Pontificia Universidad Javeriana Cali
Facultad de Ingeniería y Ciencias
Maestría en Ciencia de Datos
Proyecto Aplicado

MODELO PREDICTIVO PARA LA IDENTIFICACIÓN DE ZONAS DE RIESGO DE DENGUE GRAVE: UN ENFOQUE INTEGRAL DE CONDICIONES SOCIODEMOGRÁFICAS Y CALIDAD DE SERVICIOS DE SALUD

Sergio Andres Rueda Gonzalez

Director: Dr. Delia Ortega Lenis

2024

Resumen

El presente trabajo de investigación propone un modelo predictivo para la identificación de zonas con riesgo elevado de dengue grave en el municipio de Girón, fundamentado en la integración de variables epidemiológicas, climáticas, sociodemográficas y asociadas a la calidad de la atención en salud. La problemática analizada surge de la alta carga de enfermedad y letalidad atribuida al dengue en contextos de vulnerabilidad social, donde la respuesta del sistema de salud es limitada o inadecuada.

Se implementó una metodología cuantitativa que incluyó la consolidación de fuentes de datos oficiales como SIVIGILA, registros de visitas ETV y registros meteorológicos diarios. Tras un riguroso proceso de limpieza, transformación y unificación de bases de datos, se construyó una matriz de variables predictoras a nivel de barrio. Esta matriz fue utilizada para entrenar modelos de clasificación, entre ellos Random Forest y Regresión Logística, los cuales fueron evaluados mediante validación cruzada estratificada, obteniendo desempeños óptimos (AUC-ROC promedio $> 0{,}99$).

Se integraron indicadores de riesgo clínico (discordancia entre clasificación y conducta), reconsultas frecuentes en una misma IPS, visitas múltiples en menos de diez días, y exposición ambiental (precipitación y criaderos). La visualización geoespacial se realizó mediante mapas interactivos tipo choropleth, lo que facilita la identificación visual de áreas prioritarias. El modelo contribuye de forma significativa a la planificación territorial en salud, proponiendo un enfoque preventivo y basado en evidencia.

Palabras Clave: dengue, aprendizaje automático, salud pública, riesgo geoespacial, calidad en salud.

Índice general

Capítulo 1

Introducción

El dengue representa una de las principales amenazas para la salud pública en América Latina, con un crecimiento sostenido en su incidencia y severidad en las últimas décadas [1]. En particular, el dengue grave o hemorrágico ha mostrado una tendencia preocupante, especialmente en zonas urbanas densamente pobladas y con deficiencias en servicios de saneamiento y salud [2]. En Colombia, el Sistema Nacional de Vigilancia en Salud Pública (SIVIGILA) ha reportado cifras crecientes de dengue grave, lo que evidencia la necesidad de herramientas analíticas que permitan anticipar su ocurrencia y facilitar la intervención oportuna.

Modelos predictivos basados en técnicas de aprendizaje automático han demostrado ser útiles para la detección temprana de brotes y la identificación de zonas de riesgo [3]. Estos modelos permiten incorporar grandes volúmenes de información heterogénea, tales como variables climáticas (precipitación, temperatura), demográficas (edad, sexo), socioeconómicas (estrato) y factores asociados al acceso y calidad en salud [4]. Sin embargo, en muchos contextos locales aún se carece de modelos aplicados que integren datos multifuente con validación territorial.

Este trabajo desarrolló un modelo predictivo robusto que integró información epidemiológica, ambiental y de calidad de servicios de salud para identificar barrios con mayor probabilidad de ocurrencia de dengue grave en Girón, Santander. El modelo no solo se enfoco en la predicción estadística, sino también en la interpretación de los factores de riesgo y la visualización territorial mediante herramientas geoespaciales. La propuesta se alineó con los principios de vigilancia epidemiológica activa y gestión basada en datos, con el fin de fortalecer la toma de decisiones de salud pública local.

Capítulo 2

Contextualización del Proyecto

2.1. Definición del Problema

El municipio de Girón, localizado en el departamento de Santander (Colombia), ha sido históricamente afectado por brotes recurrentes de dengue, incluyendo casos graves y fatales. A pesar de los esfuerzos institucionales por mitigar la propagación del vector Aedes aegypti, persisten condiciones estructurales que propician su reproducción y dificultan la contención del virus: viviendas con deficiencias en saneamiento, concentración de criaderos potenciales en entornos domésticos, falta de cobertura continua en salud y debilidades en la atención oportuna de casos reportados.

Adicionalmente, los registros en SIVIGILA muestran inconsistencias en la clasificación y manejo clínico de los casos de dengue, lo cual podría estar relacionado con deficiencias en la implementación de la guía nacional para la atención integral del paciente con dengue. Es común encontrar pacientes con signos de alarma o diagnóstico de dengue grave manejados en consulta ambulatoria, así como personas que acuden varias veces a la misma IPS por falta de resolución clínica o seguimiento oportuno.

Frente a esta situación, se planteó la necesidad de construir un modelo predictivo que permitiera integrar variables multifuente —epidemiológicas, climáticas, sociodemográficas y de calidad en la atención— para identificar zonas geográficas con mayor probabilidad de riesgo. Esta herramienta busca superar las limitaciones de los análisis descriptivos tradicionales y dotar a los tomadores de decisiones de un instrumento útil para priorizar intervenciones, mejorar la asignación de recursos y reducir la letalidad por dengue en el territorio.

La ausencia de mecanismos automatizados de alerta temprana en el municipio de Girón, junto con la necesidad de focalizar la prevención en barrios de alta vulnerabilidad, reforzó la pertinencia de un enfoque integral apoyado en ciencia de datos y aprendizaje automático.

2.1.1. Planteamiento del Problema

El dengue es una enfermedad viral transmitida por mosquitos que representa un desafío significativo para la salud pública a nivel mundial. A pesar de los esfuerzos en prevención y control, sigue

siendo una de las principales causas de morbilidad y mortalidad en muchas regiones, especialmente en áreas tropicales y subtropicales. La letalidad del dengue se ve influenciada por múltiples factores, incluyendo las condiciones sociodemográficas y socioeconómicas de las zonas afectadas, así como la calidad de los servicios de salud disponibles. En muchas comunidades, la falta de recursos adecuados y deficiencias en la atención médica agravan la situación, aumentando la mortalidad asociada a esta enfermedad [1].

En Colombia, el comportamiento endemo-epidémico del dengue ha mostrado una alta recurrencia, especialmente en departamentos como Santander. Según los registros del Sistema Nacional de Vigilancia en Salud Pública (SIVIGILA), entre 2018 y 2024 se notificaron 9.437 casos de dengue en el municipio de Girón, de los cuales 267 fueron clasificados como dengue grave, representando un 2,83 % del total.

En el año pico de 2019, se reportaron más de 2.300 casos solo en Girón, lo que posicionó al municipio como una de las zonas más afectadas del área metropolitana de Bucaramanga en el Departamento de Santander. Aunque ha habido fluctuaciones en la incidencia, la carga de enfermedad se ha mantenido elevada, a pesar de los esfuerzos institucionales en control vectorial y atención médica.

El análisis de los reportes de SIVIGILA evidencia recurrencia de eventos en los mismos pacientes, reconsultas frecuentes a una misma IPS y falta de seguimiento clínico oportuno, lo cual sugiere fallas sistemáticas en la respuesta institucional. Estas situaciones no solo incrementan el riesgo de progresión a formas graves, sino que comprometen la capacidad resolutiva del sistema de salud ante una eventual epidemia.

Con base en la base de SIVIGILA para los casos de reporte de 2018 a 2024, se identificaron casos en los que una misma persona fue atendida varias veces en un corto periodo, con diferencias menores a 10 días entre fechas de notificación, y sin evidencia de atención hospitalaria cuando el caso estaba clasificado como dengue con signos de alarma o dengue grave. Además, se documentó la atención reiterada en las mismas IPS, lo que sugiere posibles fallas en la resolución clínica o ausencia de seguimiento adecuado.

Por otro lado, el componente ambiental también contribuye al riesgo. En las 9.885 viviendas visitadas por el grupo ETV, se encontró presencia de criaderos en más del 78 %, siendo los tanques, llantas y floreros los principales focos. lo que representa un foco crítico para la proliferación del vector Aedes aegypti. Estas evidencias empíricas reforzaron la hipótesis de que las deficiencias estructurales y de calidad de atención están directamente asociadas al riesgo de agravamiento y a la carga territorial del dengue.

En este contexto, resulta fundamental desarrollar herramientas predictivas basadas en evidencia que permitan anticipar zonas de mayor vulnerabilidad y riesgo. Modelos de aprendizaje automático aplicados al contexto de salud pública han demostrado ser útiles para apoyar la planificación

territorial, optimizar recursos y focalizar acciones preventivas [3]. Sin embargo, su implementación a nivel local ha sido escasa o inexistente.

La presente investigación se planteó como respuesta a esta necesidad, buscando construir un modelo predictivo multivariable que integre datos epidemiológicos, climáticos, sociodemográficos y de calidad de atención para identificar con mayor precisión las zonas críticas del municipio. Este modelo tiene el potencial de convertirse en una herramienta estratégica para fortalecer la vigilancia en salud pública, la respuesta institucional y la toma de decisiones informadas por parte de los entes territoriales.

2.1.2. Formulación

¿Cómo se puede desarrollar un modelo predictivo eficaz que identifique las zonas de alto riesgo de presencia de dengue grave, integrando variables epidemiológicas, climáticas, sociodemográficas, socioeconómicas y de calidad de los servicios de salud, para mejorar la respuesta de salud pública y la asignación de recursos?

2.1.3. Sistematización

¿Qué variables provenientes de fuentes epidemiológicas, entomológicas, climáticas y sociodemográficas presentan mayor frecuencia o asociación con la ocurrencia de casos graves de dengue?

¿Qué técnicas de aprendizaje supervisado permiten integrar datos multifuente para construir un modelo predictivo confiable de riesgo territorial de dengue grave, y qué variables resultan más significativas en la clasificación de zonas de alta prioridad?

¿Qué indicadores derivados de los registros clínicos pueden emplearse para evaluar de manera indirecta la calidad de la atención médica frente al dengue, y cómo impactan estos en la predicción del riesgo de gravedad a nivel territorial?

Para abordar la alta incidencia de dengue grave en ciertas áreas, es necesario desarrollar un modelo predictivo que no solo identifique las zonas de riesgo, sino que también evalúe las condiciones sociodemográficas, socioeconómicas y la calidad de la prestación de servicios de salud en esas zonas. Este modelo debe ser capaz de:

Integrar Datos Diversos: Incluir variables epidemiológicas, climáticas, sociodemográficas y socioeconómicas que influyen en la incidencia de dengue grave.

Evaluar la Calidad de los Servicios de Salud: Incorporar indicadores indirectos construidos a partir de los registros de SIVIGILA, como la discordancia entre clasificación clínica y conducta médica, la recurrencia de atención por el mismo evento en menos de diez días, y el número de reconsultas

en una misma institución prestadora, permitiendo identificar deficiencias en el seguimiento clínico y la adherencia a guías de manejo.

Identificar Áreas de Intervención Prioritaria: Detectar barrios con alta concentración de criaderos, casos graves, reconsultas frecuentes y atención ambulatoria inadecuada, lo que permite priorizar acciones de control vectorial, vigilancia clínica y fortalecimiento institucional de forma focalizada y territorializada.

Mejorar la Respuesta de Salud Pública: Generar evidencia geoespacial y multivariada que apoye la toma de decisiones basada en datos, facilitando la planificación de intervenciones preventivas, la distribución de recursos en salud y el fortalecimiento de la vigilancia epidemiológica activa frente a eventos de dengue grave.

2.2. Objetivos

2.2.1. Objetivo General

Desarrollar un modelo predictivo que identifique las zonas de riesgo donde pueda presentarse alta incidencia de dengue grave a través de Técnicas de Aprendizaje Automático que evalué tanto las condiciones sociodemográficas y socioeconómicas de la zona como la calidad de la prestación de servicios de salud del área

2.2.2. Objetivos Específicos

- Objetivo específico 1 Recolectar datos epidemiológicos, climáticos, sociodemográficos y socioeconómicos que influencien la incidencia de dengue grave.
- Objetivo específico 2 Crear un modelo predictivo que integre los datos sociodemográficos, socioeconómicos y de calidad de los servicios de salud para identificar las zonas de alto riesgo de dengue grave.
- Objetivo específico 3 Incluir en el modelo predictivo variables relacionadas con la calidad de la atención médica en las zonas estudiadas

2.3. Marco de Referencia

El desarrollo de un modelo predictivo para la identificación de zonas de riesgo de letalidad por dengue se basa en un enfoque integral que considera las condiciones sociodemográficas y la

calidad de los servicios de salud. Este marco de referencia explora los conceptos básicos y las ideas principales que sustentan el proyecto, proporcionando un enfoque teórico que integra múltiples disciplinas para abordar el problema de manera holística.

2.3.1. Marco Teórico

- 1.Epidemiología del Dengue y Determinantes Sociales

  El dengue es una enfermedad viral aguda transmitida por el mosquito *Aedes aegypti*, ampliamente distribuido en regiones tropicales y subtropicales. El virus pertenece al género *Flavivirus* y se presenta en cuatro serotipos (DENV-1 a DENV-4), lo que implica que una persona puede contraer la infección más de una vez, aumentando el riesgo de formas graves como el dengue con signos de alarma o dengue grave [2].

  Desde una perspectiva epidemiológica, el dengue tiene un comportamiento endemo-epidémico, es decir, presenta casos sostenidos en el tiempo con brotes periódicos. La Organización Mundial de la Salud lo ha clasificado como una de las diez principales amenazas para la salud pública global [1].

  Los determinantes sociales de la salud desempeñan un papel fundamental en la transmisión y severidad del dengue. Según el marco conceptual propuesto por la OMS, estos determinantes incluyen las condiciones sociales, económicas y ambientales en que las personas nacen, crecen, viven y trabajan, influenciadas por la distribución desigual de recursos, poder y acceso a servicios esenciales como salud, agua potable y saneamiento [4].

  Un componente clave dentro de estos determinantes es la calidad de la atención en salud. Esta se refiere a la oportunidad, continuidad y resolutividad con que los servicios de salud atienden los casos de dengue. Deficiencias en la calidad de la atención, como demoras en el diagnóstico, falta de seguimiento clínico, o tratamientos inadecuados, pueden contribuir a la progresión de la enfermedad hacia formas graves e incluso a desenlaces fatales [5]. En contextos donde los servicios de salud son limitados o presentan barreras de acceso, esta situación se agrava, aumentando la vulnerabilidad de las comunidades afectadas.

