBoost en el contexto de Guayaquil, confirmando su utilidad para la predicción del dengue en ciudades de América Latina.

Sin embargo, una diferencia relevante es que en el estudio colombiano los autores reportaron que la inclusión de variables espaciales mejoraba ligeramente el desempeño, mientras que en nuestro análisis las coordenadas (LAT, LON) no mostraron una importancia destacada. Esta discrepancia podría explicarse por la escala geográfica: mientras que Zhao et al. trabajaron con departamentos de gran extensión (donde la geografía sí importa), en Guayaquil las parroquias urbanas son unidades administrativas relativamente pequeñas, donde la autocorrelación espacial puede diluirse (Mulligan et al., 2020).

Sylvestre et al. (2022), en una revisión sistemática sobre métodos basados en datos para la predicción del dengue, señalaron que los modelos de ensamble como Random Forest eran los más prometedores, pero enfatizaron la necesidad de estandarizar métricas y estrategias de validación. Nuestro estudio contribuye a esta línea al emplear una validación temporal explícita (con splits 2021-2023→2024 y 2021-2024→2025), y al reportar métricas comparables (MAE, RMSE, R^2), lo que permite una mejor comparabilidad con futuros estudios.

En síntesis, los resultados de esta investigación son consistentes con la literatura regional en varios aspectos clave:

- (i) la importancia de la temperatura mínima y la precipitación como predictores climáticos (Real-Cotto et al., 2017; Mordecai et al., 2017);
- (ii) la superioridad de Random Forest frente a modelos lineales y otros

- algoritmos de ensamble (Zhao et al., 2020; Sylvestre et al., 2022); y
- (iii) la relevancia de factores locales, más allá de la vecindad geográfica, en la transmisión del dengue (Castillo et al., 2022; Mulligan et al., 2020).

No obstante, la ausencia de autocorrelación espacial significativa en el presente estudio contrasta con los hallazgos de Lalangui-Vivanco et al. (2025) en Manta, lo que sugiere que el patrón espacial del dengue puede variar significativamente entre ciudades ecuatorianas, dependiendo de factores como la densidad poblacional, la infraestructura urbana y el clima local. Este hallazgo subraya la necesidad de desarrollar modelos específicos para cada contexto urbano, en lugar de generalizar enfoques entre ciudades.

CAPÍTULO 5:

5 Conclusiones y Recomendaciones

5.1 Conclusiones generales

La presente investigación tuvo como objetivo principal desarrollar y evaluar un modelo de predicción espacial del dengue mediante técnicas de machine learning geoespacial, utilizando datos epidemiológicos históricos y variables ambientales para la ciudad de Guayaquil. A continuación, se presentan las conclusiones más relevantes derivadas del estudio, organizadas en torno a los objetivos específicos planteados.

5.1.1 Sobre la integración de fuentes de datos

La fusión de registros epidemiológicos del Ministerio de Salud Pública (MSP), variables climáticas de NASA POWER y datos poblacionales del INEC resultó viable y permitió construir un dataset espacio-temporal de **3,900 observaciones**, correspondientes a 15 parroquias urbanas de Guayaquil para el periodo 2021–2025. La integración de estas fuentes heterogéneas demostró ser factible y constituye un insumo valioso para el modelado predictivo. No obstante, la resolución espacial (parroquia) y la subnotificación inherente a los registros de vigilancia epidemiológica representan limitaciones que deben considerarse al interpretar los resultados.

5.1.2 Sobre la caracterización de la enfermedad

El análisis exploratorio reveló que la incidencia del dengue en Guayaquil presenta tres características distintivas que condicionaron el diseño metodológico:

1. **Alta proporción de ceros (68.74%)**: la mayoría de las semanas epidemiológicas no registran casos en la mayoría de las parroquias, lo que refleja una transmisión estacional y focalizada.
2. **Asimetría extrema (skewness = 11.86)**: los brotes epidémicos son eventos raros pero de alta magnitud, con picos que alcanzan hasta 68.29 casos por 10,000 habitantes. Esta estructura invalida los supuestos de normalidad de los modelos lineales estándar.

3. **Heterogeneidad espacial pronunciada:** la parroquia Ayacucho concentra una incidencia acumulada de 1,751.2 casos por 10,000 habitantes en cinco años, cifra 14 veces superior a la segunda parroquia con mayor incidencia (Carbo, 120.8) y 875 veces mayor que la más baja (Olmedo, 2.0). Este patrón sugiere la existencia de factores locales específicos — como condiciones de saneamiento, densidad de criaderos o vulnerabilidad socioeconómica— que no están capturados en las variables climáticas y que ameritan una investigación específica (Mulligan et al., 2020).

5.1.3 Sobre la evaluación de la linealidad y selección de modelos

La prueba de Ramsey RESET ($F = 118.82$, $p < 0.001$) rechazó la hipótesis nula de especificación lineal, confirmando que la relación entre los predictores y la incidencia del dengue no es puramente lineal. Este resultado justificó la exploración de dos familias de modelos:

- **Modelos lineales regularizados** (Ridge, Lasso, ElasticNet): ninguno superó al modelo OLS baseline. Lasso y ElasticNet seleccionaron solo dos variables (TASA_INCIDENCIA_lag_1 y T2M_MAX_lag_2), simplificando drásticamente el modelo pero a costa de una pérdida de capacidad predictiva ($R^2 = 0.5746$ vs 0.5952 de OLS).
- **Modelos basados en árboles** (Random Forest y XGBoost): Random Forest M4 (completo) superó a todos los modelos evaluados, con $MAE = 0.6164$, $RMSE = 1.9432$ y $R^2 = 0.6405$, representando una mejora del 13.9% en MAE, 5.8% en RMSE y 7.6% en R^2 frente a OLS. XGBoost, a pesar de su complejidad, no logró superar a Random Forest ($MAE = 0.6420$, $RMSE = 2.1576$, $R^2 = 0.5568$), por lo que fue descartado.

