PREDICCIÓN DE LA TASA DE DENGUE A TRAVÉS DE MÉTODOS DE MACHINE LEARNING EN EL VALLE DEL CAUCA

Víctor Hugo Cifuentes Rodríguez
María Alejandra Ibarra Calvache
Gregory David Díaz Barrios

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

Directora
Delia Ortega Lenis

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, DICIEMBRE 5 DE 2024

TABLA DE CONTENIDO

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

INTRODUCCIÓN

El dengue, una enfermedad viral transmitida por mosquitos del género *Aedes*, ha emergido como un importante problema de salud global, especialmente en regiones tropicales y subtropicales. La Organización Mundial de la Salud (OMS) estima que alrededor de 390 millones de infecciones por dengue ocurren anualmente, con una incidencia creciente y consecuencias significativas en la salud pública [1].

En América Latina, el dengue se ha convertido en una de las principales amenazas epidemiológicas, debido a las condiciones climáticas favorables para la reproducción del mosquito Aedes aegypti, el principal vector de la enfermedad. Según la Organización Panamericana de la Salud (OPS), países como Brasil, México, y Colombia se encuentran entre los más afectados, con brotes recurrentes y una tendencia ascendente en la incidencia de casos en los últimos años [2].

En Colombia, el dengue es una enfermedad endémica, y su impacto ha sido particularmente grave en regiones tropicales como el Caribe, la Amazonía y el Valle del Cauca. El Instituto Nacional de Salud (INS) ha reportado un aumento significativo en los casos durante la última década, con brotes que coinciden con la temporada de lluvias y se ven exacerbados por factores como el cambio climático y la urbanización descontrolada [3]. La población vulnerable en áreas con infraestructura deficiente enfrenta un mayor riesgo de contagio debido a la falta de medidas efectivas de control vectorial.

El Valle del Cauca, una región caracterizada por su clima cálido y húmedo, ha sido históricamente una de las zonas más afectadas por el dengue en Colombia. Su capital, Cali, concentra una gran parte de los casos reportados anualmente en el país, debido a su alta densidad poblacional y condiciones propicias para la proliferación del mosquito transmisor. Las características urbanas y los déficits en servicios básicos como el agua potable y saneamiento en algunas zonas agravan la situación, lo que hace de la región un foco de transmisión constante [4].

Ante esta situación, la capacidad de predecir y mitigar la propagación de enfermedades infecciosas emergentes o re-emergentes como el dengue se ha convertido en un área vital de investigación en salud pública. El uso de técnicas de Machine Learning ha demostrado ser una herramienta prometedora para analizar y predecir patrones epidemiológicos. Estas técnicas permiten analizar grandes conjuntos de datos y extraer patrones complejos, lo que facilita la identificación de factores predictivos clave para comprender y prevenir la propagación del virus del dengue.

El propósito principal de este proyecto es desarrollar y aplicar modelos predictivos basados en técnicas de Machine Learning para estimar el riesgo de propagación del dengue en los municipios del Valle del Cauca. Se utilizaron conjuntos de datos epidemiológicos, variables climáticas y demográficas, y se realizó un análisis descriptivo que permitió identificar patrones y factores predictivos asociados con la

incidencia de la enfermedad.

1. DEFINICION DEL PROBLEMA

1.1 PLANTEAMIENTO DEL PROBLEMA

El dengue representa un desafío significativo para la salud pública en numerosas regiones tropicales y subtropicales. Con un incremento sostenido en la incidencia de casos, se ha convertido en una prioridad de salud global debido a su impacto en la población y los sistemas de salud.

La Unidad de Vigilancia de la Salud [4], ha concluido que factores climáticos como la lluvia y la temperatura influyen directamente en la formación de charcos que sirven como hábitat e incubadora para las larvas del mosquito *Aedes Aegyptus*, transmisor del dengue, por lo que considerando el cambio climático que actualmente atraviesa el planeta, impactan en gran escala el riesgo de dengue en población que reside en zonas tropicales, lo que se deriva en un aumento significativo de atención hospitalaria.