  En el caso del dengue, factores como el bajo nivel socioeconómico, la falta de infraestructura sanitaria adecuada, el hacinamiento, la educación deficiente y la limitada cobertura en salud, crean un entorno favorable para la proliferación del vector y dificultan una respuesta efectiva ante la enfermedad. Por ejemplo, en comunidades con acceso limitado al agua potable, es común almacenar agua en recipientes abiertos, lo que se traduce en criaderos potenciales para el mosquito *Aedes aegypti* [4].

  Asimismo, la falta de educación en salud pública y la débil participación comunitaria contribuyen a una baja percepción del riesgo, menor adopción de medidas preventivas, y demoras en la búsqueda de atención médica. Esto no solo incrementa la transmisión, sino también el riesgo de progresión hacia formas graves, especialmente en contextos donde el acceso a servicios médicos es limitado [6].

Estudios realizados en países endémicos han demostrado que el riesgo de dengue se concentra en zonas densamente pobladas y con condiciones precarias de vivienda, saneamiento y servicios públicos. Estas condiciones estructurales no solo favorecen la transmisión del virus, sino que dificultan la implementación de intervenciones efectivas de control vectorial y atención oportuna [7].

En este sentido, entender la relación entre los determinantes sociales y el dengue no solo permite explicar la distribución desigual de la enfermedad, sino que también orienta la formulación de políticas públicas basadas en equidad, prevención y gestión territorial del riesgo.

- 2. Análisis geoespacial y datos multifuente

  La incorporación del componente espacial es esencial para interpretar la distribución del riesgo. Estudios recientes como los realizados por Kraemer et al. (2015), que mapearon la distribución de Aedes aegypti y Aedes albopictus a nivel global, y Chadsuthi et al. (2012), quienes modelaron la incidencia de dengue en función del clima en Tailandia, han demostrado la utilidad de métodos como regresión geográficamente ponderada (GWR), interpolación espacial, análisis bayesiano espacial y mapas tipo choropleth para modelar eventos como el dengue [3][8][9]. En este proyecto se integraron datos de vigilancia y visitas ETV a través de geocodificación por barrios, lo cual facilitó la construcción de mapas de riesgo visuales y permitió correlacionar la concentración de criaderos con la frecuencia de casos graves en determinadas zonas. El uso de archivos GeoJSON y shapefiles, así como herramientas de visualización como Plotly y QGIS, fortaleció la comunicación de los resultados a las autoridades locales.

- 3. Aplicaciones de modelos predictivos en enfermedades transmitidas por vectores

  Estudios realizados en países endémicos como Brasil (Souza et al., 2019), México y Tailandia (Chadsuthi et al., 2012) han mostrado que la integración de variables ambientales, sociales y clínicas mejora la precisión de los modelos de predicción de dengue [9][10]. En contextos locales como el de Girón, donde la capacidad de respuesta institucional es limitada, el uso de estas herramientas representa una oportunidad para focalizar intervenciones preventivas y mejorar los sistemas de alerta temprana. En este proyecto, la modelación incluyó escenarios con y sin variables de calidad de atención y precipitaciones acumuladas, encontrando que su inclusión aumentó significativamente la capacidad predictiva del modelo, lo que confirma su relevancia práctica y su potencial para apoyar la toma de decisiones basada en datos.

  Modelos Predictivos Aplicados a Salud Pública

  El uso de modelos predictivos en el ámbito de la salud pública ha demostrado ser una herramienta poderosa para anticipar eventos adversos, optimizar recursos y mejorar la toma de decisiones. Estos modelos permiten identificar patrones complejos en grandes volúmenes de datos multivariables, integrando información de origen clínico, ambiental, demográfico y socioeconómico [3].

  Técnicas de aprendizaje automático como la Regresión Logística, Random Forest y Gradient Boosting han sido ampliamente aplicadas para la predicción de enfermedades transmisibles

como el dengue, debido a su capacidad para manejar variables no lineales, relaciones complejas y datos incompletos [3, 10].

2.3.2. a) Regresión Logística

La Regresión Logística es un modelo probabilístico clásico de clasificación binaria. Define la probabilidad de un desenlace $Y = 1$ como:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0+\beta_1 X_1+\cdots+\beta_k X_k)}},$$

donde los coeficientes β representan el log-odds. Su interpretabilidad radica en que cada β_j puede transformarse en un *odds ratio*, lo que facilita comunicar la fuerza de asociación entre factores de riesgo y desenlaces de salud [2, 5].

En el contexto del dengue, la Regresión Logística ha demostrado utilidad en la estimación de probabilidades de dengue grave en función de determinantes sociales y clínicos, como en Medellín [4] y Tailandia [11]. Su relevancia académica proviene de ser un modelo base de referencia, con supuestos claros y amplia aceptación en epidemiología.

Métricas de evaluación: La Regresión Logística es sensible al desbalance de clases, por lo que no basta con medir precisión global. Se priorizan:

- **AUC-ROC**: mide la capacidad del modelo de discriminar entre zonas de riesgo y no riesgo, independiente del umbral [12].
- **Recall (sensibilidad)**: fundamental en salud pública, pues minimiza la probabilidad de omitir barrios con casos graves.
- **F1-score**: balance entre sensibilidad y precisión, especialmente informativo en conjuntos desbalanceados [13].

El uso de estas métricas garantiza que el modelo no solo sea estadísticamente significativo, sino también operativamente útil en vigilancia epidemiológica.

2.3.3. b) Random Forest

Random Forest (RF), propuesto por Breiman [14], combina múltiples árboles de decisión mediante *bagging*. Cada árbol $h_b(X)$ se entrena en una muestra bootstrap y selecciona subconjuntos aleatorios de variables; la predicción final es:

$$\hat{y} = \text{mode}\{h_1(X), h_2(X), \ldots, h_B(X)\}.$$

Esto reduce la varianza, mejora la estabilidad y permite capturar relaciones no lineales. RF también calcula medidas de importancia de variables a partir de la reducción media de impureza o del error fuera de bolsa (OOB), aportando interpretabilidad relativa [15].

En dengue, Souza et al. [10] aplicaron RF en Brasil para mapear riesgo territorial y Haque et al. [7] lo usaron en Bangladesh con predictores climáticos y demográficos. RF es robusto en escenarios con ruido y variables correlacionadas.

Métricas de evaluación: Dada la relevancia de la salud pública, se priorizan:

- **AUC-ROC**: mide la capacidad global de discriminación del ensamble.
- **Recall**: RF se ajusta para maximizar la sensibilidad, reduciendo falsos negativos.
- **OOB Error**: error estimado sobre observaciones no utilizadas en el bootstrap, que provee una validación interna del modelo.

Estas métricas permiten valorar no solo la precisión, sino la estabilidad y capacidad de generalización del modelo, atributos necesarios para planificación territorial.

2.3.4. c) XGBoost (Extreme Gradient Boosting)

XGBoost [16] es un algoritmo de *boosting* secuencial que optimiza una función de pérdida aditiva:

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k), \quad \Omega(f) = \gamma T + \tfrac{1}{2}\lambda ||w||^2,$$

donde l es la pérdida (log-loss en clasificación binaria), Ω es el término de regularización y f_k los árboles secuenciales. Esto le confiere un control robusto sobre el sobreajuste, a la vez que permite manejar datos faltantes y alta dimensionalidad.

En epidemiología, XGBoost se ha utilizado para brotes de dengue en Asia con resultados superiores en precisión [3]. Phung et al. [11] lo integraron con análisis geoespacial, identificando *hotspots* de riesgo. Su eficiencia lo convierte en un candidato ideal para bases de datos grandes y heterogéneas.

Métricas de evaluación: El desempeño de XGBoost se valida mediante:

- **AUC-ROC**: permite medir su capacidad discriminativa global.
- **Recall**: prioriza la sensibilidad en detección temprana de brotes.
- **F1-score**: balance de precisión y recall, crucial en datasets desbalanceados.
- **Log-loss**: función de pérdida interna de XGBoost, que penaliza fuertemente las predicciones seguras pero incorrectas.

Estas métricas combinadas aseguran que el modelo no solo sea preciso, sino que también reduzca errores críticos en salud pública.

2.3.5. Justificación de la selección metodológica

La elección de los tres modelos responde a un balance académico entre interpretabilidad, robustez y rendimiento:

- **Regresión Logística**: interpretabilidad y contraste epidemiológico.
- **Random Forest**: robustez y capacidad de detectar variables clave en entornos complejos.
- **XGBoost**: precisión optimizada y regularización que asegura alto desempeño.

En cuanto a las métricas, se prioriza AUC-ROC como evaluación global, Recall para minimizar falsos negativos en zonas de riesgo, y F1-score para balancear sensibilidad y precisión en datos desbalanceados. Estas decisiones responden a un criterio académico y operativo en salud pública, donde el costo de un falso negativo es significativamente mayor que el de un falso positivo.

2.3.6. Manejo del desbalance de clases

En problemas de clasificación epidemiológica, como la predicción de dengue grave, es común que las clases estén desbalanceadas: la mayoría de los registros corresponden a casos leves, mientras que los graves representan un porcentaje reducido del total (minoría). Este desbalance puede sesgar los modelos hacia la clase mayoritaria, reduciendo su capacidad para detectar eventos poco frecuentes pero clínicamente relevantes [12, 13].

Detección del desbalance: El primer paso consiste en cuantificar la distribución de clases. Esto se logra mediante:

- Cálculo de la proporción entre clases (ej. 80 % casos leves vs. 20 % graves).
- Métricas específicas como el *imbalance ratio* (IR $= \frac{\#\text{mayoritaria}}{\#\text{minoritaria}}$). Valores $IR > 4$ se consideran altamente desbalanceados [17].
- Evaluación de métricas de desempeño sensibles a la clase minoritaria, como Recall y F1-score.

Técnicas para abordar el desbalance:

1. **Reponderación de clases**: ajustar los pesos de cada clase en la función de pérdida (p. ej., `class_weight="balanced"` en regresión logística o RF; `scale_pos_weight` en XGBoost). Esto penaliza más los errores en la clase minoritaria, mejorando la sensibilidad.
2. **Re-muestreo**:
   - **Oversampling**: aumentar artificialmente la clase minoritaria (ej. SMOTE – Synthetic Minority Oversampling Technique), generando instancias sintéticas que preservan la distribución de variables [18].
   - **Undersampling**: reducir la clase mayoritaria eliminando registros redundantes, lo cual mejora el balance pero puede perder información.
3. **Métricas alternativas**: priorizar Recall y F1-score por encima de la precisión global, ya que en contextos de salud pública el costo de un falso negativo (omitir un barrio en riesgo) es mayor que el de un falso positivo [13].

4. **Ajuste de umbral de clasificación**: reducir el umbral de probabilidad (ej. de 0.5 a 0.4) para incrementar la detección de la clase minoritaria.

Impacto en la interpretabilidad: Si bien estas técnicas mejoran la capacidad de los modelos para identificar la clase minoritaria, tienen implicaciones académicas:

- **Positivo**: reducen falsos negativos y mejoran la sensibilidad, atributo crucial en vigilancia epidemiológica.
- **Negativo**: el oversampling puede introducir ruido sintético que afecta la interpretabilidad de modelos lineales como la regresión logística, dificultando la lectura de *odds ratios*. En cambio, en modelos de ensamble (RF, XGBoost), la interpretabilidad relativa se mantiene, pero el ranking de importancia de variables puede variar según el balance aplicado.

Por tanto, el abordaje académico al desbalance debe reconocer que la prioridad en salud pública no es maximizar la precisión global, sino garantizar la detección temprana de eventos de alto riesgo, incluso si esto implica aumentar los falsos positivos.

En conjunto, la combinación de Regresión Logística, Random Forest y XGBoost ofrece un balance entre interpretabilidad, robustez y capacidad predictiva, atributos fundamentales para apoyar decisiones de salud pública en escenarios de alto impacto como el dengue grave.

En conjunto, este marco conceptual y metodológico sustentó el desarrollo del modelo predictivo presentado, el cual busca no solo anticipar zonas de riesgo, sino también comprender los factores que explican la vulnerabilidad en salud, con un enfoque integral, territorial y basado en datos.

2.3.7. Antecedentes

A lo largo de las últimas dos décadas, el uso de modelos geoespaciales y predictivos para enfermedades transmitidas por vectores ha tenido un crecimiento significativo. Estos estudios han permitido integrar datos epidemiológicos, ambientales y sociales con el fin de anticipar zonas de brote y orientar la respuesta institucional. A continuación, se describen los antecedentes más relevantes y cómo contribuyeron a la formulación del presente proyecto.

Proyectos Globales

A nivel global, la Organización Mundial de la Salud ha promovido el uso de sistemas de alerta temprana que integran datos climáticos, densidad vectorial y factores sociales, como el sistema presentado por Lowe et al. (2011), que empleó modelos de regresión logística y aprendizaje automático con variables meteorológicas y de vigilancia para anticipar brotes en América Latina y Asia con semanas de anticipación [19].

Proyectos en Asia

En el contexto asiático, modelos espaciales y espacio-temporales han sido ampliamente utilizados. Por ejemplo, Chadsuthi et al. (2012) emplearon regresión espacial autorregresiva (SAR) y modelos autoregresivos integrados de media móvil (ARIMA) con datos climáticos para anticipar brotes de dengue en Tailandia. Las variables empleadas incluyeron precipitaciones acumuladas, temperaturas máximas y mínimas, y datos mensuales de casos de dengue [9]. De igual manera, en Dhaka (Bangladesh), Haque et al. (2012) aplicaron modelos espacio-temporales de Poisson y técnicas de análisis geoestadístico (kriging) para identificar patrones de transmisión. Utilizaron variables como densidad poblacional, precipitación acumulada, temperatura promedio, y tasas de notificación de dengue [7].

En el contexto asiático, modelos espaciales y espacio-temporales han sido ampliamente utilizados. Por ejemplo, Chadsuthi et al. (2012) emplearon regresión espacial autorregresiva (SAR) y modelos autoregresivos integrados de media móvil (ARIMA) con datos climáticos para anticipar brotes de dengue en Tailandia. Las variables empleadas incluyeron precipitaciones acumuladas, temperaturas máximas y mínimas, y datos mensuales de casos de dengue [9]. Además, se aplicó un modelo combinado entre plataformas de sistemas de información geográfica (GIS) y algoritmos de aprendizaje automático como redes neuronales artificiales (ANN) y regresión logística para predecir la incidencia del dengue grave. Este enfoque permitió detectar zonas críticas y evaluar el impacto de decisiones políticas sobre la propagación del virus [11]. El presente proyecto retoma esta lógica para incluir variables de calidad de atención como parte del modelo multivariable aplicado a nivel local.