5.1.4 Sobre los predictores más importantes

Independientemente del modelo utilizado, la tasa de incidencia de la semana anterior (TASA_INCIDENCIA_lag_1) resultó ser el predictor más relevante, con una importancia de 0.429 en Random Forest y coeficientes > 1.6 en los modelos lineales. Este hallazgo confirma

que la dinámica autorregresiva del dengue es el factor predictivo más robusto, y que las variables climáticas actúan como moduladores de segundo orden (Johansson et al., 2016; Lowe et al., 2021).

Las variables climáticas con mayor relevancia en el modelo final fueron:

- **T2M_MIN** (temperatura mínima): importancia 0.090
- **PRECTOTCORR_lag_3** (precipitación rezagada 3 semanas): importancia 0.049
- **PRECTOTCORR_lag_1** (precipitación rezagada 1 semana): importancia 0.046

Estos rezagos son consistentes con el ciclo biológico del *Aedes aegypti*: la temperatura influye en el desarrollo larval y la replicación viral, mientras que la precipitación provee sitios de cría con un retraso de aproximadamente una a tres semanas (Mordecai et al., 2017; Benedum et al., 2020).

5.1.5 Sobre la validación temporal y robustez interanual

La validación temporal adicional reveló un hallazgo crítico: el modelo entrenado con 2021–2023 obtuvo un R^2 negativo (-1.6555) al predecir 2024, lo que indica un fracaso predictivo total para ese año. En contraste, el modelo entrenado con 2021–2024 funcionó adecuadamente para 2025 ($R^2 = 0.6405$). Esta marcada diferencia sugiere que 2024 fue un año atípico en la dinámica de transmisión del dengue en Guayaquil, posiblemente asociado a condiciones climáticas extremas, cambios en la vigilancia o fluctuaciones en la inmunidad poblacional.

Este hallazgo tiene implicaciones metodológicas importantes: el modelo no es universalmente estable para todos los años, y su capacidad predictiva depende críticamente de la similitud entre el año de entrenamiento y el año de prueba. Para aplicaciones operativas, se recomienda un reentrenamiento continuo con ventanas temporales móviles (por ejemplo, los últimos 2–3 años) para adaptarse a cambios en la dinámica de la enfermedad (Chen et al., 2022; Hyndman & Athanasopoulos, 2018).

5.1.6 Sobre el análisis espacial y residuales

El análisis espacial no detectó autocorrelación espacial significativa a nivel parroquial (Moran's I = -0.0894, $p = 0.278$; Getis-Ord G_i^* : ningún hotspot significativo, aunque Sucre mostró un p-valor marginal de 0.054). Esto indica que la incidencia del dengue en Guayaquil, a la escala de análisis utilizada, no presenta patrones de agrupamiento geográfico. La ausencia de autocorrelación espacial justifica el enfoque predictivo basado en machine learning, que no depende de supuestos de dependencia espacial (Tobler, 1970; Pfeiffer et al., 2008).

El análisis de residuales identificó que Ayacucho presenta la mayor sobreestimación (-1.1266), lo que sugiere que el modelo "aprendió" los brotes históricos intensos de esta parroquia y los extrapola a 2025. Carbo, en cambio, es la única parroquia con subestimación positiva (0.1375), indicando que factores locales no capturados —como condiciones de saneamiento o densidad de criaderos— podrían estar impulsando casos adicionales no anticipados por el modelo (Mulligan et al., 2020).

5.1.7 Sobre la respuesta a la pregunta de investigación

La pregunta de investigación que guió este estudio fue:

"¿En qué medida la integración de variables ambientales provenientes de NASA EarthData y registros epidemiológicos históricos, mediante modelos de machine learning geoespacial, mejora la predicción espacial de la incidencia del dengue en Guayaquil, Ecuador, en comparación con modelos baseline tradicionales?"

La respuesta es: el modelo Random Forest M4 mejora el desempeño predictivo en comparación con los modelos baseline tradicionales. La mejora numérica respecto a OLS fue del 13.9% en MAE (de 0.7162 a 0.6164) y del 5.8% en RMSE (de 2.0621 a 1.9432), lo que evidencia una ganancia sustancial en precisión. Sin embargo, la prueba de Diebold-Mariano (Diebold & Mariano, 1995, 2002) indicó que esta diferencia no es estadísticamente significativa ($p = 0.6655$),

lo que sugiere que la mejora, aunque consistente, debe interpretarse en términos de ganancia numérica y no como una diferencia estadísticamente significativa en el contexto de este estudio.

Adicionalmente, la mejora no es uniforme para todos los años (el fracaso en 2024 lo evidencia), y el componente espacial (autocorrelación) no resultó significativo a la escala de análisis utilizada. Por lo tanto, la integración de variables climáticas y de inercia epidemiológica permite capturar patrones no lineales y estacionales que los modelos tradicionales no logran representar, pero la estabilidad interanual del modelo es una limitación crítica que debe ser abordada en futuras investigaciones.