Según la Organización Mundial de la Salud (OMS) y la Organización Panamericana de la Salud (OPS) [5], en América cerca de 500 millones de personas están en riesgo de contraer dengue; además, se ha evidenciado un incremento en el número de infecciones durante las últimas 4 décadas, llegando a cifras de hasta 16.2 millones de contagios. El Ministerio de Salud de Colombia [6], reporta que cada año se registran alrededor de 50 millones de casos de dengue en el mundo, con aproximadamente 500,000 personas hospitalizadas debido a formas graves de la enfermedad y unas 20,000 víctimas fatales. Durante las epidemias, hasta el 80 o 90% de los individuos susceptibles pueden resultar afectados, y la tasa de mortalidad puede exceder el 2%, siendo los más afectados los niños menores de 15 años. En el 2010, Colombia presentó la más grande epidemia por dengue descrita en su historia, con más de 150.000 casos, 9.482 de ellos graves y 217 muertes confirmadas.

El control y prevención del dengue se basan en estrategias de vigilancia epidemiológica y vectorial. Sin embargo, la naturaleza compleja y multifactorial de la propagación del dengue dificulta su predicción y control efectivos. Los métodos tradicionales de vigilancia han demostrado limitaciones para anticipar y responder de manera proactiva a los brotes, lo que conlleva a respuestas reactivas en lugar de preventivas, ante la propagación de la enfermedad.

El reto clave radica en la capacidad limitada de predecir y comprender la dinámica de propagación del dengue, considerando la interacción de múltiples variables, como factores epidemiológicos, climáticos y socioeconómicos. Por lo tanto, la investigación propuesta se centró en la creación de modelos predictivos que utilicen datos epidemiológicos, climáticos y demográficos para predecir la propagación de dengue.

En la literatura, se evidenció que la predicción del riesgo de dengue ha tenido -mayoritariamente- un enfoque desde la estadística y matemática clásica, no obstante, el desafío principal en este proyecto consistió en utilizar técnicas avanzadas de Machine Learning para analizar conjuntos de datos complejos y multidimensionales, a fin de identificar patrones y relaciones entre variables que permitan la predicción del riesgo en términos del clima y variables socioeconómicas, en áreas geográficas específicas del territorio colombiano. El objetivo fue, además, proporcionar una herramienta práctica para la toma de decisiones en salud pública y la implementación de estrategias preventivas más eficientes.

1.2 FORMULACIÓN DEL PROBLEMA

- **¿Cómo predecir la tasa de Dengue en el Valle del Cauca, usando técnicas de Machine Learning?**
- ¿Cómo obtener información sobre variables climáticas, sociales y demográficas disponibles en el Valle del Cauca?
- ¿Cuáles variables son relevantes para predecir el riesgo de dengue?
- ¿Cuál modelo de Machine Learning proporciona mejores métricas de desempeño en la predicción del riesgo de dengue en el Valle del Cauca?

2. OBJETIVOS DEL PROYECTO

2.1 OBJETIVO GENERAL

Desarrollar un modelo predictivo de la tasa de dengue en el Valle del Cauca, utilizando técnicas de Machine Learning que integre variables climáticas, sociales y demográficas.

2.2 OBJETIVOS ESPECIFICOS

- Construir la base de datos a través de información histórica de incidencia de dengue, a partir de factores climáticos, sociales y demográficos.
- Realizar un análisis exploratorio detallado de los datos para identificar patrones, correlaciones y posibles variables relevantes que influyan en la incidencia de dengue.
- Entrenar modelos de Machine Learning para predecir el riesgo de dengue a partir de variables climáticas, sociales y demográficas.
- Evaluar los modelos de Machine Learning, mediante validación cruzada y métricas de desempeño, para elegir el mejor modelo de predicción de riesgo de dengue en el Valle del Cauca.

3. MARCO TEÓRICO Y ANTECEDENTES

3.1 DENGUE Y SU CONTEXTO EPIDEMIOLÓGICO

- **Definición de Dengue:** Según D. J. Gubler (1998) [7], el dengue es una enfermedad viral transmitida por mosquitos, caracterizada por síntomas como fiebre aguda y dolores musculares y articulares. Es causada por el virus del dengue, del género Flavivirus, y se transmite principalmente por mosquitos del género Aedes, especialmente *Aedes Aegypti*. La Organización Mundial de la Salud [8] indica que la enfermedad presenta un patrón estacional, con un mayor número de casos en la primera parte del año en el hemisferio sur y en la segunda mitad en el hemisferio norte. En Colombia, el dengue se clasifica en tres formas principales según su severidad: (1) dengue sin signos de alarma, que es la forma más leve; (2) dengue con signos de alarma, que presenta síntomas como dolor abdominal severo, vómitos persistentes y sangrado de mucosas, y requiere hospitalización; y (3) dengue grave, que es una emergencia médica y puede causar complicaciones mortales como choque hipovolémico o hemorragias masivas, lo que requiere atención en unidades de cuidados intensivo [6].