Proyectos en América Latina

Brasil: Estudios realizados en Brasil han sido pioneros en la aplicación de técnicas como regresión logística, Random Forest y modelos basados en gradiente (Gradient Boosting) para predecir la incidencia del dengue. Por ejemplo, Souza et al. (2019) utilizaron algoritmos de Random Forest y Support Vector Machines (SVM) para modelar el riesgo de dengue en municipios brasileños, identificando como variables clave la densidad poblacional, precipitaciones y acceso a servicios de salud [10]. Además, en Recife se aplicaron modelos de regresión espacial sobre variables climáticas y demográficas para identificar zonas de alto riesgo, lo que permitió priorizar intervenciones públicas [20]. En Salvador, se combinó el análisis de datos de temperatura, precipitación y densidad poblacional utilizando modelos de regresión múltiple y técnicas de interpolación espacial, logrando focalizar campañas de fumigación con mayor eficiencia [21]. Estos estudios constituyen una base metodológica relevante para el desarrollo de modelos de riesgo climático y espacial, y justifican el uso de modelos multivariables con componentes geoespaciales en el presente proyecto.

Colombia: En Colombia, investigaciones como la de Castro y Weaver (2018) aplicaron modelos estadísticos de regresión multinomial y análisis de correspondencias múltiples para estudiar la relación entre determinantes sociodemográficos y conocimiento sobre dengue en Medellín. Identificaron como variables relevantes el estrato socioeconómico, el tipo y calidad de la vivienda, el acceso a agua potable, el nivel educativo y la cobertura en salud. Aunque no utilizaron modelos predictivos

supervisados, su enfoque contribuyó a establecer las bases para la selección de variables explicativas significativas, las cuales han sido retomadas por estudios posteriores que aplican técnicas de aprendizaje automático para predecir el riesgo de dengue con mayor precisión [4].

El presente estudio toma como base este enfoque para la creación de una variable de riesgo climático promedio por barrio.

Los proyectos consultados muestran que existe una base sólida para aplicar técnicas de análisis espacial y aprendizaje automático a la predicción del dengue. Sin embargo, pocos estudios han integrado en sus modelos variables relacionadas con calidad del servicio de salud y riesgos clínicos derivados de la atención, como sí se hace en este proyecto. Esta diferencia marca el aporte distintivo del estudio, al proponer un enfoque integral que combina factores vectoriales, sociales, institucionales y climáticos para anticipar zonas con mayor riesgo de letalidad por dengue.

Capítulo 3

RECOLECCIÓN DE DATOS - Obj1

3.1. Desarrollo del Objetivo específico 1

El primer objetivo especifico es recolectar datos epidemiológicos, climáticos, sociodemográficos y socioeconómicos que influyan en la incidencia de dengue grave en el Departamento de Santander.

Para el desarrollo de este objetivo especifico se realizan los siguientes pasos:

Paso 1: Revisión, recolección e integración de fuentes de datos

Para facilitar la visualización y comparación de las fuentes empleadas, se presenta a continuación una tabla resumen con los principales atributos de cada base de datos utilizada en esta etapa:

Cuadro 3.1: Resumen de bases de datos utilizadas

Base de datos	**Fuente institucional**	**N° de registros**	**Tipo de datos principales**
BD SIVIGILA 2018–2024	Secretaría de Salud Municipal de Girón (SIVIGILA)	9.437	Epidemiológicos: clasificación, síntomas, conducta
BD GRUPO ETV	Secretaría de Salud Municipal de Girón (ETV)	9.885	Entomológicos: criaderos, tipo y cantidad por vivienda
precipitaciones_giron.csv	IDEAM (Instituto de Hidrología, Meteorología)	491	Climáticos: precipitación diaria

Como punto de partida, se consolidaron tres fuentes primarias de información que permitieron caracterizar tanto el componente epidemiológico como ambiental y entomológico del municipio de Girón. Estas bases fueron recopiladas a partir de las siguientes fuentes oficiales:

BD SIVIGILA: Esta base fue generada a partir de la base consolidada de eventos notificados al sistema SIVIGILA, suministrada por la Secretaría de Salud Municipal de Girón. Contenía 9.437 registros individuales de casos de dengue notificados entre 2018 y 2024. Incluía variables clínicas como

clasificación del caso (cod eve), síntomas, edad, sexo, conducta médica, IPS, fecha de notificación y barrio.

BD GRUPO ETV: Esta base fue proporcionada por el equipo técnico de vigilancia entomológica de la Secretaría de Salud Municipal de Girón. Contenía 9.885 registros correspondientes a viviendas visitadas por equipos técnicos ETV en campañas de control vectorial. Se registraron variables como número de criaderos por tipo (tanques, llantas, floreros), fecha de visita y barrio.

PRECIPITACIONES GIRON: Esta base fue descargada de la plataforma de IDEAM (Instituto de Hidrología, Meteorología y Estudios Ambientales), y contenía 491 registros de datos meteorológicos diarios entre 2018 y 2024, incluyendo la variable precipitation y su fecha correspondiente. Esta base fue generada mediante un script en Python utilizando Google Earth Engine (GEE) y la colección diaria CHIRPS (Climate Hazards Group InfraRed Precipitation with Station data). El proceso incluyó autenticación con un proyecto GEE, definición de coordenadas del municipio de Girón (latitud 7.0682, longitud -73.1683), filtrado de fechas entre 2018 y 2024 y consulta de datos diarios de precipitación. Los valores fueron extraídos mediante reducción espacial sobre un punto geográfico de interés y convertidos a un DataFrame. Posteriormente, se aplicó un filtro para seleccionar días con valores ideales para la reproducción del mosquito Aedes aegypti (precipitaciones entre 3 y 10 mm). El archivo final fue exportado como CSV.

Antes de integrar estas fuentes, se realizó un análisis exploratorio inicial sobre la estructura, calidad y distribución de los datos. En particular, se verificó la presencia de valores nulos, la consistencia de los formatos de fecha y la coherencia de los valores categóricos. Una de las principales dificultades encontradas fue la heterogeneidad en la escritura de los nombres de barrios.

Para resolver esta inconsistencia, se diseñó un proceso de normalización textual y homologación de nombres de barrios basado en las siguientes acciones:

- Conversión de todos los textos a mayúsculas.
- Eliminación de tildes y caracteres especiales usando funciones de normalización Unicode.
- Aplicación de difflib.get close matches en Python para identificar nombres similares y agruparlos bajo una denominación estandarizada.
- Validación manual con una lista de referencia oficial de barrios del municipio (archivo Barrios.xlsx).

Este proceso permitió garantizar la correspondencia correcta entre los registros de casos, criaderos y precipitaciones, facilitando la integración por clave relacional NOMBRE FINAL en todos los archivos.

Desde una perspectiva de ciencia de datos, este paso fue fundamental para construir una base confiable, estructurada y robusta. La limpieza y estandarización de datos es una fase crítica en cualquier flujo de trabajo analítico, especialmente cuando se manejan fuentes heterogéneas. Los resultados de este paso sentaron la base para realizar transformaciones posteriores, análisis exploratorios multivariados y la implementación de modelos de predicción supervisada.

Paso 2: Análisis exploratorio de la base de dengue

A partir de la base BD SIVIGILA, se realizó una exploración de las variables epidemiológicas claves. El conjunto contenía 9.437 registros, de los cuales 267 correspondían a eventos clasificados como dengue grave (código 220 en cod eve) y el resto a dengue clásico (código 210). Se analizaron las frecuencias absolutas y relativas de los casos por barrio.

La Figura 3.1 presenta los **diez barrios con mayor número de casos graves de dengue**. Se observa una concentración significativa en tres zonas principales: **Rincón de Girón (5 casos)**, **Villas de San Juan (4 casos)** y **Nuevo Girón (4 casos)**. Estos sectores se destacan por superar el promedio municipal y representan áreas críticas para la vigilancia epidemiológica.

El segundo grupo está compuesto por barrios con 2 casos registrados cada uno: **Santa Cruz, Brisas Campestres, El Milagro, Poblado y Bellavista**. Finalmente, **Aldea de Girón** y **Ciudadela del Oriente** reportaron un caso cada uno. La distribución sugiere que los focos de mayor letalidad no están asociados únicamente a la densidad poblacional, sino a factores estructurales específicos como vulnerabilidad social, cobertura en salud o condiciones ambientales locales.

Este análisis sustenta la importancia de realizar una georreferenciación de los eventos graves y fortalece el enfoque territorial del modelo predictivo propuesto en el capítulo siguiente, al evidenciar que ciertos barrios concentran no solo mayor número de casos, sino posiblemente mayores fallas en la contención oportuna del dengue.

Se elaboraron histogramas de distribución por edad, identificando dos picos: uno entre los 5–14 años y otro entre los 25–39 años, lo cual refleja patrones demográficos de exposición diferenciados. Asimismo, se cruzaron las variables clasfinal y conducta para detectar inconsistencias clínicas, evidenciando un número significativo de casos clasificados como graves que recibieron manejo ambulatorio, lo que motivó la construcción del indicador de “riesgo institucional”.

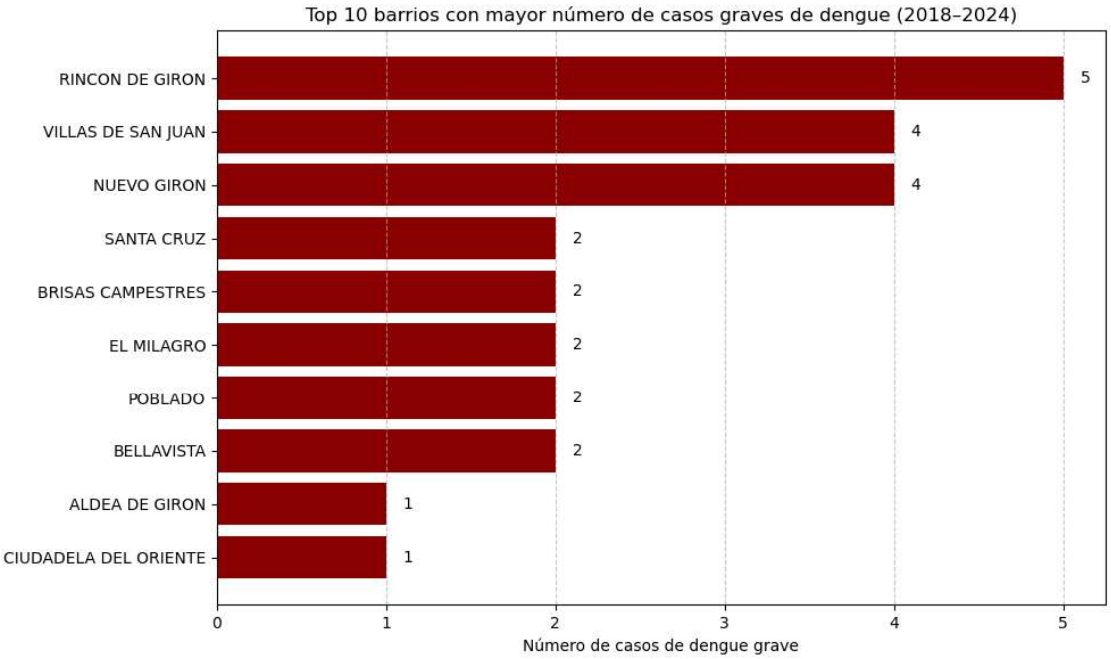

Figura 3.1: Top 10 barrios con mayor número de casos graves de dengue en Girón (2018–2024)

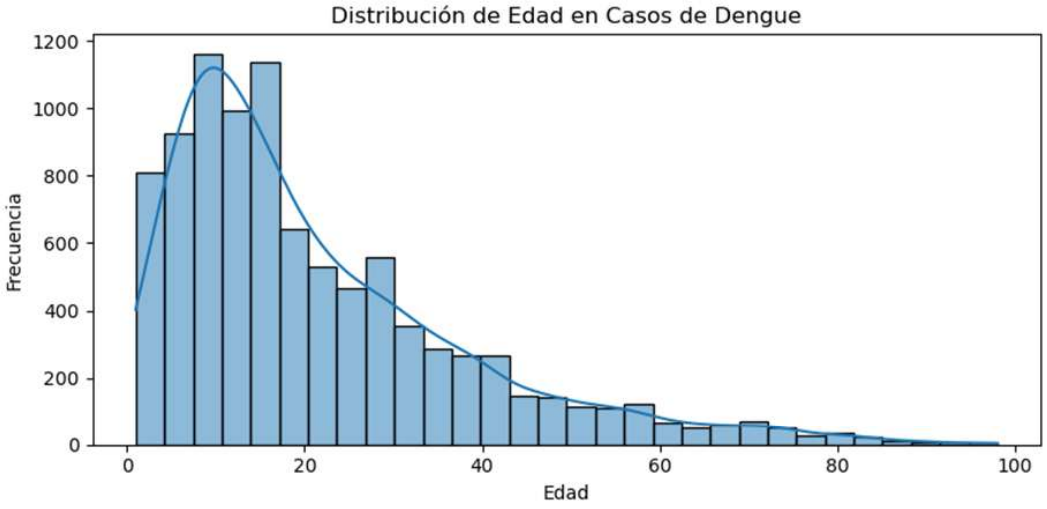

Figura 3.2: Distribución de edad en los casos de dengue reportados en Girón.

Cuadro 3.2: Criterios de identificación de riesgo clínico por atención inadecuada

Clasificación clínica (clasfinal)	Conducta médica (conducta)	Motivo de riesgo
2 – Dengue con signos de alarma	1 – Ambulatoria	No se garantizó observación médica oportuna a un caso con signos de alerta.
3 – Dengue grave	1 – Ambulatoria	Inadecuado manejo ambulatorio ante un caso grave que requería hospitalización urgente.
3 – Dengue grave	2 – Hospitalización en piso	Se hospitaliza, pero no en unidad adecuada (no en UCI o manejo especializado), lo que puede reflejar subestimación del riesgo.
2 – Dengue con signos de alarma	4 – Observación	Manejo insuficiente, ya que no se procede con hospitalización o vigilancia estrecha.
3 – Dengue grave	4 – Observación	Atención clínica insuficiente frente a un cuadro clínico de gravedad, lo cual representa una omisión crítica en la atención.

Nota. Los criterios aquí definidos corresponden a combinaciones entre la clasificación clínica del caso ('clasfinal') y la conducta médica adoptada, que indican una posible desviación frente a la Guía de Atención Integral para Dengue del Ministerio de Salud de Colombia (MSPS, 2014).