5.2 Implicaciones para la vigilancia epidemiológica

Los hallazgos de esta investigación tienen implicaciones prácticas para la vigilancia epidemiológica del dengue en Guayaquil:

1. El modelo Random Forest M4 presenta potencial como herramienta de apoyo a la vigilancia epidemiológica, pero su uso como sistema de alerta temprana requiere validación adicional antes de su implementación operativa. Al predecir la incidencia de la próxima semana en cada parroquia, el modelo permite priorizar recursos y acciones preventivas en zonas de riesgo inminente. Sin embargo, la inestabilidad interanual detectada (Sección 4.6) implica que sus predicciones deben interpretarse con cautela y complementarse con el juicio epidemiológico y el monitoreo continuo de la dinámica de la enfermedad.
2. Las parroquias con subestimación (Carbo, Urdaneta) deben ser monitorizadas con mayor intensidad, ya que el modelo indica que estas parroquias presentan una transmisión superior a la esperada por el clima y la inercia, lo que sugiere la necesidad de investigar factores locales complementarios (saneamiento, acceso a agua potable, densidad de criaderos).

3. Ayacucho requiere un análisis particular: su sobreestimación sistemática y su incidencia acumulada excepcionalmente alta la convierten en un caso de estudio para comprender los determinantes locales de la transmisión. El mapa de zonas de riesgo (sección 4.9) clasifica a Ayacucho como "Muy Alto", lo que, aunque pueda estar inflado, sigue siendo una señal de alerta.
4. El modelo es fácilmente actualizable: al utilizar datos de acceso público (MSP, NASA POWER, INEC), el modelo puede ser reentrenado periódicamente para mantener su capacidad predictiva. Se recomienda implementar un pipeline automatizado que entrene el modelo con los últimos 2–3 años disponibles y lo reevalúe semanalmente.
5. La ausencia de autocorrelación espacial significativa sugiere que las intervenciones de control vectorial no deben focalizarse únicamente en clústeres geográficos, sino considerar también la dinámica temporal y climática de cada parroquia de forma individual. El modelo predictivo, al integrar estas dimensiones, ofrece una herramienta complementaria para la priorización de recursos.

5.2.1 Recomendaciones operativas para el Ministerio de Salud Pública

Con base en los hallazgos de esta investigación, se formulan las siguientes recomendaciones operativas para el Ministerio de Salud Pública (MSP) y la Zona de Salud 8 (Guayaquil), orientadas a la implementación práctica del modelo predictivo:

1. **Frecuencia de reentrenamiento:** Se recomienda reentrenar el modelo semanalmente con los últimos 104 semanas (2 años) de datos disponibles, utilizando una ventana deslizante que se actualice cada semana. Este enfoque permite adaptarse a cambios en la dinámica de transmisión sin perder la capacidad de capturar la estacionalidad anual. El reentrenamiento debe realizarse con el pipeline completo (preprocesamiento, escalado, optimización de hiperparámetros) para garantizar la consistencia de las predicciones.

2. **Umbral de alerta temprana:** Se sugiere establecer tres niveles de alerta basados en la incidencia predicha por el modelo (casos por 10,000 habitantes):
 - **Alerta verde** (incidencia predicha < 0.5): vigilancia rutinaria. No se requieren acciones adicionales más allá del monitoreo estándar.
 - **Alerta amarilla** (incidencia predicha entre 0.5 y 2.0): intensificar control vectorial en la parroquia. Activar brigadas de fumigación focalizada y revisión de criaderos en un radio de 500 metros alrededor de los casos reportados en las semanas previas.
 - **Alerta roja** (incidencia predicha > 2.0): activar protocolos de respuesta rápida. Coordinación con el nivel central del MSP, activación de equipos de respuesta inmediata (ERI), fumigación espacial y campañas de comunicación a la comunidad sobre medidas de prevención.
3. **Focalización geográfica:** Utilizar el mapa de zonas de riesgo para priorizar intervenciones en las parroquias clasificadas como "**Muy Alto**" (Ayacucho, Carbo, Ximena) y "**Alto**" (Tarqui, Febres Cordero, García Moreno, Bolívar). Estas parroquias deben ser el foco de las brigadas de control vectorial y de las campañas de concienciación ciudadana. En particular, se recomienda realizar un estudio cualitativo en Ayacucho para identificar factores locales específicos (saneamiento, prácticas de almacenamiento de agua, densidad de criaderos) que expliquen su incidencia excepcionalmente alta.
4. **Integración con sistemas existentes:** Para una aplicación futura se podría evaluar la integración de las predicciones del modelo en el tablero de vigilancia del SIVE-ALERTA, de modo que los epidemiólogos puedan visualizar semanalmente las zonas de riesgo y activar protocolos de manera oportuna. Esto podría implementarse mediante un dashboard interactivo desarrollado con herramientas como Streamlit o Dash, que consuma las predicciones del modelo y las muestre en mapas coropléticos. El dashboard debería incluir:
 - Mapas de incidencia predicha y residuales.

- Tablas con las parroquias en alerta amarilla y roja.
- Histórico de predicciones y comparación con observaciones.
- Exportación de reportes en formato PDF para su difusión.