- **Epidemia:** De acuerdo con el observatorio de Medicina de la Pontificia Universidad Católica de Chile [9], se define como "una enfermedad que se propaga rápida y activamente, aumentando significativamente el número de casos en un área geográfica concreta."

- **Anticuerpos para la detección del dengue:** Los anticuerpos juegan un papel clave en la detección del dengue, principalmente los anticuerpos IgM e IgG, los cuales se generan en respuesta a la infección por el virus. Los anticuerpos IgM son los primeros en aparecer y son indicativos de una infección reciente, mientras que los IgG indican una infección pasada o inmunidad. Las pruebas de detección de anticuerpos IgM e IgG, como los ensayos inmunoenzimáticos (ELISA), son métodos estándar para confirmar la presencia del virus en los pacientes [10].

3.2 METODOLOGÍA DE INVESTIGACIÓN

Metodología CRISP-DM: Según Schröern (2021) [11]esta metodología se describe como un modelo de proceso independiente de la industria para la minería de datos, ampliamente utilizado en ciencia de datos. Consta de seis fases iterativas:

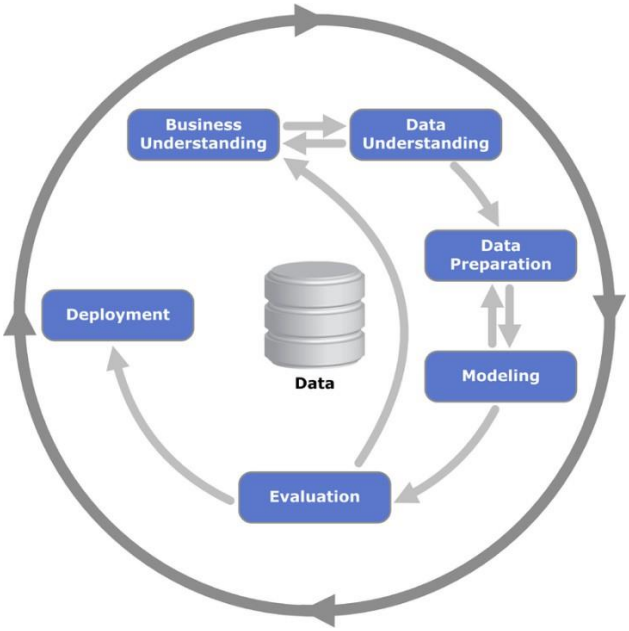

Figura 1: Metodología CRIPS-DM

Fuente: https://healthdataminer.com/data-mining/crisp-dm-una-metodologia-para-mineria-de-datos-en-salud/

.1 Comprensión del problema: Definir claramente el problema, requisitos y limitaciones.
.2 Comprensión de datos: Recolección, descripción y exploración de datos.
.3 Preparación de datos: Limpieza y transformación de datos.
.4 Modelado: Ajuste de modelos a implementar y selección de datos de prueba.
.5 Evaluación: Métricas de desempeño y selección del modelo óptimo.
.6 Implementación: Documentación y monitoreo de acciones futuras.

3.3 MODELOS DE PREDICCIÓN PARA ENFERMEDADES EPIDEMIOLOGICAS:

- **Long Short-Term Memory (LSTM):** El modelo Long Short-Term Memory (LSTM) es una variante de red neuronal recurrente (RNN) diseñada específicamente para aprender dependencias a largo plazo en secuencias de datos. Debido a su capacidad para retener y filtrar información relevante en secuencias temporales, el LSTM se ha convertido en una herramienta eficaz en la predicción de series temporales, como la propagación de epidemias y enfermedades transmisibles, incluyendo el dengue.