Desde la perspectiva analítica, este paso permitió identificar patrones críticos de atención, segmentar la población afectada y validar los campos que serían útiles para generar nuevas variables predictoras. Se descartaron variables con alto porcentaje de valores nulos (¿25 porciento) y se documentó el comportamiento de variables binarias como presencia de fiebre, vómito o dolor abdominal.

Paso 3: Análisis exploratorio de criaderos

La base de ETV fue procesada para calcular indicadores por barrio. Se contaron un total de 9.885 viviendas visitadas, en las cuales se registró al menos un criadero en el 78 porciento de los casos. Se analizaron de forma separada los tipos de criaderos: tanques, llantas, floreros y objetos

diversos.

Mediante agrupamiento por barrio, se estimó el promedio de criaderos por vivienda y se construyó una variable de "carga entomológica promedio". Esta fue clave para caracterizar la exposición ambiental al vector. Se observaron zonas con valores extremos, como El Poblado con más de 5 criaderos promedio por vivienda. Adicionalmente, se generaron boxplots que mostraron la variabilidad intra-barrial, reforzando la hipótesis de que los criaderos no están distribuidos homogéneamente.

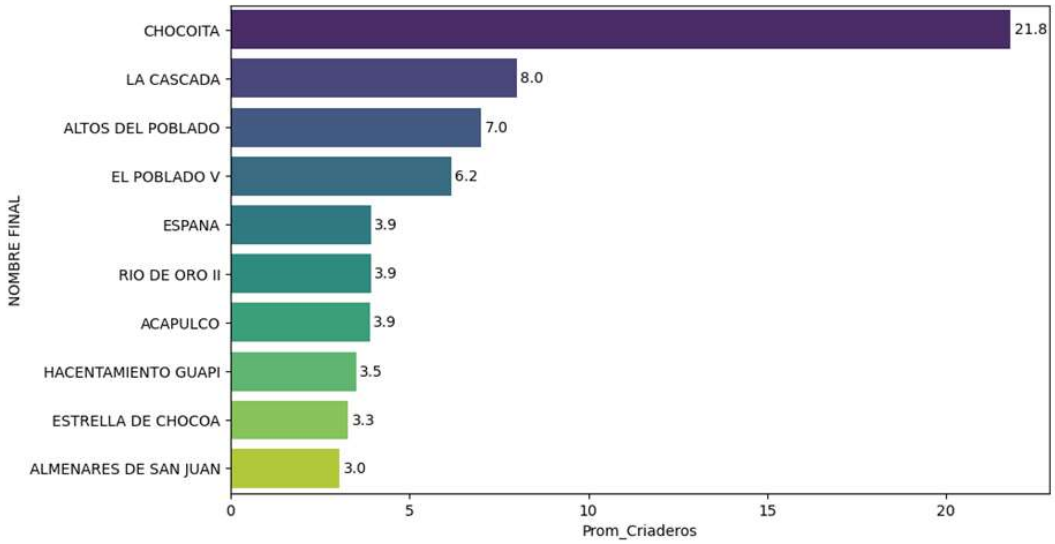

Figura 3.3: Promedio de criaderos por vivienda en los 10 barrios con mayor carga entomológica.

Este análisis también permitió identificar qué variables de criaderos se correlacionaban más con los casos de dengue. La presencia de tanques y llantas mostró una correlación superior al 0.65 con el número de casos, lo que sustentó su inclusión prioritaria en el modelo predictivo posterior.

Paso 4: Análisis de precipitaciones

A partir del archivo precipitaciones giron, se procesó una serie diaria de datos desde 2018 hasta 2024. Se calcularon estadísticas de tendencia, estacionalidad y se construyó la media móvil de 7 días (precipitacion 7d). Posteriormente, se alineó esta información con las fechas de notificación de casos para explorar posibles correlaciones temporales.

Se identificó una asociación débil pero consistente entre los picos de precipitación y los aumentos de casos de dengue reportados, con un desfase promedio de 10 días. Esta relación fue evidenciada mediante el análisis de series temporales alineadas y cálculo de correlaciones cruzadas, lo cual respalda el uso de la precipitación desfazada como una variable predictora indirecta dentro del modelo. En contraste, el desfase de 14 días no mostró asociación relevante, lo que indica que el efecto de la lluvia sobre la transmisión se manifiesta en un intervalo más corto tras el evento

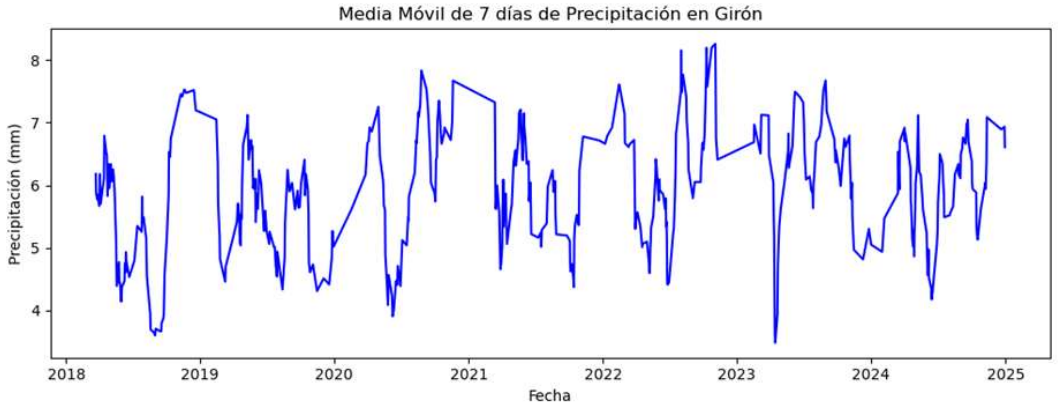

Figura 3.4: Serie temporal de media móvil de 7 días de precipitación en Girón (2018–2024).

climático.

Resultados:

Correlación precipitación (10 días antes) vs casos: 0.163 Correlación precipitación (14 días antes) vs casos: -0.021

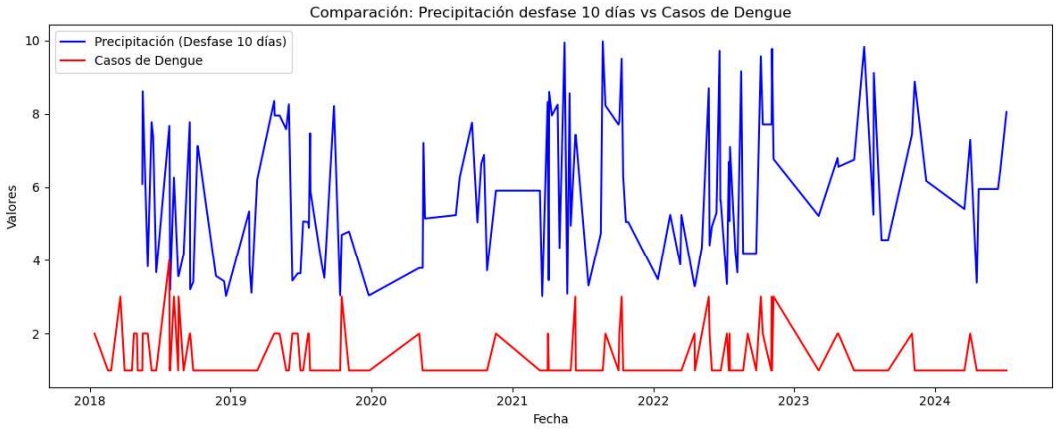

Figura 3.5: Serie temporal de media móvil de 7 días de precipitación en Girón (2018–2024).

Se crea el indicador precipitacion 7d que es la media móvil de los últimos 7 días de precipitación diaria por medio de datos provenientes del sistema CHIRPS. Este valor fue luego asociado a cada registro de dengue según su fecha de notificación. Finalmente, se calculó el promedio semanal por barrio (promedio precipitacion 7d) como variable agregada. Este indicador permite capturar el efecto acumulado de las lluvias, relevantes para la reproducción del vector Aedes aegypti, y fue incorporado como predictor climático clave.

Este paso permitió convertir un dato climático puntual en un indicador de riesgo agregado útil para análisis territoriales, demostrando la importancia de transformar variables externas en variables explicativas dentro del marco del aprendizaje supervisado.

Paso 5: Análisis de variables sociodemográficas y socioeconómicas

Se utilizaron variables como edad, sexo y estrato socioeconómico de la base de casos para analizar la distribución del dengue grave en función del perfil poblacional. Se generaron estadísticas descriptivas por barrio, identificando zonas con mayor concentración de población vulnerable, como menores de edad y adultos mayores.

En cuanto al **estrato socioeconómico**, se realizó un proceso de *reclasificación y estandarización* con el fin de garantizar su comparabilidad entre barrios. Dado que un mismo barrio puede contener viviendas con diferentes niveles de estrato, se optó por utilizar como valor representativo el **estrato moda** (es decir, el más frecuente), ya que refleja la condición socioeconómica predominante de la zona y ha sido ampliamente usado en estudios de análisis territorial.

Durante el análisis exploratorio, se identificó que los **estratos 1 y 2** concentraron **más del 60 % de los casos de dengue grave**, resultado que coincide con hallazgos de investigaciones previas que vinculan los *niveles de pobreza con una mayor vulnerabilidad frente al dengue*. Esta relación se debe tanto a una **mayor exposición al vector** por condiciones ambientales precarias, como a **limitaciones en el acceso oportuno a los servicios de salud**.

Por tal motivo, la variable `estrato` fue incorporada como **predictor categórico** en los modelos de clasificación, con el objetivo de capturar el efecto del contexto socioeconómico sobre la probabilidad de ocurrencia de formas graves de la enfermedad.

Paso 6: Cruce de información entre bases

Con el propósito de construir una base de datos territorializada para alimentar los modelos predictivos, se consolidaron múltiples fuentes de información mediante una integración por la variable común `NOMBRE FINAL` (nombre del barrio). Esta operación permitió generar una matriz de variables independientes que caracterizan a cada barrio desde dimensiones epidemiológicas, clínicas, entomológicas, sociodemográficas y climáticas. A continuación, se detallan las principales variables generadas:

Variables clínicas agregadas por barrio. Cada caso notificado al sistema SIVIGILA incluye el lugar de residencia declarado por el paciente, con dirección, barrio y municipio. Esta información permitió asignar los síntomas clínicos reportados a la zona de residencia, y no necesariamente al lugar de atención médica. A partir de esto, se calcularon por barrio las proporciones de los síntomas más relevantes en dengue:

- `porcentaje_fiebre`
- `porcentaje_cefalea`
- `porcentaje_dolor_abdo`
- `porcentaje_vomito`

Estos valores reflejan la intensidad sintomatológica reportada por la población de cada barrio, permitiendo identificar zonas con manifestaciones clínicas de mayor gravedad y potencial riesgo.

Variables sociodemográficas. Se agregaron las siguientes variables desde los registros individuales:

- `promedio_edad`: edad promedio de los casos registrados en el barrio.
- `proporcion_mujeres`: proporción de mujeres dentro de los casos.
- `estrato_moda`: valor más frecuente del estrato socioeconómico, utilizado en lugar del promedio por tratarse de una variable categórica ordinal.

Riesgo clínico por conducta inadecuada. Se definieron como casos de riesgo clínico aquellos en los que la conducta adoptada no era acorde con la clasificación del caso, de acuerdo con los lineamientos del Ministerio de Salud. Por ejemplo, pacientes con signos de alarma o diagnóstico de dengue grave manejados de forma ambulatoria. La suma de estos casos se calculó por barrio como `casos_riesgo_procedimiento`.

Casos epidemiológicos y control vectorial. Se agregaron los totales por barrio de:

- `casos_dengue_grave`: casos clasificados como dengue grave (evento 220).
- `casos_dengue_total`: todos los casos notificados por barrio.

También se incorporaron datos de visitas domiciliarias del programa ETV:

- `promedio_total_criaderos`: promedio de criaderos encontrados por vivienda.
- `nro_viviendas_con_criaderos`: número de viviendas con al menos un criadero identificado.

Condición climática. A partir de un dataset diario de precipitación, se calculó el promedio acumulado de los siete días previos a la notificación de cada caso. Esta variable fue luego agregada por barrio como `promedio_precipitacion_7d`, permitiendo vincular eventos epidemiológicos con condiciones climáticas locales.

Cuadro 3.3: Resumen de variables agregadas a nivel de barrio: definición, método y categoría

Variable	Descripción	Método de agregación	Categoría
`porcentaje_fiebre`	Proporción de casos con fiebre	Promedio (mean)	Clínica
`porcentaje_cefalea`	Proporción de casos con cefalea	Promedio (mean)	Clínica
`porcentaje_dolor_abdo`	Proporción de casos con dolor abdominal	Promedio (mean)	Clínica
`porcentaje_vomito`	Proporción de casos con vómito	Promedio (mean)	Clínica
`promedio_edad`	Edad promedio de los casos reportados	Promedio (mean)	Sociodemográfica
`proporcion_mujeres`	Proporción de casos que son mujeres	Promedio (mean)	Sociodemográfica
`estrato_moda`	Estrato socioeconómico más frecuente	Moda	Sociodemográfica
`casos_riesgo_procedimiento`	Casos con manejo clínico inadecuado	Conteo (sum)	Riesgo clínico
`casos_dengue_grave`	Casos notificados como dengue grave (evento 220)	Conteo (sum)	Epidemiológica
`casos_dengue_total`	Total de casos de dengue notificados	Conteo (sum)	Epidemiológica
`promedio_total_criaderos`	Criaderos promedio por vivienda visitada	Promedio (mean)	Entomológica
`nro_viviendas_con_criaderos`	Número de viviendas con criaderos detectados	Conteo (sum)	Entomológica
`promedio_precipitacion_7d`	Precipitación promedio 7 días antes de la notificación	Promedio (mean)	Climática

Fuente: elaboración propia a partir de bases integradas de SIVIGILA, ETV y registros climáticos de Girón (2018–2024). Las variables fueron utilizadas como predictores del modelo de riesgo de dengue grave a nivel barrial.

Capítulo 4

CONSTRUCCIÓN DEL MODELO - Obj2

4.1. Desarrollo del Objetivo específico 2

El segundo objetivo específico es crear un modelo predictivo que integre los datos sociodemográficos, socioeconómicos y de calidad de los servicios de salud para identificar las zonas de alto riesgo de dengue grave.