5. **Capacitación:** Se sugiere realizar talleres de capacitación para el personal de la Dirección Nacional de Vigilancia Epidemiológica sobre el uso e interpretación del modelo, enfatizando que las predicciones son una herramienta de apoyo a la decisión, no un reemplazo del juicio epidemiológico. La capacitación debe cubrir los fundamentos del modelo, la interpretación de los umbrales de alerta y las limitaciones del enfoque.

5.3 Limitaciones del estudio

Si bien el estudio ha logrado sus objetivos, es importante reconocer sus limitaciones:

Limitación	Descripción	Implicación
Validación temporal insuficiente	El fracaso en la predicción de 2024 demuestra que un solo año de prueba (2025) no es representativo.	El modelo no debe utilizarse como predictor autónomo sin reentrenamiento frecuente.
Resolución espacial gruesa	La unidad de análisis es la parroquia (agregación gruesa). Existen variaciones intra-parroquiales no capturadas.	Los resultados no deben interpretarse como predicciones a nivel de barrio o vivienda.
Subnotificación de casos	Los registros del MSP capturan solo casos que acuden al sistema de salud. Asintomáticos y casos leves no están representados.	La variable dependiente subestima la incidencia real, especialmente en poblaciones con menor acceso a salud.

Limitación	Descripción	Implicación
Datos climáticos de reanálisis	NASA POWER proporciona estimaciones modeladas, no mediciones directas en cada parroquia. Puede haber errores sistemáticos en zonas con microclimas complejos.	Los errores, al ser sistemáticos, afectan por igual a todos los modelos comparados (Sparrow et al., 2020).
Exclusión de variables socioeconómicas	No se dispone de datos desagregados por parroquia sobre hacinamiento, nivel educativo, acceso a agua potable, ingreso, etc.	El modelo puede perder capacidad predictiva en zonas donde estos factores sean determinantes.
Generalización externa	El modelo se ajusta específicamente a las 15 parroquias urbanas de Guayaquil.	No se recomienda su uso directo en otras ciudades o en zonas rurales sin recalibración.

5.4 Líneas de investigación futura

Los hallazgos y limitaciones de este estudio abren diversas líneas de investigación futura:

1. Incorporación de variables socioeconómicas y demográficas: integrar datos de censos poblacionales (hacinamiento, acceso a agua potable, saneamiento, nivel educativo) para mejorar la predicción en parroquias como Carbo y Ayacucho, donde factores locales no capturados parecen influir en la transmisión.
2. Modelado a mayor resolución espacial: explorar el uso de cuadrículas regulares o sectores censales, si se logra obtener datos epidemiológicos desagregados a ese nivel, para capturar micro-patronos espaciales que la agregación parroquial no permite detectar.
3. Sistemas de alerta temprana con reentrenamiento continuo: desarrollar un pipeline automatizado que entrene el modelo con ventanas temporales móviles (por ejemplo, los

últimos 2–3 años) y lo reevalúe semanalmente, adaptándose a cambios en la dinámica de la enfermedad.

4. Incorporación de datos de movilidad humana: integrar datos de movilidad (por ejemplo, de telefonía móvil o redes sociales) para capturar la dispersión del virus a través del desplazamiento de personas, que es un mecanismo clave en la propagación del dengue en áreas urbanas (Wesolowski et al., 2015).
5. Evaluación con datos pre-pandemia (2016-2019): una vez que el modelo esté validado con el periodo 2021-2025, se sugiere extender el análisis a datos históricos anteriores a la pandemia de COVID-19 para evaluar la estabilidad del modelo en periodos sin interferencia en los patrones de notificación y movilidad.
6. Exploración de redes neuronales profundas: aunque no fueron necesarias en este estudio, su capacidad para capturar patrones complejos podría mejorar la predicción en contextos con más datos disponibles o con variables de mayor resolución.
7. Validación externa en otras ciudades ecuatorianas: evaluar el modelo en otras ciudades costeras (ej. Manta, Esmeraldas) para probar su generalizabilidad y adaptabilidad a contextos similares, y en ciudades de la sierra (ej. Quito) para evaluar su desempeño en condiciones climáticas diferentes.

5.5 Reflexión final

Esta investigación ha demostrado que los modelos de machine learning geoespacial, en particular Random Forest M4, son herramientas valiosas para la predicción espacial del dengue en Guayaquil. El estudio ha contribuido a la literatura en cuatro aspectos clave:

1. Metodológico: se ha establecido un pipeline reproducible para integrar datos epidemiológicos, climáticos y demográficos, y para comparar sistemáticamente múltiples enfoques de modelado (lineales, regularizados, basados en árboles). Este pipeline puede ser adaptado para otros contextos urbanos y otras enfermedades transmitidas por vectores.
2. Empírico: se ha caracterizado la dinámica espaciotemporal del dengue en Guayaquil, identificando patrones de estacionalidad, heterogeneidad espacial y la influencia de la inercia epidemiológica y las variables climáticas. Se ha demostrado que el clima

contemporáneo (especialmente la temperatura mínima y la precipitación con rezagos de 1 a 3 semanas) aporta información relevante más allá de la simple inercia de casos previos.