  Según Hochreiter y Schmidhuber (1997) [12], el LSTM supera las limitaciones de las RNN tradicionales al utilizar una arquitectura basada en "celdas de memoria" y mecanismos de puerta (entrada, olvido y salida) que regulan el flujo de información a través de la red. Cada puerta desempeña una función específica: la puerta de entrada controla qué

información de la entrada actual se debe almacenar en la celda, la puerta de olvido decide qué parte de la información pasada debe ser desechada, y la puerta de salida regula qué información de la celda se utiliza en la salida actual. Estas características permiten al LSTM capturar patrones temporales cruciales, resultando especialmente útil en contextos donde los cambios a lo largo del tiempo son fundamentales, como en la evolución de brotes epidémicos.

En el contexto de la predicción del dengue, el modelo LSTM puede procesar datos históricos de casos, condiciones climáticas y variables socioeconómicas para identificar patrones temporales complejos y predecir la aparición de futuros brotes. La capacidad del LSTM para manejar datos no lineales y de alta dimensionalidad ha sido confirmada en diversos estudios epidemiológicos, los cuales demuestran su eficacia en modelar factores con relaciones temporales complejas y en brindar precisión en la predicción de enfermedades contagiosas [13].

- **Regresión Lineal**: La regresión lineal es uno de los modelos más básicos y fundamentales en estadística y aprendizaje automático. Busca modelar la relación entre una variable dependiente Y y una o más variables independientes X a través de una ecuación lineal:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon$$

Aquí, β representa los coeficientes de regresión que cuantifican el efecto de cada variable independiente en Y, mientras que ϵ es el término de error.
La técnica más común para ajustar los coeficientes es el método de mínimos cuadrados ordinarios (OLS), que minimiza la suma de los errores al cuadrado. A pesar de su simplicidad, la regresión lineal tiene limitaciones: es sensible a la multicolinealidad entre las variables independientes y no es adecuada para capturar relaciones no lineales sin transformaciones adicionales [14], [15].
Además, para evaluar la bondad de ajuste, se utilizan métricas como el coeficiente de determinación (R2R^2R2) y el error cuadrático medio (MSE), siendo esenciales en el análisis de los resultados [16].

- **Ridge Regression**: Ridge Regression aborda el problema de la multicolinealidad agregando una penalización L2 al error cuadrático. Su función objetivo es:

$$MSE + \lambda \sum_{i=1}^{n} \beta_i^2$$

Donde λ es un hiperparámetro que controla la cantidad de regularización. Este método reduce la magnitud de los coeficientes, evitando que variables correlacionadas tengan un impacto desproporcionado en el modelo.

La regularización L2 es particularmente útil en contextos de alta dimensionalidad, como la genómica o análisis de mercados financieros, donde las variables predictoras suelen ser redundantes. Sin embargo, a diferencia de Lasso, no realiza una selección de variables explícita, ya que no puede reducir coeficientes exactamente a cero [17], [18]

- **Lasso Regression**: Lasso (Least Absolute Shrinkage and Selection Operator) mejora la regresión lineal añadiendo una penalización L1:

$$MSE + \lambda \sum_{i=1}^{n} |\beta_i|$$

A diferencia de Ridge, Lasso puede forzar algunos coeficientes a cero, lo que equivale a eliminar variables irrelevantes del modelo. Este atributo lo hace ideal en situaciones donde hay muchas variables, pero solo unas pocas son significativas.
Lasso es particularmente útil en áreas como la biología, donde los datos suelen tener una gran cantidad de predictores, muchos de los cuales son irrelevantes. Sin embargo, puede ser menos estable que Ridge si las variables están altamente correlacionadas [17], [19].

- **Decision Tree**: Los árboles de decisión son modelos no paramétricos que dividen los datos en subconjuntos homogéneos utilizando reglas basadas en valores de las variables predictoras. Se construyen seleccionando la característica que maximiza la ganancia de información o minimiza la impureza (medida con métricas como la entropía o el índice Gini) en cada nodo.

Aunque son interpretables y útiles para problemas tanto de clasificación como de regresión, los árboles de decisión son propensos al sobreajuste, especialmente si no se limitan parámetros como la profundidad máxima del árbol. Por ello, se suelen combinar con métodos de ensamblado como Random Forest y Gradient Boosting para mejorar su generalización [18], [20].