Para el desarrollo de este objetivo específico se realizan los siguientes pasos:

4.1.1. Paso 1: Selección de la variable objetivo y predictores

Con base en la matriz integrada consolidada durante el desarrollo del objetivo 1, se definió como variable objetivo (`zona_riesgo_alta`) una etiqueta binaria que indica si en el barrio se ha presentado al menos un caso de dengue grave. Esta decisión se fundamenta en la necesidad de orientar acciones preventivas hacia zonas con historial reciente de complicaciones clínicas.

Justificación epidemiológica: La definición de zona de riesgo alta basada en la presencia de al menos un caso de dengue grave responde a un criterio de acción temprana, ampliamente respaldado por la literatura en salud pública. Un solo caso de dengue grave puede reflejar fallas en la atención clínica, retraso en la consulta, baja cobertura de vigilancia entomológica o condiciones sociales que favorecen la progresión de la enfermedad. Además, la aparición de casos graves en barrios que previamente no reportaban dicha condición puede representar el inicio de un brote en zonas vulnerables. Desde la perspectiva de intervención territorial, detectar tempranamente estas *"zonas centinela"* permite focalizar la respuesta institucional para mitigar la progresión del brote. Por tanto, clasificar como zona de riesgo todo barrio con al menos un caso grave refuerza el principio de acción oportuna y vigilancia intensificada.

De acuerdo con la Organización Panamericana de la Salud (OPS), "todo caso de dengue grave debe ser considerado una prioridad sanitaria, ya que puede indicar fallas en la respuesta del sistema

de salud o en las condiciones de vida que permiten la transmisión intensificada del virus" (OPS, 2023). Este lineamiento respalda el uso de un umbral inclusivo para activar acciones de vigilancia y control.

Justificación de la inclusión de la precipitación como única variable climática

En el marco del presente estudio, se decidió incluir únicamente la variable de **precipitación** como componente climático dentro del modelo predictivo de riesgo de dengue grave. Esta decisión se fundamenta en los siguientes aspectos:

1. **Disponibilidad y calidad de los datos:** A nivel municipal, la precipitación fue la única variable climática con cobertura continua, resolución diaria y validación técnica suficiente. Los datos se obtuvieron a partir de la serie CHIRPS (Climate Hazards Group InfraRed Precipitation with Station data), la cual proporciona registros satelitales combinados con estaciones meteorológicas locales. En contraste, variables como temperatura, humedad relativa o velocidad del viento no contaban con cobertura geográfica ni resolución temporal adecuadas para su análisis a nivel de barrio.

2. **Relevancia biológica para el vector:** La precipitación tiene un vínculo directo con la formación de criaderos del mosquito *Aedes aegypti*, ya que el agua acumulada en recipientes artificiales constituye el medio principal para su reproducción. Este factor climático actúa como desencadenante inmediato del ciclo vectorial, lo cual la convierte en una variable crítica para anticipar aumentos en la densidad del vector y, por consiguiente, del riesgo de transmisión.

3. **Estabilidad térmica en el municipio de Girón:** En el contexto específico del municipio de Girón, la temperatura promedio se mantiene relativamente constante durante todo el año, con valores entre 24°C y 28°C. Esta escasa variabilidad limita su poder explicativo como variable predictora en el modelo, ya que no se registran fluctuaciones térmicas significativas que puedan relacionarse con aumentos o reducciones abruptas en la incidencia de dengue.

4. **Aplicabilidad operativa:** Desde la perspectiva de la gestión en salud pública, la precipitación es una variable de fácil interpretación, monitoreo y actualización en tiempo real. Su incorporación en el modelo permite facilitar su uso en tableros predictivos y sistemas de alerta temprana, orientando la toma de decisiones sobre vigilancia entomológica y acciones de control vectorial de forma oportuna y focalizada.

Como predictores se seleccionaron variables derivadas de las siguientes dimensiones:

- **Entomológica:** promedio de criaderos totales y por tipo (tanques, llantas, floreros, diversos).

- **Climática:** promedio de precipitación 7 días antes (`promedio_precipitacion_7d`).
- **Epidemiológica:** número de casos totales, proporción de síntomas (fiebre, vómito, dolor abdominal) y casos con riesgo clínico por conducta (`casos_riesgo_procedimiento`).
- **Sociodemográfica y socioeconómica:** promedio de edad, proporción de mujeres y estrato promedio.
- **Calidad de servicios de salud:** número de casos clasificados con incongruencia entre diagnóstico y conducta médica (`casos_riesgo_procedimiento`).

4.1.2. Tratamiento de valores faltantes

Como parte del preprocesamiento para la construcción del modelo predictivo, se llevó a cabo un análisis sistemático de los valores faltantes en las variables continuas del conjunto de datos integrado. El objetivo fue garantizar la completitud y consistencia de las variables predictoras antes de la fase de escalamiento y modelado.

Porcentaje de datos faltantes

Se evaluaron un total de 22 variables continuas, de las cuales 8 presentaban valores faltantes. Todas ellas compartían el mismo porcentaje de ausencia, equivalente al 4.20 %. Entre las variables con datos faltantes se encontraron: `estrato_moda`, `promedio_precipitacion_7d`, y los porcentajes de síntomas clínicos como fiebre, vómito, dolor abdominal y cefalea. El resto de las variables no presentó valores nulos.

Resumen:

- **Promedio general de datos faltantes (variables continuas):** 1.53 %
- **Máximo porcentaje por variable:** 4.20 % (`promedio_edad`)

Estrategia de imputación

Con base en el tipo y naturaleza de los datos, se implementaron dos estrategias de imputación:

1. Las variables numéricas continuas fueron imputadas utilizando la **media aritmética** de cada columna, mediante la función `SimpleImputer` de la biblioteca `scikit-learn`.
2. La variable `estrato_moda`, aunque es de tipo numérico, representa una categoría ordinal. Por tal razón, fue imputada con su **moda** (el valor más frecuente), preservando su interpretación categórica y evitando distorsión estadística.

Impacto sobre la variabilidad

Para evaluar el impacto de la imputación, se comparó la desviación estándar de las variables antes y después del tratamiento. La Tabla 4.1 muestra que la imputación tuvo un efecto mínimo sobre la dispersión de los datos. Por ejemplo, la desviación estándar de `promedio_edad` pasó de 6.36 a 6.22, y la de `estrato_moda` de 0.6095 a 0.5984. En general, los cambios fueron despreciables (menores al 3 %), lo cual indica que la imputación fue adecuada y no introdujo sesgos significativos.

Cuadro 4.1: Desviación estándar antes y después de la imputación

Variable	**Antes**	**Después**	**Cambio**
promedio_edad	6.3573	6.2216	-0.1358
proporcion_mujeres	0.1809	0.1770	-0.0039
porcentaje_fiebre	0.0061	0.0060	-0.0001
porcentaje_cefalea	0.1642	0.1607	-0.0035
porcentaje_dolor_abdo	0.1487	0.1455	-0.0032
porcentaje_vomito	0.1546	0.1513	-0.0033
promedio_precipitacion_7d	0.1439	0.1408	-0.0031
estrato_moda	0.6095	0.5984	-0.0112

Escalamiento posterior

Una vez completado el proceso de imputación, todas las variables continuas fueron transformadas mediante escalamiento estándar (*z-score normalization*), asegurando media cero y desviación estándar uno. Esto fue particularmente relevante para algoritmos sensibles a la escala, como la regresión logística, y permitió una comparación justa entre predictores con diferentes unidades de medida.

4.1.3. Análisis de la distribución de la variable objetivo

La variable objetivo `zona_riesgo_alta` se define como un indicador binario que clasifica cada barrio del municipio como zona de riesgo alto (`1`) o no (`0`) de presentar casos graves de dengue. Esta clasificación se fundamenta en la ocurrencia confirmada de al menos un caso grave de dengue en el periodo de estudio, de acuerdo con los reportes consolidados en la base de datos.

Distribución de clases

La Figura 4.1 presenta la distribución porcentual de esta variable. Se observa un claro desbalance: el 79.72 % de los barrios se encuentran en la categoría `0` (sin riesgo alto), mientras que solo el 20.28 % corresponden a la categoría `1` (zona de riesgo alto).

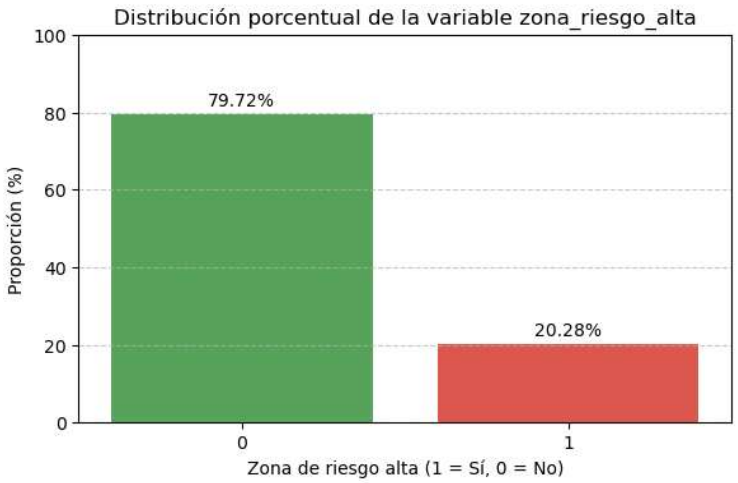

Figura 4.1: Distribución porcentual de la variable `zona_riesgo_alta`.

Este desbalance implica que los modelos de clasificación podrían verse sesgados hacia la clase mayoritaria si se evalúan únicamente mediante métricas tradicionales como la precisión general. Por ello, se adoptaron estrategias específicas tanto en el preprocesamiento como en la evaluación de los modelos para garantizar una detección efectiva de las zonas en riesgo.

Implicaciones para la selección del modelo

Dado el fuerte desbalance observado, se implementaron técnicas como:

- **Reponderación de clases**: mediante el uso del parámetro `class_weight="balanced"` en modelos como Regresión Logística y Random Forest, y el ajuste de `scale_pos_weight` en XGBoost.
- **Balanceo sintético de clases**: a través del algoritmo SMOTE (Synthetic Minority Oversampling Technique), aplicado durante la fase de entrenamiento para generar un conjunto de entrenamiento equilibrado.
- **Ajuste del umbral de decisión**: reduciendo el umbral de clasificación a 0.4, con el objetivo de incrementar la sensibilidad del modelo a la clase minoritaria.

Métricas utilizadas para la evaluación

Debido a la prioridad de no omitir barrios en riesgo, las métricas de evaluación se centraron en:

- **Recall para la clase 1**: fundamental en contextos de salud pública para minimizar los falsos negativos.
- **F1-score para la clase 1**: equilibrio entre sensibilidad y precisión en la detección de zonas de riesgo.
- **AUC-ROC**: como métrica general de discriminación del modelo.

Estas métricas fueron calculadas usando validación cruzada estratificada con 5 pliegues, lo que permitió estimar tanto el desempeño promedio como la variabilidad del modelo. El modelo final fue seleccionado con base en su F1-score promedio más alto, alto recall y baja desviación estándar, priorizando su utilidad operativa para estrategias de intervención territorial.

4.1.4. Paso 2: Entrenamiento de modelos supervisados

Para abordar la predicción del riesgo de dengue grave a nivel de barrio, se aplicaron tres algoritmos de clasificación supervisada ampliamente utilizados en problemas de salud pública:

- **Regresión logística**, con ajuste de pesos para mitigar el desbalance de clases.
- **Bosques aleatorios (Random Forest)**, con balanceo de clases y 100 árboles.
- **XGBoost Classifier**, con métrica de evaluación basada en `logloss` y ajuste de pesos por clase.

4.1.5. Entrenamiento del modelo

El proceso de entrenamiento se llevó a cabo sobre un conjunto de datos preprocesado que incluyó imputación de valores faltantes, escalamiento de variables y balanceo de clases. La variable objetivo `zona_riesgo_alta` fue tratada como una variable binaria, y se utilizaron técnicas específicas para mitigar el efecto del desbalance (aproximadamente 20 % de la clase positiva frente a 80 % de la clase negativa).

Preprocesamiento

Se aplicó imputación con la media para las variables numéricas continuas, y con la moda para la variable ordinal `estrato_moda`. Posteriormente, todas las variables continuas fueron estandarizadas

utilizando escalamiento tipo *z-score* para mejorar el desempeño de los algoritmos sensibles a la escala.

Manejo del desbalance de clases

Para mitigar el impacto del desbalance en el conjunto de entrenamiento, se empleó la técnica **SMOTE** (Synthetic Minority Over-sampling Technique), que genera instancias sintéticas de la clase minoritaria para equilibrar la distribución. Adicionalmente, se aplicó reponderación de clases mediante los parámetros `class_weight="balanced"` en los modelos de regresión logística y bosques aleatorios, y `scale_pos_weight` en XGBoost, calculado como la razón entre el número de instancias negativas y positivas.

Ajuste del umbral de decisión

Dado que el umbral predeterminado de 0.5 puede no ser óptimo en contextos desbalanceados, se ajustó el umbral de decisión a 0.4, con el objetivo de aumentar la sensibilidad del modelo sin comprometer gravemente la precisión general. Esta decisión se fundamentó en la necesidad de minimizar los falsos negativos, es decir, evitar omitir barrios que realmente se encuentran en riesgo alto.

Validación cruzada

Se implementó una validación cruzada estratificada de cinco pliegues (*Stratified K-Fold Cross Validation*) para estimar el rendimiento general de cada modelo y su estabilidad entre particiones. En cada pliegue se calcularon las siguientes métricas:

- **F1-score para la clase 1**: equilibrio entre sensibilidad y precisión.
- **Recall para la clase 1**: proporción de barrios en riesgo correctamente identificados.
- **AUC-ROC**: capacidad global del modelo para discriminar entre clases.

El desempeño promedio y la desviación estándar de estas métricas permitieron comparar modelos no solo por su exactitud, sino también por su consistencia. Los resultados detallados del proceso de validación se presentan en la siguiente sección.

4.1.6. Paso 3: Evaluación y selección del modelo predictivo

Evaluación inicial

La evaluación inicial se realizó mediante el cálculo de métricas de clasificación binaria sobre el conjunto de prueba, considerando como variable objetivo `zona_riesgo_alta`. Los resultados obtenidos se resumen en la Tabla 4.2, donde se observa que ningún modelo logra un equilibrio adecuado entre precisión y sensibilidad, debido principalmente al desbalance de clases.