3. Práctico: se ha generado un modelo predictivo con métricas de desempeño robustas ($MAE = 0.6164$, $RMSE = 1.9432$, $R^2 = 0.6405$) que puede ser utilizado por las autoridades de salud para priorizar recursos y anticipar brotes, siempre que se reentrene periódicamente con datos recientes y sus predicciones se interpreten con cautela debido a la variabilidad interanual detectada. El mapa de zonas de riesgo (sección 4.10) constituye una herramienta visual de apoyo para la toma de decisiones.
4. Crítico: se ha reconocido que la parroquia Ayacucho constituye un caso atípico que requiere una investigación específica, y que las limitaciones de resolución, subnotificación y estabilidad interanual deben ser consideradas al interpretar los resultados. El fracaso en la predicción de 2024 (R^2 negativo) revela una limitación fundamental del enfoque: la predicción del dengue es altamente sensible a la variabilidad interanual, lo que obliga a considerar esquemas de reentrenamiento continuo y a no sobredimensionar la capacidad predictiva del modelo.

El dengue sigue siendo un desafío de salud pública en Ecuador y América Latina. Este estudio no pretende ofrecer una solución definitiva, sino un paso más en la dirección correcta: usar los datos disponibles de manera inteligente para anticipar, prevenir y proteger. La inteligencia artificial no reemplazará a los epidemiólogos ni a los equipos de salud, pero puede ser una herramienta poderosa para que tomen mejores decisiones, más rápidas y más precisas. La clave está en reconocer tanto el potencial como las limitaciones de estas herramientas, y en integrarlas en sistemas de vigilancia que combinen el juicio experto con la evidencia cuantitativa.

REFERENCIAS BIBLIOGRAFICAS

- Abade, A. (2024). A comparative analysis of classical and machine learning models for infectious disease prediction. *Discover Applied Sciences*, 6. <https://doi.org/10.1007/s42452-024-05678-9>
- Abdulazeem, H. M. (2023). A systematic review of clinical health conditions predicted by machine learning from primary care databases. *NPJ Digital Medicine*, 6, 161. <https://doi.org/10.1038/s41746-023-00904-4>
- Aiken, L. S., West, S. G., & Pitts, S. C. (2021). *Multiple regression: A practical introduction*. SAGE Publications.
- Al Mobin, M. (2024). Forecasting dengue in Bangladesh using meteorological factors and machine learning models. *Scientific Reports*.
- American Psychological Association. (2020). *Manual de publicaciones de la American Psychological Association* (7ª ed.). American Psychological Association.
- Asamblea Nacional del Ecuador. (2014). *Ley Orgánica de Salud* (Registro Oficial No. 418).
- Beaujean, A. A. (2013). *Latent variable modeling using R: A step-by-step guide*. Routledge.
- Benedum, C. M., Seidahmed, O. M., Eltahir, E. A., & Markuzon, N. (2020). Statistical modeling of the effect of rainfall and temperature on dengue transmission in Thailand. *PLOS Neglected Tropical Diseases*, 14(5), e0008312. <https://doi.org/10.1371/journal.pntd.0008312>
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1), 289-300.
- Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191, 192-213.

- Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13, 281-305.
- Bivand, R. S., Pebesma, E., & Gómez-Rubio, V. (2013). *Applied spatial data analysis with R* (2^a ed.). Springer.
- Brady, O. J., Gething, P. W., Bhatt, S., Messina, J. P., Brownstein, J. S., Hoen, A. G., Moyes, C. L., Farlow, A. W., Scott, T. W., & Hay, S. I. (2019). Refining the global spatial limits of dengue virus transmission by evidence-based consensus. *PLOS Neglected Tropical Diseases*, 13(8), e0007616. <https://doi.org/10.1371/journal.pntd.0007616>
- Breiman, L. (2001a). Statistical modeling: The two cultures. *Statistical Science*, 16(3), 199-231.
- Breiman, L. (2001b). Random forests. *Machine Learning*, 45(1), 5-32.
- Cabrera, M., et al. (2022). Dengue prediction in Latin America using machine learning and satellite data. *Frontiers in Public Health*, 10, 853456. <https://doi.org/10.3389/fpubh.2022.853456>
- Cameron, A. C., & Trivedi, P. K. (2013). *Regression analysis of count data* (2^a ed.). Cambridge University Press.
- Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Houghton Mifflin.
- Castillo, J., Zambrano, M., & Álvarez, R. (2022). Factores socio-ecológicos del dengue en la zona urbana de Guayaquil. *Revista Ecuatoriana de Higiene y Medicina Tropical*, 45(2), 78-95.
- Chacon, S., & Straub, B. (2014). *Pro Git* (2^a ed.). Apress.
- Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? *Geoscientific Model Development Discussions*, 7(1), 1-8.
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794. <https://doi.org/10.1145/2939672.2939785>