- **Random Forest**: Este algoritmo extiende los árboles de decisión mediante el ensamblado, construyendo múltiples árboles independientes sobre diferentes subconjuntos de datos y características. Al promediar o votar las predicciones de todos los árboles, Random Forest reduce tanto la varianza como el riesgo de sobreajuste.

Su capacidad para manejar grandes conjuntos de datos y relaciones complejas lo convierte en una opción popular en aplicaciones como detección de fraudes, clasificación de imágenes y predicción de resultados médicos. Sin embargo, el modelo puede volverse menos interpretable debido al gran número de árboles involucrados [19], [21].

- **Gradient Boosting**: Gradient Boosting utiliza una técnica secuencial para optimizar la precisión del modelo. Cada árbol subsiguiente corrige los errores residuales del árbol anterior. La función objetivo se ajusta minimizando una pérdida utilizando gradientes descendentes.

Este enfoque captura relaciones no lineales complejas y es efectivo en competiciones de

ciencia de datos. Sin embargo, es computacionalmente intensivo y requiere ajustes cuidadosos de hiperparámetros como la tasa de aprendizaje y el número de árboles para evitar el sobreajuste [20], [22].

- **Support Vector Regression (SVR)**: SVR es una extensión de las máquinas de soporte vectorial (SVM) para problemas de regresión. Intenta encontrar una función que minimice el error dentro de un margen tolerable, lo que lo hace eficaz para problemas donde se busca una predicción precisa,pero se toleran pequeños errores. SVR es adecuado para manejar datos no lineales mediante el uso de trucos de kernel, lo que le permite encontrar patrones en datos de alta dimensionalidad, como en el reconocimiento de imágenes o el análisis financiero [23] .

- **K-Nearest Neighbors (KNN**): KNN es un modelo basado en la similitud de datos. Para una predicción, busca los "K" puntos de datos más cercanos en el espacio de características y calcula el promedio (para regresión) o la mayoría (para clasificación). Aunque es fácil de implementar y útil en aplicaciones con datos bien distribuidos, KNN puede volverse ineficiente con conjuntos de datos grandes, ya que necesita calcular distancias a todos los puntos. Suele combinarse con reducción de dimensionalidad para mejorar su eficiencia en espacios de alta dimensión [24] .

Estructura de datos tipo panel: Los datos tipo panel son un conjunto de observaciones en las que se combina información temporal y transversal. Es decir, se recogen mediciones para múltiples unidades (como individuos, empresas, o regiones) a lo largo del tiempo. Este formato permite estudiar tanto las diferencias entre unidades como los cambios que ocurren en ellas a través del tiempo, ofreciendo ventajas significativas frente a los datos puramente transversales o de series temporales [25].

Las características de este tipo de datos tienen en cuenta:

a) Dimensionalidad Dual: Cada observación incluye un identificador para la unidad de análisis y un indicador de tiempo. Esto resulta en una estructura bidimensional.
b) Heterogeneidad: Los datos tipo panel permiten capturar diferencias no observadas entre las unidades, ya que combinan la información transversal con la temporal [26].
c) Eficiencia Estadística: Dado que contienen más información que los datos únicamente transversales o temporales, permiten estimaciones más precisas y robustas [27]
d) Análisis Dinámico: Se pueden analizar relaciones dinámicas y efectos retardados debido a la naturaleza temporal de los datos.
e) Tamaño y Complejidad: El tamaño y la complejidad de los datos aumentan con el número de unidades y períodos analizados, lo que puede influir en la capacidad de los modelos para procesar esta información [28]

Sin embargo, se cuenta con una serie de limitaciones de los Datos Tipo Panel, las cuales

contemplan:

a) Ausencia de Datos: Es común que existan valores faltantes debido a problemas de seguimiento de las unidades o registros incompletos, lo que puede sesgar los resultados si no se maneja adecuadamente.
b) Heterogeneidad No Observada: Aunque los modelos tipo panel controlan algunos efectos no observados, pueden surgir variables no incluidas que influyan en los resultados [29].
c) Dependencia Temporal y Espacial: Las observaciones pueden estar correlacionadas en el tiempo o entre las unidades, lo que puede requerir métodos avanzados para evitar inferencias incorrectas.
d) Complejidad Computacional: Con grandes volúmenes de datos, los cálculos estadísticos y econométricos pueden ser computacionalmente intensivos.
e) Suposiciones del Modelo: Los métodos empleados, como modelos de efectos fijos o aleatorios, pueden depender de supuestos estrictos que, de no cumplirse, afectarán la validez de los resultados [30].
f) Importancia de los Datos Tipo Panel
A pesar de sus limitaciones, los datos tipo panel son fundamentales para analizar relaciones complejas y dinámicas en áreas como economía, sociología, y epidemiología. Su capacidad para controlar heterogeneidades no observadas y modelar efectos temporales los convierte en una herramienta poderosa para la investigación aplicada y teórica.

- **METRICAS DE EVALUACIÓN DE MODELOS:**
Para evaluar el rendimiento de los modelos predictivos en el contexto epidemiológico, se utilizaron varias métricas de desempeño. A continuación, se presentan las más comunes:

1. **Coeficiente de determinación (R^2)**: Según Draper y Smith (1998) [31] el R^2 mide la proporción de la varianza en la variable dependiente que es explicada por el modelo. Un valor de R^2 cercano a 1 indica que el modelo tiene un buen ajuste.

2. **Error Absoluto Medio (MAE)**: El MAE representa el promedio de los errores absolutos entre los valores predichos y los reales. Es una medida de la precisión de las predicciones y, a diferencia del MSE, no amplifica los errores grandes.

3. **Error Cuadrático Medio (MSE)**: El MSE calcula la media de los errores al cuadrado entre los valores predichos y observados. Es más sensible a errores grandes que el MAE, ya que amplifica los errores mayores al elevarlos al cuadrado.

4. **Raíz del Error Cuadrático Medio (RMSE)**: El RMSE es la raíz cuadrada del MSE y se utiliza para medir la diferencia promedio entre los valores observados y predichos. Es útil

porque proporciona una interpretación en la misma escala de los datos originales.

3.4 TECNOLOGÍAS EMERGENTES EN EPIDEMOLOGÍA

Web Scraping: Esta técnica permite la extracción automatizada de datos de páginas web, convirtiéndose en una herramienta valiosa en la investigación epidemiológica y vigilancia de enfermedades como el dengue. En el contexto del presente estudio, se utiliza para complementar la recopilación de datos sobre dengue en el Valle del Cauca, accediendo a información de fuentes que no están disponibles en medios tradicionales.

El proceso de Web Scraping implica identificar fuentes confiables, desarrollar scripts de extracción (por ejemplo, utilizando bibliotecas de Python como BeautifulSoup o Scrapy), limpiar y transformar los datos, y finalmente integrarlos con la base de datos existente. Este enfoque no solo amplía el acceso a datos para el análisis predictivo, sino que también destaca la importancia de las tecnologías emergentes en la vigilancia de salud pública y la toma de decisiones en el control de enfermedades infecciosas.

3.5 ANTECEDENTES

Para la revisión de antecedentes, se centró en el eje fundamental del proyecto: la predicción del dengue y el uso de modelos de Machine Learning. En este sentido, se identificaron artículos, tesis y proyectos tanto nacionales como internacionales que implementan métodos estadísticos clásicos y técnicas de Machine Learning en un contexto epidemiológico específico. La búsqueda se enfocó en trabajos recientes.

Un artículo estudiado se enfocó en la dinámica geométrica de los casos de dengue en Colombia entre 1990 y 2006, utilizando un enfoque probabilista mediante una caminata al azar. Esta metodología permitió calcular la predicción temporal de los casos anuales de dengue, logrando predecir los casos para el año 2007 con un porcentaje de acierto del 90.4%. Sin embargo, este enfoque presenta limitaciones al no considerar factores adicionales que influyen en la propagación del dengue, adoptando una perspectiva puramente matemática sin integrar variables externas como las condiciones climáticas y sociales. En este caso, las variables utilizadas fueron el número de casos anuales de dengue, pero el modelo no consideró variables socioeconómicas ni climáticas [32].