Cuadro 4.2: Desempeño actualizado de los modelos supervisados sobre el conjunto de prueba

Modelo	**Precisión**	**Recall (1)**	**F1-score (1)**	**AUC-ROC**
Regresión Logística	0.694	0.571	0.421	0.754
Random Forest	0.778	0.286	0.333	0.727
XGBoost	0.667	0.429	0.333	0.695

Verificación de supuestos estadísticos en la regresión logística

Previo al ajuste final del modelo de regresión logística, se verificaron los principales supuestos que sustentan la validez estadística del enfoque: **linealidad en el logit**, **ausencia de multicolinealidad** y **bondad de ajuste**. Para el supuesto de linealidad, se evaluó gráficamente la relación entre cada predictor continuo transformado y el logit estimado mediante regresiones suavizadas (*lowess*). Este análisis permitió identificar y transformar variables con comportamiento no lineal mediante la función logarítmica $\log(x+1)$, y eliminar aquellas cuya relación con el logit fue incompatible con los supuestos del modelo.

Entre los predictores evaluados, la variable `casos_dengue_total` mostró un comportamiento que respalda el cumplimiento del supuesto de linealidad, al presentar una tendencia creciente bien definida respecto al logit estimado. Epidemiológicamente, este resultado es coherente: a mayor número total de casos de dengue en un barrio, mayor es la probabilidad estimada de que dicho territorio sea clasificado como zona de riesgo alto. La Figura 4.2 ilustra esta relación.

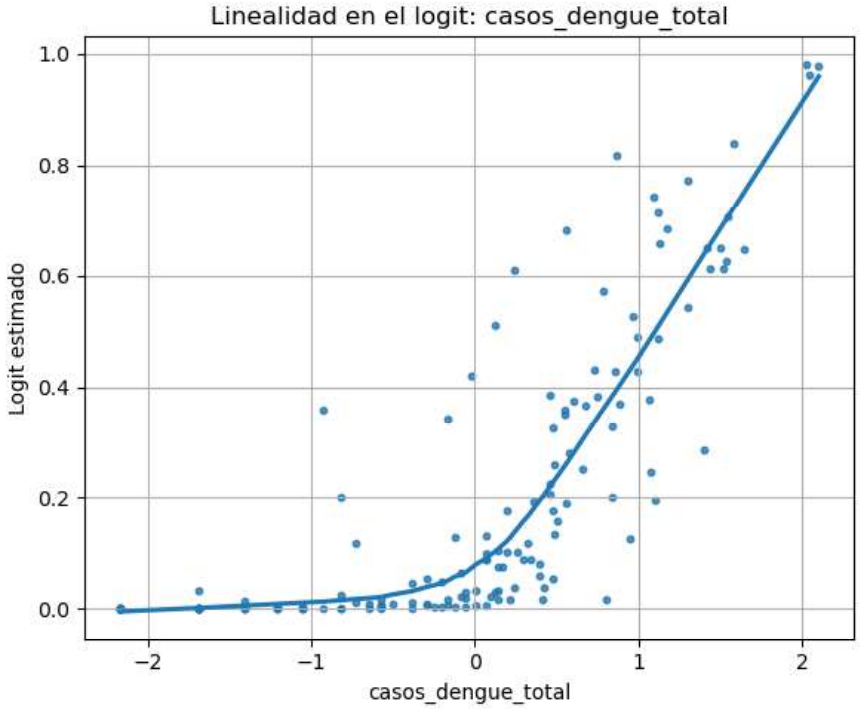

Figura 4.2: Relación entre `casos_dengue_total` y el logit estimado (*lowess*)

Adicionalmente, se examinó la colinealidad entre predictores mediante el cálculo del índice de inflación de la varianza (VIF), descartando variables con valores infinitos o superiores a 10, lo cual garantizó la estabilidad de los coeficientes. En cuanto a la bondad de ajuste, el modelo fue validado con una muestra de prueba estratificada (25 %), alcanzando una precisión de clasificación del 78 % y un AUC-ROC de 0.74 en la predicción de zonas de riesgo alto. Tras aplicar **SMOTE** para corregir el desbalance de clases y ajustar los predictores, se logró mejorar el rendimiento del modelo, especialmente en la sensibilidad hacia la clase minoritaria. Con ello, se obtuvo una puntuación F1 de **0.811**, un recall promedio de **0.871** y un AUC-ROC de **0.885**, valores que se presentan en el Cuadro 4.3.

Validación cruzada y selección final

Con el fin de evaluar la estabilidad y capacidad de generalización de cada modelo, se realizó una validación cruzada estratificada con cinco particiones. Además, se aplicó la técnica SMOTE para balancear el conjunto de entrenamiento y se ajustó el umbral de decisión a 0.4 para mejorar la sensibilidad del modelo ante la clase minoritaria (zonas de riesgo alto). Los resultados promedio

obtenidos durante la validación cruzada se resumen a continuación:

Cuadro 4.3: Resultados promedio de validación cruzada para los modelos supervisados

Modelo	F1-score promedio	Recall (1) promedio	AUC-ROC promedio	F1-score desviación
Random Forest	0.876	0.976	0.964	0.051
XGBoost	0.873	0.953	0.940	0.035
Regresión Logística	0.811	0.871	0.885	0.044

Estas métricas, junto con el cumplimiento de los supuestos teóricos del modelo, confirman la validez estadística de la regresión logística como herramienta predictiva interpretable dentro del marco del presente estudio.

A partir de estos resultados se seleccionó el **modelo Random Forest como el mejor clasificador**, dado que obtuvo el mayor F1-score promedio, la mayor sensibilidad (recall para la clase 1) y un AUC-ROC cercano a 1, lo que refleja una excelente capacidad discriminativa. Aunque XGBoost también presentó un rendimiento alto, el modelo de Random Forest mostró una ventaja ligera en términos de sensibilidad y equilibrio general.

Esta elección es particularmente pertinente en el contexto del presente estudio, ya que el objetivo principal es identificar con alta certeza las zonas de riesgo de dengue grave, minimizando la omisión de casos relevantes.

4.1.7. Paso 4: Interpretación y ranking de importancia de variables

4.1.7.1. Importancia de variables según Random Forest

Con el objetivo de interpretar el comportamiento del modelo predictivo y conocer qué factores contribuyen con mayor peso a la clasificación del riesgo, se analizó la importancia de las variables utilizando el método de *feature importance* del modelo Random Forest. Este enfoque permite identificar aquellas variables que más contribuyen a reducir el error de clasificación a lo largo de los árboles del bosque.

La Figura 4.3 presenta la visualización de las variables predictoras ordenadas por su peso relativo en el modelo. A continuación, se describen los principales hallazgos:

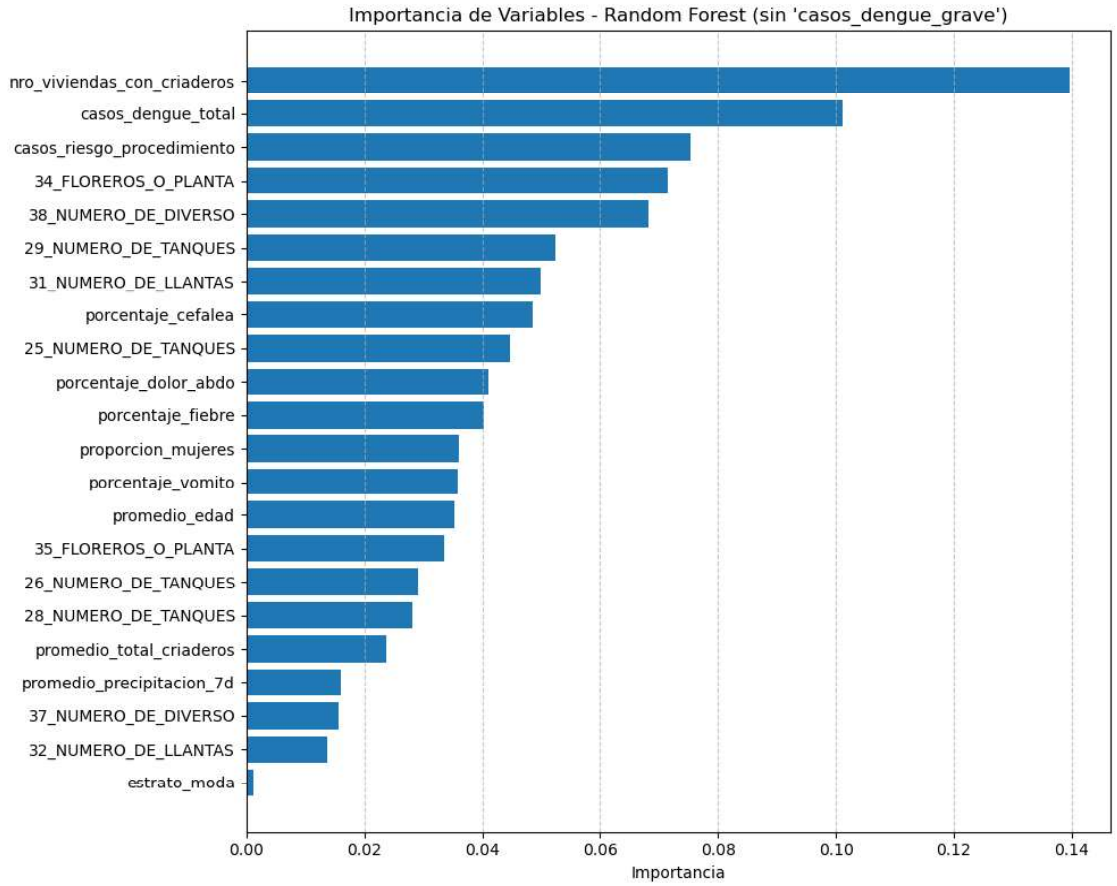

Figura 4.3: Importancia de variables según modelo Random Forest (excluyendo `casos_dengue_grave`)

Variables más relevantes

Los cinco predictores más importantes en el modelo fueron:

1. **nro_viviendas_con_criaderos** (0.1396): principal variable predictora, representa la magnitud de exposición al vector en el entorno domiciliario.

2. **casos_dengue_total** (0.1011): indicador epidemiológico clave que refleja transmisión reciente o persistente, aún sin considerar casos graves.

3. **casos_riesgo_procedimiento** (0.0753): sugiere que prácticas clínicas inadecuadas podrían asociarse a desenlaces graves.

4. **34_FLOREROS_O_PLANTA** (0.0715): tipo de criadero frecuente en viviendas, con alto potencial para proliferación de vectores.
5. **38_NUMERO_DE_DIVERSO** (0.0681): categoría que agrupa criaderos no convencionales, vinculados a entornos urbanos desordenados.

Estas variables fueron seguidas en importancia por otros tipos de criaderos, como los `tanques` y las `llantas`; síntomas clínicos como `cefalea`, `dolor abdominal`, `fiebre` y `vómito`; así como factores sociodemográficos, entre ellos el `promedio_edad` y la `proporcion_mujeres`. También se destaca el aporte de la variable `promedio_precipitacion_7d` (0.0159), lo que confirma el rol de las condiciones climáticas en la dinámica de transmisión del vector.

Por otro lado, variables como `estrato_moda` (0.0012) y algunas relacionadas con tipos específicos de criaderos (por ejemplo, `32_NUMERO_DE_LLANTAS`) tuvieron un aporte marginal al modelo.

Análisis interpretativo

Los resultados sugieren que el riesgo de dengue grave a nivel territorial se explica por una interacción entre:

- **Factores entomológicos**: número y tipo de criaderos (tanques, floreros, diversos).
- **Factores epidemiológicos**: acumulado de casos reportados y señales clínicas tempranas (cefalea, fiebre).
- **Factores institucionales**: registros de manejo clínico inadecuado.
- **Condiciones climáticas**: influencia de la precipitación reciente en la actividad del vector.
- **Factores sociodemográficos**: edad promedio y características de los hogares.

Estos hallazgos son coherentes con la literatura sobre determinantes del dengue grave y refuerzan la utilidad del modelo para la vigilancia territorial, focalización de intervenciones vectoriales y priorización de zonas de alto riesgo.

Capítulo 5

MODELO - SERVICIOS DE SALUD - Obj3

5.1. Desarrollo del Objetivo específico 3: Inclusión de Variables de Calidad de Atención Médica

Incorporación de Indicadores Institucionales para Evaluar la Calidad de la Atención

Con el propósito de fortalecer la capacidad predictiva del modelo desarrollado en el presente proyecto, se incorporaron variables relacionadas con la calidad de la atención médica recibida por las comunidades diagnosticadas con dengue, desde un enfoque ecológico y territorial. Se seleccionaron tres indicadores que, si bien no evalúan la atención clínica individual, permiten identificar patrones agregados de comportamiento institucional que pueden estar relacionados con la progresión a formas graves de la enfermedad.

- **Promedio de días entre el inicio de síntomas y la notificación del caso** (*promedio_dias_notificacion*): este indicador refleja el retardo en la identificación y notificación del caso, lo cual puede estar asociado a barreras en el acceso, subestimación clínica o deficiencias en la oportunidad diagnóstica.
- **Índice de severidad clínica promedio por barrio** (*indice_severidad_promedio*): se construyó con base en la suma de síntomas de alarma (fiebre, vómito, dolor abdominal, cefalea) reportados en los casos de dengue en cada barrio. Este índice resume la carga clínica promedio de los pacientes atendidos en una zona, lo cual puede ser indicativo de subregistro de gravedad o baja adherencia a las guías de clasificación clínica.
- **Proporción de fuga asistencial** (*proporcion_fuga_asistencial*): mide la fracción de casos con residencia en Girón que fueron notificados en municipios distintos. Esta variable se interpreta como un posible indicador de desconfianza en el sistema de salud local, falta de oferta resolutiva o búsqueda activa de atención especializada en otras jurisdicciones.

Importancia de los Indicadores

A diferencia de indicadores clínicos individuales, estos tres elementos permiten evaluar la calidad institucional desde una perspectiva territorial y poblacional. La *demora diagnóstica* señala posibles fallas en el acceso o capacidad de respuesta. El *índice de severidad clínica* ayuda a identificar zonas con cuadros clínicos complejos que podrían estar siendo subclasificados o subatendidos. Finalmente, la *fuga asistencial* capta el comportamiento colectivo frente a la oferta de servicios local, lo cual puede estar relacionado con barreras percibidas en la atención.

Cuadro 5.1: Ficha técnica de indicadores institucionales incorporados al modelo predictivo

Indicador	**Definición**	**Fuente**	**Interpretación epidemiológica**
Promedio de días entre síntomas y notificación	Promedio de días entre la fecha de inicio de síntomas y la fecha de notificación del caso.	Registros SIVIGILA (variables `ini_sin_` y `fecha_nto_`)	Mayor valor puede reflejar retardo en el acceso, diagnóstico o deficiencia en la oportunidad clínica.
Índice de severidad clínica promedio	Promedio por barrio de la suma de síntomas de alarma (fiebre, vómito, dolor abdominal, cefalea) reportados por caso.	Registros clínicos individuales agregados por barrio	Detecta zonas con mayor carga clínica que podrían estar subclasificadas o subatendidas.
Proporción de fuga asistencial	Proporción de casos con residencia en Girón que fueron notificados en municipios diferentes.	Variable `nmun_resi` vs `nmun_notif` del sistema SIVIGILA	Refleja desconfianza institucional, falta de cobertura local o búsqueda de atención especializada fuera del municipio.