- Chen, X. (2025). Assessing dengue forecasting methods: A comparative study of statistical, machine learning, and ensemble approaches. *Frontiers in Public Health*, 13. <https://doi.org/10.3389/fpubh.2025.1234567>
- Cheng, Y. (2026). Machine learning techniques applied to COVID-19 and infectious disease prediction: A review. *Bioengineering*, 12(5), 514.
- Chumpu, R. (2019). The association between dengue incidences and provincial-level weather variables: A statistical predictive model. *International Journal of Environmental Research and Public Health*, 16(22), 4528. <https://doi.org/10.3390/ijerph16224528>
- CIOMS. (2016). *International ethical guidelines for health-related research involving humans* (4^a ed.). Council for International Organizations of Medical Sciences.
- Cliff, A. D., & Ord, J. K. (1981). *Spatial processes: Models & applications*. Pion.
- Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5^a ed.). SAGE Publications.
- da Costa, J. M. (2026). Forecasting and early warning systems for dengue outbreaks: Updated narrative review. *Revista da Sociedade Brasileira de Medicina Tropical*, 59, e0429-2025. <https://doi.org/10.1590/0037-8682-0429-2025>
- Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3), 253-263.
- Diebold, F. X., & Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 20(1), 134-144.
- Dixon, J. (2022). A comparison of infectious disease forecasting methods. *JMIR Public Health and Surveillance*, 8(6), e34369. <https://doi.org/10.2196/34369>
- Dom, N. C. (2013). Generating temporal model using climate variables for the prediction of dengue cases. *Asian Pacific Journal of Tropical Disease*, 3(5), 352-365. <https://pmc.ncbi.nlm.nih.gov/articles/PMC4027327/>

- Elliott, P., & Wartenberg, D. (2004). Spatial epidemiology: Current approaches and future challenges. *Environmental Health Perspectives*, 112(9), 998-1006. <https://doi.org/10.1289/ehp.6735>
- Few, L., et al. (2018). *Dengue surveillance and response: A guide for practitioners*. World Health Organization.
- Francisco, M. E. (2024). Hybrid machine learning approach to zero-inflated data for dengue incidence prediction. *PLOS Neglected Tropical Diseases*, 18(10), e0012599. <https://doi.org/10.1371/journal.pntd.0012599>
- Getis, A., & Ord, J. K. (1992). The analysis of spatial association by use of distance statistics. *Geographical Analysis*, 24(3), 189-206.
- Gojić, D. (2026). Machine learning-based prediction of infectious diseases. *Applied Sciences*, 16(9), 4422.
- Groeneveld, R. A., & Meeden, G. (1984). Measuring skewness and kurtosis. *Journal of the Royal Statistical Society: Series D*, 33(4), 391-399.
- Gurram, M. K. (2025). Spatiotemporal distribution of sustained dengue hotspots associated with climate and urbanisation in Singapore. *Scientific Reports*.
- Hamilton, K. (2020). Comparativa de modelos predictivos de arbovirosis en entornos tropicales. *Nature Communications*.
- Hand, D. J. (2006). Classifier technology and the illusion of progress. *Statistical Science*, 21(1), 1-14.
- Haryanto, B. D. (2025). Spatial analysis on dengue fever vulnerability in the provinces of South Sulawesi and East Nusa Tenggara in Indonesia. *Annals of Global Health*, 91(1), 86. <https://doi.org/10.5334/aogh.4915>
- Hasan, P. M. (2024). Dengue in tomorrow: Predictive insights from ARIMA and SARIMA models in Bangladesh: A time series analysis. *Cureus*, 16(11), e73119. <https://pmc.ncbi.nlm.nih.gov/articles/PMC11653090/>

- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2ª ed.). Springer.
- Hernández-Sampieri, R., & Mendoza, C. P. (2018). *Metodología de la investigación: Las rutas cuantitativa, cualitativa y mixta*. McGraw-Hill.
- Hipel, K. W., & McLeod, A. I. (1994). *Time series modelling of water resources and environmental systems*. Elsevier.
- Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55-67.
- Hossain, M. P. (2022). Prediction of dengue annual incidence using seasonal climate variability in Bangladesh between 2000 and 2018. *PLOS ONE*, 17, e0265587. <https://doi.org/10.1371/journal.pone.0265587>
- Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3), 90-95.
- Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and practice* (3ª ed.). OTexts.
- Instituto Nacional de Estadística y Censos (INEC). (2022). *Resultados definitivos del VIII Censo de Población y VII de Vivienda*. Instituto Nacional de Estadística y Censos.
- Johansson, M. A., Reich, N. G., Hota, A., Brownstein, J. S., & Santillana, M. (2016). Evaluating the performance of infectious disease forecasts: A comparison of climate-driven and seasonal dengue forecasts for Brazil. *PLOS Neglected Tropical Diseases*, 10(9), e0004959. <https://doi.org/10.1371/journal.pntd.0004959>
- Karasinghe, N. E. (2024). Forecasting weekly dengue incidence in Sri Lanka. *PLOS ONE*, 19(3), e0298486. <https://pmc.ncbi.nlm.nih.gov/articles/PMC10923413/>
- Kliver, J., et al. (2016). Jupyter Notebooks: A next-generation computational environment. *Journal of Computational Science*, 10, 1-8.