En otro estudio, se desarrolló una metodología predictiva para estimar la proporción de casos de dengue grave en relación con el total anual de infectados en cada departamento de Colombia, utilizando la teoría de la probabilidad. El análisis de datos desde 2005 hasta 2010 mostró una predicción precisa para el año 2011, con un acierto del 93.3%. Las variables consideradas en este estudio incluyeron la proporción de dengue grave en relación con los infectados y la evolución espacio-temporal de los casos. No obstante, este enfoque no incorporó variables externas como los cambios climáticos y las condiciones sociales, aunque permitió identificar la relación espacio-temporal en la probabilidad de casos de dengue [33].

Un artículo reciente revisó la literatura científica sobre variables y métodos de aprendizaje automático utilizados para detectar la infección por dengue. Se evaluaron 247 artículos, de los cuales 33 mostraron una alta frecuencia de palabras clave relevantes. Entre los hallazgos, se destacó que las redes neuronales convolucionales (CNN) y los perceptrones multicapa (MLP) fueron los métodos más utilizados para la detección del dengue, mientras que los modelos ARIMA y las series de tiempo dominaron en la predicción de brotes. Las variables más comúnmente utilizadas fueron las relacionadas con el clima, como la temperatura, la humedad y las precipitaciones, que se consideraron factores determinantes en el comportamiento de los casos de dengue a nivel global. Esta revisión resaltó la importancia de integrar variables climáticas, como la temperatura y la humedad, en la construcción de modelos predictivos [34].

Un estudio realizado en Costa Rica analizó cómo las variables climáticas influyen en los casos de dengue reportados entre 2007 y 2017. Los análisis estadísticos revelaron una alta correlación no lineal entre las variables de precipitación, temperatura y humedad relativa, lo que llevó a los autores a descartar los modelos de regresión lineal para la predicción del riesgo de dengue. En su lugar, ajustaron un modelo de Random Forest, que mostró un desempeño adecuado en la predicción de casos de dengue en 2017. Además de las variables climáticas, los autores sugirieron la importancia de considerar variables socioeconómicas como el nivel educativo y las condiciones de urbanización para mejorar la precisión del modelo predictivo [35] .

Un artículo adicional revisó la literatura sobre métodos de aprendizaje automático en la detección de la infección por dengue y encontró que, aunque las redes neuronales convolucionales y los perceptrones multicapa fueron predominantes en la detección, los modelos estadísticos clásicos como ARIMA seguían siendo los más empleados para predecir brotes. Este enfoque tradicional en la predicción fue impulsado por un fuerte interés en estudiar y correlacionar las variables climáticas con los casos de dengue a nivel mundial, aunque también se evidenció la necesidad de integrar nuevas variables, como las sociales y demográficas, para obtener modelos más robustos [36] .

En Brasil, un estudio realizado por Mussumeci y Coelho comparó el rendimiento de modelos de pronóstico multivariados a gran escala para la predicción de casos de dengue, utilizando LSTM (Long Short-Term Memory) y regresión de bosques aleatorios (Random Forest). Los autores encontraron que, si bien ambos modelos proporcionaron estimaciones útiles, el LSTM resultó ser superior en la captura de patrones no lineales y temporales en los datos de casos de dengue, lo que permitió una mejor representación de la dinámica temporal de la enfermedad. Las variables utilizadas en el modelo incluyeron factores climáticos, epidemiológicos y demográficos, enfatizando la relevancia de integrar múltiples tipos de datos para mejorar la precisión predictiva. Este estudio refuerza la importancia de emplear modelos avanzados de Machine Learning que incorporen variables multivariadas y capturen relaciones complejas en los datos para mejorar la estabilidad y efectividad de los modelos predictivos [37] .

Finalmente, otro estudio en Costa Rica relacionó variables climáticas y el fenómeno de El Niño-

Oscilación del Sur (ENOS) con los casos de dengue entre 2007 y 2017, señalando la correlación significativa de estas variables con el dengue y recomendando modelos de Machine Learning para una predicción más precisa, aunque las métricas de desempeño resultaron inestables [38].

Estos antecedentes muestran un creciente interés en el uso de aprendizaje automático para predecir y monitorear el dengue, integrando variables climáticas, sociales y demográficas. Sin embargo, persisten desafíos en la estabilidad y robustez de los modelos, lo que sugiere la necesidad de enfoques más integrales que consideren una mayor diversidad de factores que inciden en la propagación del dengue.