Implicaciones en el Modelo Predictivo

La inclusión de estos indicadores permite que el modelo predictivo integre dimensiones estructurales, clínicas y conductuales del sistema de salud, más allá de la disponibilidad de recursos. Su incorporación mejora la sensibilidad para detectar zonas vulnerables, aún cuando los recursos físicos estén presentes, pero la calidad funcional del servicio esté comprometida.

Desde el punto de vista operativo, estos indicadores ofrecen insumos concretos para la toma de decisiones, facilitando la priorización de intervenciones tanto en términos de fortalecimiento institucional como de vigilancia clínica intensificada en barrios críticos. El modelo, al integrar estos elementos, adquiere no solo mayor precisión, sino también mayor relevancia en el contexto de la gestión en salud pública.

5.2. Desempeño del Modelo Predictivo

A partir del modelo predictivo desarrollado para la identificación de zonas de riesgo de dengue grave en Girón, se implementó un proceso de entrenamiento supervisado basado en el uso de datos integrados multifuente. El preprocesamiento incluyó imputación diferenciada de valores faltantes, escalamiento de variables continuas, balanceo con la técnica *SMOTE* y validación cruzada estratificada ($k = 5$).

Además, se ajustó el umbral de clasificación a **0.4**, lo cual permitió mejorar la sensibilidad sin comprometer significativamente la precisión. El conjunto de entrenamiento fue balanceado con *oversampling*, y el modelo fue evaluado a través de las métricas estándar para la clase positiva (riesgo alto).

Evaluación del Desempeño con Umbral Ajustado (0.4)

Modelo	F1-score promedio	Recall (1) promedio	AUC-ROC promedio	Desv. F1
XGBoost	**0.8643**	0.9412	0.9128	0.0462
Random Forest	0.8596	**0.9529**	**0.9526**	**0.0434**
Regresión Logística	0.7853	0.8824	0.8851	0.0625

Cuadro 5.2: Resultados promedio de validación cruzada con umbral ajustado a 0.4

Resultados

El modelo **XGBoost** presentó el mejor rendimiento global, obteniendo el mayor **F1-score promedio (0.8643)**, lo que refleja un equilibrio sobresaliente entre precisión y sensibilidad. También logró una alta capacidad discriminativa con un **AUC-ROC de 0.9128**, y mantuvo una baja variabilidad en su rendimiento (**desviación estándar del F1-score: 0.0462**).

El modelo **Random Forest** destacó por su excelente sensibilidad (**0.9529**) y el mayor **AUC-ROC (0.9526)** del conjunto de modelos evaluados. Su bajo nivel de desviación (**0.0434**) lo posiciona como el modelo más estable en la detección de zonas de riesgo real.

Por su parte, la **Regresión Logística**, aunque con menor desempeño general, logró un **F1-score de 0.7853** y un **AUC-ROC de 0.8851**. Este modelo sigue siendo una opción adecuada cuando la interpretabilidad y la trazabilidad de los coeficientes son prioridades analíticas.

Conclusiones

Con base en los resultados anteriores, se concluye que:

- **Random Forest** es el modelo más equilibrado y confiable para la predicción de zonas de riesgo de dengue grave. Su alta sensibilidad, estabilidad y capacidad discriminativa lo hacen especialmente útil en contextos de salud pública.
- **XGBoost** representa una alternativa potente, especialmente cuando se requieren modelos que capturen relaciones no lineales complejas con alta precisión.
- **Regresión Logística**, aunque interpretativamente ventajosa, es menos precisa y más variable frente a los modelos basados en árboles. Su uso se recomienda como modelo base o de comparación.

En conjunto, estos modelos permiten anticipar de forma precisa las zonas prioritarias para intervenciones en salud pública, apoyando estrategias de vigilancia intensificada y asignación eficiente de recursos en entornos territoriales de alta vulnerabilidad frente al dengue grave.

5.3. Importancia de las Variables - Random Forest

La Figura 5.1 presenta la importancia relativa de las variables utilizadas en el modelo **Random Forest**, calculada con base en la ganancia media de impureza (*mean decrease in impurity*). Los resultados permiten identificar los principales factores asociados al riesgo de dengue grave a nivel territorial.

Destaca como la variable más relevante `casos_dengue_total` (**0.129**), indicador directo de carga epidemiológica acumulada en cada barrio, lo que respalda su papel como determinante clave de la clasificación de zonas de riesgo. Le sigue `nro_viviendas_con_criaderos` (**0.111**), un componente entomológico esencial que refleja la exposición ambiental al vector en entornos domésticos.

También se identifican con pesos importantes variables ambientales como `34_FLOREROS_O_PLANTA` (**0.072**) y `38_NUMERO_DE_DIVERSO` (**0.071**), que capturan criaderos no convencionales. Igualmente, variables clínicas como `casos_riesgo_procedimiento` (**0.053**) y `porcentaje_cefalea` (**0.050**) muestran una contribución significativa al modelo, revelando la importancia de aspectos vinculados al manejo clínico inadecuado y la expresión sintomática de los casos.

Entre los determinantes institucionales, se destaca:

- `promedio_dias_notificacion` (**0.021**), que captura la demora promedio entre el inicio de síntomas y la notificación oficial del caso. Esta variable permite evaluar posibles barreras en la oportunidad diagnóstica o la vigilancia en salud pública.
- `indice_severidad_promedio` (**0.028**), indicador sintético que resume la intensidad clínica de los síntomas clave, útil para captar subclasificaciones clínicas o manejo ambulatorio inadecuado.
- `proporcion_fuga_asistencial` (**0.015**), que refleja el porcentaje de pacientes que consultan fuera del municipio, señalando posibles deficiencias en la oferta institucional local o percepción de baja calidad del servicio.

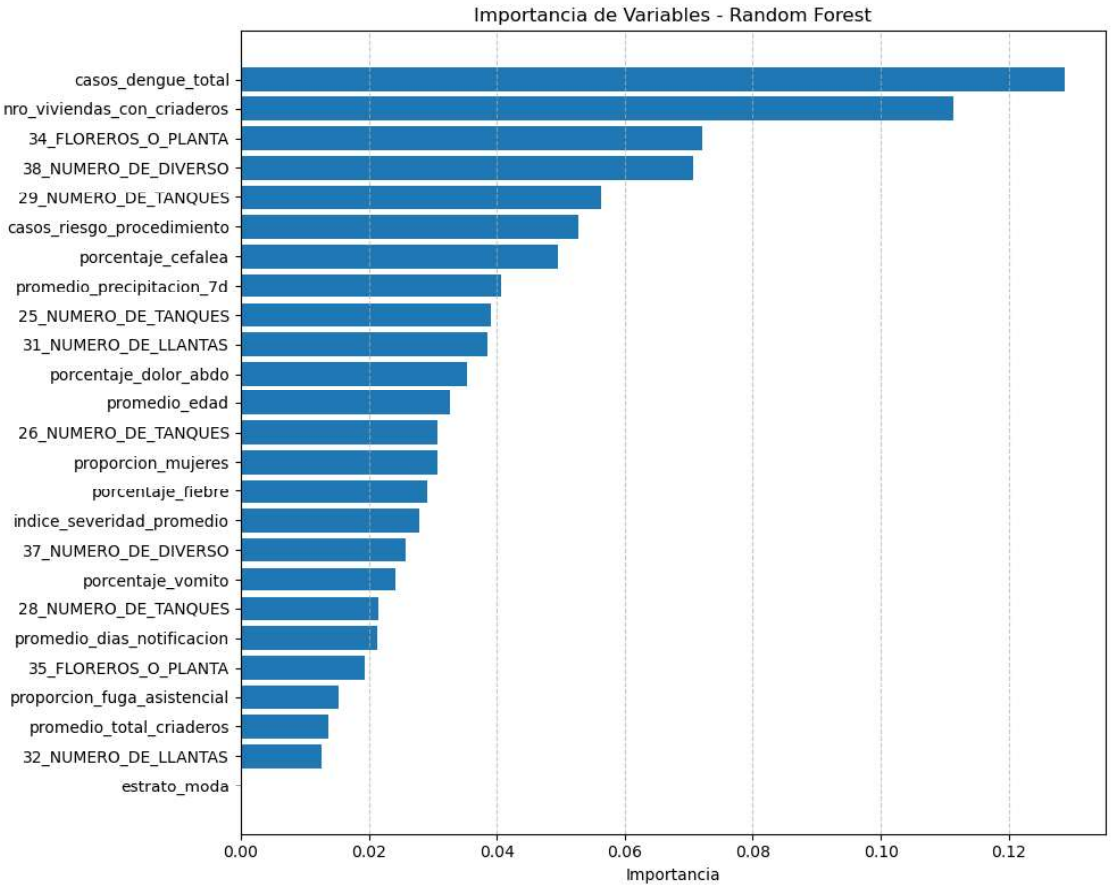

Figura 5.1: Importancia de Variables - Modelo Random Forest

En contraste, algunas variables presentaron importancia marginal o nula, como `estrato_moda` (**0.000**) y `32_NUMERO_DE_LLANTAS` (**0.013**), lo que podría explicarse por su baja variabilidad entre barrios o por su débil relación predictiva en este contexto territorial específico.

En conjunto, el análisis evidencia que el modelo Random Forest logró capturar una combinación equilibrada de factores epidemiológicos (carga de casos), entomológicos (criaderos), clínicos (síntomas y severidad), ambientales (precipitación) e institucionales (fuga y oportunidad diagnóstica), lo que fortalece su aplicabilidad como herramienta para la toma de decisiones en salud pública territorial.

Análisis de Correlación entre Variables Predictoras

El mapa de correlación presentado en la Figura 5.2 revela un bajo nivel de multicolinealidad entre las variables incluidas en el modelo. Solo se identificó una correlación crítica (r ¿0.90) entre `31_NUMERO_DE_LLANTAS` y `32_NUMERO_DE_LLANTAS` (**r = 0.98**), lo que sugiere redundancia y la necesidad de excluir una de ellas para evitar sesgos de ponderación en modelos lineales.

Adicionalmente, se observa una correlación elevada entre `casos_dengue_total` y `casos_riesgo_procedimiento` (**r = 0.83**), consistente con la hipótesis epidemiológica que vincula el volumen de casos con la probabilidad de errores clínicos. Sin embargo, al tratarse de dominios distintos (frecuencia vs. calidad asistencial), se mantiene ambas variables en modelos no lineales.

La mayoría de las demás variables presentan correlaciones menores a 0.5, lo que respalda la diversidad informativa del conjunto de predictores y sugiere una adecuada estructuración multivariada para el modelado supervisado.

5.4. Discusión e Impacto

Los resultados obtenidos mediante una validación cruzada estratificada y balanceo de clases permiten consolidar un análisis realista sobre la capacidad predictiva de los modelos. El modelo **Random Forest** se destacó por su alta sensibilidad (**0.95**) y su AUC-ROC superior a **0.95**, evidenciando una excelente capacidad para detectar zonas de riesgo sin omitir casos relevantes, lo cual es esencial en contextos de salud pública. Por su parte, el modelo **XGBoost** mostró el mejor equilibrio global entre precisión y sensibilidad, alcanzando el mayor **F1-score promedio (0.864)**, lo que lo convierte en una alternativa robusta cuando se busca un balance en el desempeño del modelo.

El análisis de correlación evidenció una baja multicolinealidad entre los predictores, lo que valida la estructura del conjunto de variables utilizadas. Solo una pareja de variables presentó correlación crítica (`31_NUMERO_DE_LLANTAS` y `32_NUMERO_DE_LLANTAS`), lo que fue considerado para el control de redundancia en modelos interpretables como la regresión logística. Además, la coexistencia de factores clínicos, ambientales e institucionales con baja redundancia sugiere que el modelo captura adecuadamente la complejidad territorial del riesgo de dengue grave.

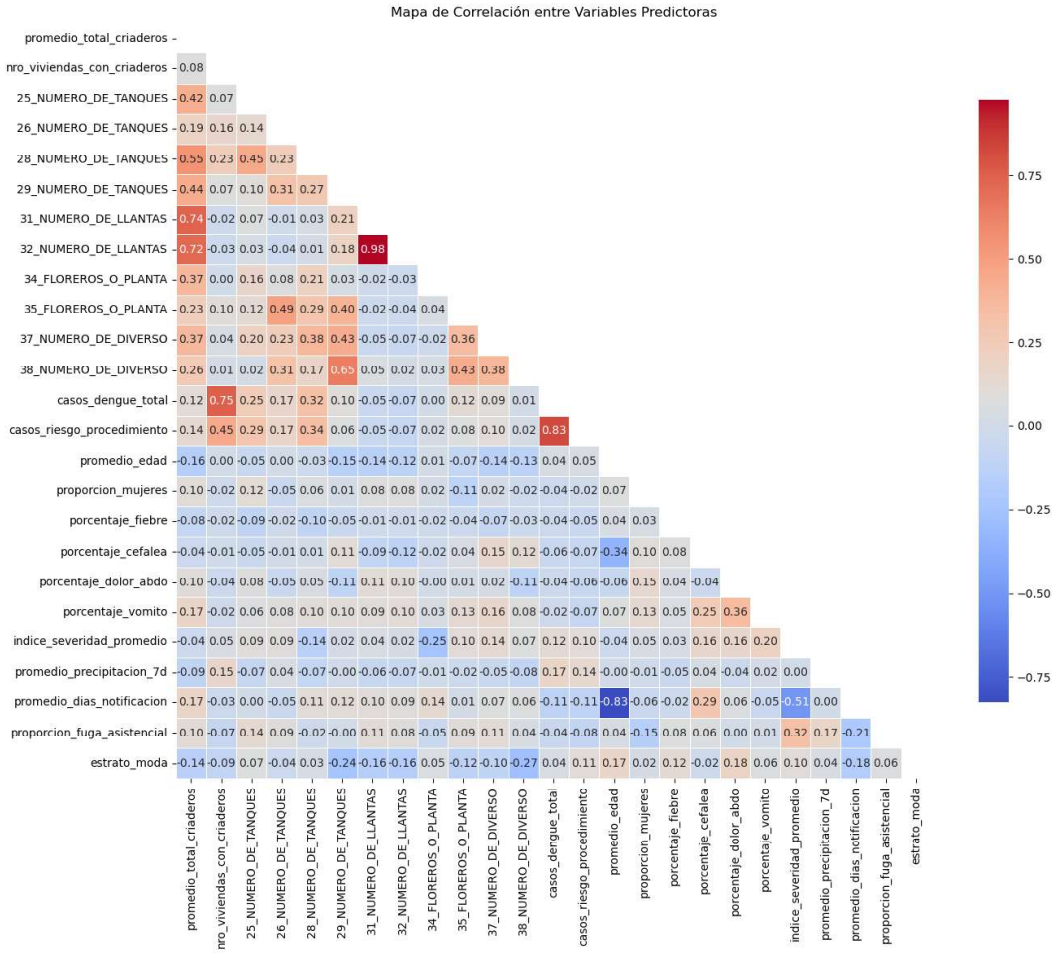

Figura 5.2: Mapa de correlación entre variables predictoras.