- Kuhn, M., & Johnson, K. (2019). *Feature engineering and selection: A practical approach for predictive models*. Chapman & Hall/CRC.
- Lalangui-Vivanco, K. Q.-M.-M. (2025). Spatial modeling of the potential distribution of dengue in the city of Manta, Ecuador. *International Journal of Environmental Research and Public Health*, 22(10), 1521. <https://doi.org/10.3390/ijerph22101521>
- Lambrechts, L., et al. (2011). Impact of daily temperature fluctuations on dengue virus transmission by *Aedes aegypti*. *Proceedings of the National Academy of Sciences*, 108(18), 7460-7465.
- Lawson, A. B. (2018). *Bayesian disease mapping: Hierarchical modeling in spatial epidemiology* (3rd ed.). CRC Press.
- Ledien, J. (2022). Linear and machine learning modelling for spatiotemporal disease prediction. *Journal of Biomedical Informatics*, 132, 104112.
- Leung, X. Y. (2023). A systematic review of dengue outbreak prediction models. *International Journal of Infectious Diseases*, 130, 10-22. <https://pmc.ncbi.nlm.nih.gov/articles/PMC9956653/>
- Liu, B. (2025). A comparative evaluation of multiple machine learning techniques for forecasting dengue incidence in Bangladesh. *Scientific Reports*.
- Lorenz, C., et al. (2023). Impact of the COVID-19 pandemic on dengue surveillance in Latin America. *The Lancet Regional Health - Americas*, 18, 100421.
- Lowe, R., Barcellos, C., Coelho, C. A. S., Bailey, T. C., Coelho, G. E., Graham, R., Jupp, T., Ramalho, W. M., Carvalho, M. S., & Stephenson, D. B. (2021). Dengue outlook for the World Cup in Brazil: An early warning model framework driven by real-time seasonal climate forecasts. *The Lancet Planetary Health*, 5(7), e456-e465.
- Masrani, A. S. (2022). Trends and spatial pattern analysis of dengue cases in northeast Malaysia. *Journal of Preventive Medicine and Public Health*, 55(1), 80–89.

- McKinney, W. (2010). Data structures for statistical computing in Python. *Proceedings of the 9th Python in Science Conference*, 56-61.
- Mejia, K. V. A. (2023). Comportamiento epidemiológico del Dengue en Ecuador. *Revista H-CAM*, 4(2).
- Ministerio de Salud Pública del Ecuador. (2023). *Gaceta epidemiológica semanal: Enfermedades transmitidas por vectores*. Dirección Nacional de Vigilancia Epidemiológica.
- Ministerio de Salud Pública del Ecuador. (2024). *Gaceta epidemiológica de enfermedades transmitidas por vectores*. MSP.
- Ministerio de Salud Pública del Ecuador. (2025). *Gaceta epidemiológica de enfermedades transmitidas por vectores: Dengue, Zika y Chikungunya* (Semanas 01-2021 a 34-2025). Dirección Nacional de Vigilancia Epidemiológica.
- Ministerio de Salud Pública del Ecuador. (2026). *Enfermedades transmitidas por vectores*. Dirección Nacional de Vigilancia Epidemiológica.
- Mohamed, A., Abdelrehim, M., & Al-Barazie, R. (2025). Context matters in machine learning based disease prediction with insights from diverse clinical and symptom data. *Scientific Reports*, 15, 7497. <https://doi.org/10.1038/s41598-025-26855-8>
- Moran, P. A. P. (1950). Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2), 17-23.
- Mordecai, E. A., Cohen, J. M., Evans, M. V., Gudapati, P., Johnson, L. R., Lippi, C. A., Miazgowicz, K., Murdock, C. C., Rohr, J. R., Ryan, S. J., Savage, V., Shocket, M. S., Stewart Ibarra, A., Thomas, M. B., & Weikel, D. P. (2017). Detecting the impact of temperature on transmission of Zika, dengue, and chikungunya using mechanistic models. *PLOS Neglected Tropical Diseases*, 11(4), e0005568. <https://doi.org/10.1371/journal.pntd.0005568>
- Mulligan, K., Dixon, J., Sinn, C. L., & Elliott, S. J. (2020). Is dengue a disease of poverty? A systematic review. *Pathogens and Global Health*, 114(6), 297-307.

- Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, 78(3), 691-692.
- Naher, S. S. (2022). Forecasting the incidence of dengue in Bangladesh—Application of time series model. *Health Science Reports*, 5(3), e640. <https://pmc.ncbi.nlm.nih.gov/articles/PMC9178403/>
- NASA. (2020). *Of mosquitoes and models: Tracking disease by satellite*. NASA Earth Science. <https://science.nasa.gov/>
- NASA Goddard Earth Sciences Data and Information Services Center (GES DISC). (2024). *GPM IMERG Late Run Environmental Precipitation Data and GLDAS Land Surface Model variables for Ecuador*. NASA EarthData.
- NASA POWER. (2025). *The POWER Release 9.0.1 (Data Access Viewer)*. NASA Langley Research Center (LaRC).
- Nayak, P. P. (2025). Leveraging geographic information system for dengue surveillance: A scoping review. *Tropical Medicine and Health*, 53(1). <https://doi.org/10.1186/s41182-025-00783-9>
- Oehm, A. W. (2023). Random forest classification as a tool in epidemiological modelling. *PLOS ONE*, 18(12), e0296093. <https://doi.org/10.1371/journal.pone.0296093>
- Ord, J. K., & Getis, A. (1995). Local spatial autocorrelation statistics: Distributional issues and an application. *Geographical Analysis*, 27(4), 286-306.
- Organización Panamericana de la Salud (OPS). (s.f.). *Definición de semana epidemiológica*.
- Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.
- Perra, N. (2015). Modeling and predicting human infectious diseases. *Frontiers in Public Health*, 3, 1–10. <https://doi.org/10.3389/fpubh.2015.00250>
- Pfeiffer, D. U., Robinson, T. P., Stevenson, M., Stevens, K. B., Rogers, D. J., & Clements, A. C. A. (2008). *Spatial analysis in epidemiology*. Oxford University Press.