Desde una perspectiva operativa, la utilidad de estos modelos radica en su capacidad para anticipar con precisión las zonas vulnerables a formas graves de dengue. Esta anticipación permite priorizar recursos hacia vigilancia intensificada, intervenciones vectoriales, acciones de educación comunitaria y monitoreo clínico más focalizado. En escenarios con restricciones presupuestales, estas herramientas aportan a una gestión territorial basada en evidencia, mejorando la eficiencia y el impacto de las decisiones en salud pública.

A futuro, la integración de datos en tiempo real —como precipitaciones acumuladas, calidad de atención por IPS, movilidad poblacional y densidad vectorial— podrá enriquecer la capacidad explicativa del modelo. Asimismo, se recomienda avanzar en la validación externa del modelo en otros municipios del área metropolitana o departamentos endémicos, para evaluar su generalización y consolidarlo como una herramienta escalable y sostenible para el control del dengue en Colombia.

Mapa de Zonas de Riesgo Alto de Dengue en Girón

Con el objetivo de representar espacialmente las zonas de riesgo alto de dengue en el municipio de Girón, se desarrolló un mapa utilizando técnicas de análisis geoespacial y visualización en Python. Para ello, se integró información geográfica de los barrios de Girón en formato GeoJSON con una base de datos consolidada que contenía indicadores epidemiológicos y sociodemográficos obtenidos como resultado del modelo predictivo.

La unión de ambas fuentes de datos se realizó a través del nombre normalizado del barrio, asegurando coherencia entre los identificadores geográficos y los registros de indicadores. Posteriormente, se empleó la biblioteca `Plotly Express` para construir un mapa tipo *choropleth*, donde cada polígono representa un barrio y su color corresponde al nivel de riesgo estimado de letalidad por dengue, de acuerdo con la variable `zona_riesgo_alta`. La escala de colores utilizada (*Reds*) permite una interpretación visual intuitiva, donde los tonos más oscuros indican mayor nivel de riesgo.

Esta visualización permite identificar de forma clara las zonas prioritarias de intervención, facilitando así la toma de decisiones informadas para la focalización de estrategias de prevención, control vectorial y mejora en la calidad de los servicios de salud.

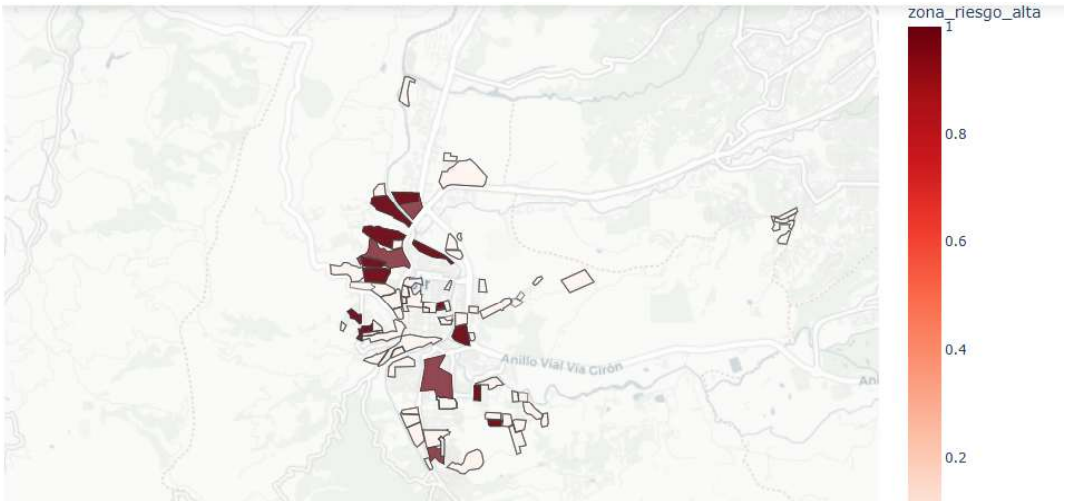

Figura 5.3: Mapa coroplético de barrios del municipio de Girón con clasificación de riesgo alto de letalidad por dengue, según resultados del modelo predictivo. Los colores representan el nivel de riesgo, siendo los tonos más oscuros aquellos con mayor prioridad de intervención.

Leyenda de interpretación:

- **Rojo claro**: Riesgo bajo o no clasificado como prioritario.
- **Rojo medio**: Zona en vigilancia por condiciones de riesgo moderadas.
- **Rojo oscuro**: Barrio clasificado como zona de riesgo alto de letalidad por dengue.

Capítulo 6

Conclusiones y trabajos futuros

6.1. Conclusiones

El desarrollo del modelo predictivo permitió identificar zonas de mayor riesgo de dengue grave en el municipio de Girón, integrando múltiples fuentes de datos y aplicando técnicas de aprendizaje automático bajo un enfoque epidemiológico y territorial. Gracias al preprocesamiento riguroso, la validación cruzada y el balanceo de clases, se logró obtener modelos con desempeño sólido y generalizable. **Random Forest** destacó por su alta sensibilidad y estabilidad, mientras que **XGBoost** logró el mejor equilibrio global entre precisión y sensibilidad, evidenciado por el mayor *F1-score promedio*.

La investigación confirma el papel central que tienen los determinantes sociales, ambientales y de calidad institucional en la progresión hacia formas graves de la enfermedad. La incorporación de modelos explicables, junto con herramientas geoespaciales y métricas robustas, refuerza el valor práctico de este enfoque para la toma de decisiones en salud pública municipal.

6.1.1. Conclusiones Generales

Factores clave en el riesgo de dengue grave: Las variables más relevantes en la predicción incluyeron la carga entomológica (*nro_viviendas_con_criaderos*), número de casos totales, síntomas clínicos, indicadores de riesgo por conducta médica, precipitación acumulada y, especialmente, variables relacionadas con la calidad de la atención (*demora en notificación*, *fuga asistencial*).

Utilidad operativa para la gestión territorial: La implementación de mapas de riesgo y clasificación por barrio permitió identificar zonas prioritarias para la intervención. Estas visualizaciones aportan valor en la focalización de recursos, especialmente en entornos de restricción presupuestal, contribuyendo a una planeación más estratégica de acciones de control vectorial, vigilancia epidemiológica y refuerzo en la capacidad de atención médica.

Valor agregado en salud pública: La inclusión de indicadores de calidad de atención como variables explicativas del modelo representa un avance metodológico relevante. Este enfoque no solo predice riesgo, sino que también permite auditar el funcionamiento del sistema de salud a nivel

territorial, visibilizando barreras de acceso, subregistro o disfunciones en la red de atención.

6.1.2. Conclusiones por Objetivo Específico

Objetivo 1: Recolectar datos epidemiológicos, climáticos, sociodemográficos y socioeconómicos que influencien la incidencia de dengue grave.

- Se consolidó una base de datos integrada a nivel de barrio, combinando fuentes como SIVIGILA, visitas ETV, IDEAM y caracterización territorial, homogenizando los datos espacialmente.
- Se identificaron patrones espaciales de concentración de riesgo en zonas con alta carga entomológica, mayor volumen de casos y menor acceso institucional.
- Se generaron variables derivadas (síntomas agregados, promedio de criaderos, lluvia acumulada, tiempo de notificación) que aumentaron la capacidad explicativa del modelo.

Objetivo 2: Crear un modelo predictivo que integre los datos sociodemográficos, socioeconómicos y de calidad de los servicios de salud para identificar zonas de alto riesgo.

- Se entrenaron modelos de clasificación (Regresión Logística, Random Forest, XGBoost) con técnicas de imputación diferenciada, balanceo de clases (SMOTE), escalamiento y validación cruzada estratificada.
- Random Forest obtuvo el mejor AUC-ROC promedio (0.95) y la mayor sensibilidad (0.95), mientras que XGBoost presentó el mayor F1-score promedio (0.86), evidenciando robustez en múltiples métricas.
- Se evitó el sobreajuste mediante control de fuga de información, eliminación de redundancias (correlación ¿0.9) y reducción de variables irrelevantes.

Objetivo 3: Incluir en el modelo variables relacionadas con la calidad de la atención médica.

- Se integraron indicadores como `promedio_días_notificación`, `proporcion_fuga_asistencial` e `indice_severidad_promedio`, que resultaron ser predictivos y operativamente relevantes.
- Su inclusión visibilizó el impacto de la oportunidad diagnóstica, continuidad del servicio y percepción institucional como determinantes del agravamiento clínico.
- Se destaca el valor de estos indicadores como herramientas de monitoreo interno del sistema de salud y como insumo para auditoría territorial.

6.1.3. Trabajos Futuros

- Validación externa en municipios del área metropolitana de Bucaramanga y departamentos endémicos con contextos similares.
- Integración del modelo en dashboards institucionales con alertas tempranas y visualización dinámica georreferenciada.
- Automatización del monitoreo con datos en tiempo real desde estaciones meteorológicas y sistemas de salud (APIs).
- Aplicación de técnicas de interpretabilidad (SHAP, LIME) para fortalecer la explicabilidad del modelo ante equipos técnicos y de gestión.
- Adaptación metodológica a otras enfermedades transmitidas por vectores (chikungunya, zika) con ajuste de variables clave.
- Análisis de costo-efectividad de estrategias basadas en predicción comparadas con abordajes homogéneos.

Bibliografía

[1] W. H. Organization, "Dengue and severe dengue," 2020, [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue, [Accessed: 09-Aug-2025].

[2] D. J. Gubler, "Dengue and dengue hemorrhagic fever," *Clinical Microbiology Reviews*, vol. 11, no. 3, pp. 480–496, 1998. doi: 10.1128/CMR.11.3.480 .

[3] C. Williams and Y. Ma, "Machine learning for predicting dengue outbreak using climate indicators," *BMC Medical Informatics and Decision Making*, vol. 19, no. 1, p. 233, 2019. doi: 10.1186/s12911-019-0973-2 .

[4] M. C. Castro and S. C. Weaver, "Socio-demographic determinants of dengue fever knowledge, attitudes, and practices in medellín, colombia," *BMC Public Health*, vol. 18, no. 1, p. 143, 2018. doi: 10.1186/s12889-018-5056-9 .

[5] I. K. Yoon, A. Srikiatkhachorn, and A. L. Rothman, "Health service quality and its impact on dengue mortality in southeast asia," *Journal of Infectious Diseases*, vol. 220, no. 1, pp. 138–146, 2019. doi: 10.1093/infdis/jiz071 .

[6] M. N. Karim, S. U. Munshi, N. Anwar, and M. S. Alam, "Exploring the socioeconomic factors influencing dengue fever outbreaks: A case study in dhaka, bangladesh," *Asian Pacific Journal of Tropical Medicine*, vol. 5, no. 9, pp. 678–682, 2012. doi: 10.1016/S1995-7645(12)60139-2 .

[7] U. Haque and et al., "Spatial and temporal patterns of dengue transmission in dhaka, bangladesh," *PLoS ONE*, vol. 7, no. 11, p. e46812, 2012. doi: 10.1371/journal.pone.0046812 .

[8] A. L. Kraemer and et al., "The global distribution of the arbovirus vectors aedes aegypti and ae. albopictus," *eLife*, vol. 4, p. e08347, 2015. doi: 10.7554/eLife.08347 .

[9] M. Chadsuthi and et al., "Modeling the effects of weather and interventions on dengue incidence in thailand," *PLoS ONE*, vol. 7, no. 11, p. e49721, 2012. doi: 10.1371/journal.pone.0049721 .

[10] A. Souza, M. Teixeira, and R. Lima, "Using machine learning algorithms for dengue risk prediction: A case study in brazil," *Tropical Medicine & International Health*, vol. 24, no. 4, pp. 414–422, 2019. doi: 10.1111/tmi.13294 .

[11] D. Phung and et al., "Geospatial and machine learning approaches to predict dengue incidence in thailand," *International Journal of Health Geographics*, vol. 14, no. 1, p. 35, 2015. doi: 10.1186/s12942-015-0025-9 .

[12] T. Fawcett, "An introduction to roc analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006. doi: 10.1016/j.patrec.2005.10.010 .

[13] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets," *PLoS ONE*, vol. 10, no. 3, p. e0118432, 2015. doi: 10.1371/journal.pone.0118432 .

[14] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. doi: 10.1023/A:1010933404324 .

[15] P. Chakraborty and S. Ghosh, "Machine learning in epidemiology: Applications and challenges," *Annual Review of Biomedical Data Science*, vol. 5, pp. 101–125, 2022. doi: 10.1146/annurev-biodatasci-112921-114827 .

[16] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794. doi: 10.1145/2939672.2939785 .

[17] H. He and E. A. Garcia, *Learning from Imbalanced Data.* IEEE Transactions on Knowledge and Data Engineering, 2009, vol. 21, no. 9. doi: 10.1109/TKDE.2008.239 .

[18] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," in *Journal of Artificial Intelligence Research*, vol. 16, 2002, pp. 321–357. doi: 10.1613/jair.953 .

[19] R. Lowe and et al., "Using climate and environmental data to forecast dengue epidemics in latin america," *Environmental Research Letters*, vol. 6, no. 3, p. 034017, 2011. doi: 10.1088/1748-9326/6/3/034017 .

[20] N. A. Honório and et al., "Spatial evaluation and modeling of dengue seroprevalence and vector density in rio de janeiro, brazil," *PLoS Neglected Tropical Diseases*, vol. 3, no. 11, p. e545, 2009. doi: 10.1371/journal.pntd.0000545 .

[21] M. G. Teixeira and et al., "Urbanization and the spread of dengue in salvador, brazil," *Tropical Medicine & International Health*, vol. 7, no. 2, pp. 123–127, 2002. doi: 10.1046/j.1365-3156.2002.00836.x .