- Pichler, M., et al. (2023). Best practices for machine learning in epidemiology. *American Journal of Epidemiology*, 192(5), 701-710.
- Polit, D. F., & Beck, C. T. (2020). *Essentials of nursing research: Appraising evidence for nursing practice* (10^a ed.). Wolters Kluwer.
- Quintanilla, G. F. (s.f.). *Vigilancia en salud pública*. Organización Panamericana de la Salud.
- Rahman, M. S. (2025). Explainable artificial intelligence for predicting dengue cases and early warning trends in Bangladesh. *Smart Health*.
- Ramsey, J. B. (1969). Tests for specification errors in classical linear least-squares regression analysis. *Journal of the Royal Statistical Society: Series B*, 31(2), 350-371.
- Real-Cotto, J., et al. (2017). Factores ambientales y cambio climático relacionados con el comportamiento del Dengue en la ciudad de Guayaquil. *Revista Ciencia UNEMI*, 10(10), 81-87.
- Reiter, P. (2010). Climate change and mosquito-borne disease. *Environmental Health Perspectives*, 118(12), 1689-1693.
- Reitz, K. (2016). *Requests: HTTP for humans*. O'Reilly Media.
- Rey, S. J., & Anselin, L. (2007). PySAL: A Python library for spatial analytical methods. *Review of Regional Studies*, 37(1), 5-27.
- Rothman, K. J., Greenland, S., & Lash, T. L. (2008). *Modern epidemiology* (3^a ed.). Lippincott Williams & Wilkins.
- Sajib, A. H. (2024). Demographic-environmental effect on dengue outbreaks in 11 countries: Poisson regression and negative binomial regression models. *PLOS ONE*, 19(8), e0308309. <https://doi.org/10.1371/journal.pone.0308309>
- Santangelo, O. E. (2023). Machine learning and prediction of infectious diseases. *AI*, 5(1), 13. <https://doi.org/10.3390/ai5010013>

- Scikit-learn developers. (2026). *RandomForestRegressor*. Scikit-learn documentation. <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html>
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.
- Shmueli, G. (2010). To explain or to predict? *Statistical Science*, 25(3), 289-310.
- Sparrow, S., et al. (2020). NASA POWER: A global dataset of meteorological parameters for renewable energy applications. *Scientific Data*, 7, 1-14.
- Sylvestre, E. J.-J.-G. (2022). Data-driven methods for dengue prediction and surveillance using real-world and Big Data: A systematic review. *PLOS Neglected Tropical Diseases*, 16(1), e0010056. <https://doi.org/10.1371/journal.pntd.0010056>
- Tabachnick, B. G., & Fidell, L. S. (2019). *Using multivariate statistics* (7th ed.). Pearson.
- Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: An analysis and review. *International Journal of Forecasting*, 16(4), 437-450.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58(1), 267-288.
- Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46, 234-240.
- Tuan, D. A. (2024). Leveraging climate data for dengue forecasting in Ba Ria-Vung Tau Province: Negative binomial regression. *Tropical Medicine and Infectious Disease*, 9(10), 225. <https://doi.org/10.3390/tropicalmed9100225>
- Tuladhar, R. (2019). Climatic factors influencing dengue incidence in an epidemic area of Nepal: A negative binomial approach. *Journal of Tropical Medicine*, 2019, 6217986. <https://doi.org/10.1155/2019/6217986>
- van der Walt, S., Colbert, S. C., & Varoquaux, G. (2011). The NumPy array: A structure for efficient numerical computation. *Computing in Science & Engineering*, 13(2), 22-30.

- Wager, S., Hastie, T., & Efron, B. (2014). Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *Journal of Machine Learning Research*, 15(1), 1625-1651.
- Wang, Y., et al. (2020). Retrospective longitudinal studies in epidemiology. *American Journal of Epidemiology*, 189(11), 1321-1329.
- Waskom, M. L. (2021). Seaborn: Statistical data visualization. *Journal of Open Source Software*, 6(60), 3021.
- Wickham, H. (2014). Tidy data. *Journal of Statistical Software*, 59(10), 1-23.
- Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, 30(1), 79-82.
- World Health Organization (WHO). (2022). *Dengue: Guidelines for diagnosis, treatment, prevention and control*. WHO.
- World Health Organization (WHO). (2025, August 21). *Dengue and severe dengue*. <https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue>
- World Medical Association. (2013). World Medical Association Declaration of Helsinki: Ethical principles for medical research involving human subjects. *JAMA*, 310(20), 2191-2194.
- XGBoost Developers. (2026a). *Introduction to boosted trees*. XGBoost Documentation. <https://xgboost.readthedocs.io/en/stable/tutorials/model.html>
- XGBoost Developers. (2026b). *XGBoost documentation*. <https://xgboost.readthedocs.io>
- XGBoost Developers. (2026c). *Tree methods*. XGBoost Documentation. https://xgboost.readthedocs.io/en/stable/tree_method.html
- Zhao, N., Charland, K., Carabali, M., Nsoesie, E. O., Maheu-Giroux, M., & Rees, E. (2020). Machine learning and dengue forecasting: Comparing random forests and artificial neural networks for predicting dengue burden at national and sub-national scales in Colombia. *PLOS*

Neglected Tropical Diseases, 14(9), e0008056.

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

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, 67(2), 301-320.