4. METODOLOGÍA

La metodología utilizada en este proyecto se basó en un enfoque de recopilación y análisis de datos múltiples, integrando técnicas de minería de datos, análisis predictivo y recolección de información mediante web scraping. A continuación, se describen los principales componentes metodológicos aplicados:

1. **Selección de fuentes de datos:** Se identificaron varias fuentes de datos primarias y secundarias para el análisis. Los datos epidemiológicos del dengue se obtuvieron del Sistema de Vigilancia en Salud Pública (SIVIGILA) y del Instituto Nacional de Salud (INS). Para obtener información sobre las tendencias de búsqueda de dengue, se seleccionó Google Trends como fuente principal de datos web.

2. **Recopilación de datos epidemiológicos y climáticos:** Los datos epidemiológicos abarcaron casos confirmados de dengue en los 42 municipios del Valle del Cauca durante el período comprendido entre enero de 2000 y diciembre de 2019. Estos casos fueron confirmados en laboratorio mediante análisis de hemogramas y pruebas de inmunoglobulina M (IgM), un anticuerpo que indica infecciones recientes. Asimismo, se recopilaron variables climáticas mensuales, como la temperatura media, humedad relativa y precipitación, obtenidas del Servicio de Cambio Climático de Copernicus (C3S).

3. **Factores socioeconómicos:** Los datos socioeconómicos se derivaron del censo de 2018 realizado por el Departamento administrativo nacional de estadística (DANE). Se incluyeron variables como el acceso a servicios básicos (agua y alcantarillado), la proporción de población urbana y la clasificación de los hogares por los diferentes estratos socioeconómicos.

4. **Web scraping:** Se implementaron técnicas de extracción automática de datos desde páginas web mediante scripts desarrollados en Python, utilizando bibliotecas especializadas como BeautifulSoup y Scrapy. Estos scripts se diseñaron para obtener índices de búsqueda del término "dengue" en la región del Valle del Cauca. La recolección de estos datos permitió identificar patrones temporales y regionales de interés.

5. **Limpieza y preprocesamiento de datos:** Tras la extracción de los datos, se procedió a la limpieza y transformación de estos. Este proceso incluyó la eliminación de valores atípicos, el manejo de datos faltantes y la normalización de las variables para asegurar la consistencia y calidad de los datos antes de su integración en el análisis predictivo.

6. **Análisis y modelado predictivo:** Utilizando los datos recopilados, se desarrollaron modelos predictivos basados en técnicas de aprendizaje automático para estimar el número de casos de dengue. Se realizó un análisis exploratorio y se probaron diferentes algoritmos supervisados para seleccionar el modelo con mejor desempeño predictivo.

7. **Visualización de datos:** Finalmente, se utilizaron herramientas de visualización para comunicar los resultados del análisis. Estas visualizaciones permitieron identificar correlaciones y tendencias a lo largo del tiempo, facilitando la comprensión de los factores determinantes en la propagación del dengue.

4.1 ESTRUCTURA Y DESCRIPCIÓN DE BASE DE DATOS

Dando cumplimiento al objetivo número 1 del proyecto *<<Construir la base de datos a través de información histórica de incidencia de dengue, a partir de factores climáticos, sociales y demográficos.>>* se expone la procedencia de la fuente de información para obtener cada uno de los registros utilizados en el proyecto:

- Los casos confirmados y notificados se obtuvieron del sistema de vigilancia en salud pública (SIVIGILA) para cada uno de los 42 municipios del Valle del Cauca entre enero de 2000 y diciembre de 2019. Los datos se descargaron del sitio web del Instituto Nacional de Salud (INS). Todos los casos se confirmaron en laboratorio con pruebas de hemograma y de inmunoglobulina M [IgM]. Para el cálculo de las tasas mensuales, se descargaron proyecciones de población para cada municipio y año del Departamento Administrativo Nacional de Estadística (DANE). Estos datos son de dominio público y se obtuvieron de la siguiente manera:
- Casos de dengue: se descargaron los datos de casos confirmados de Dengue del Instituto Nacional de Salud, de acceso libre [39].
- Variables climáticas: se trabajó con los datos del Copernicus Climate Change Service (C3S, version 1.0) para el periodo de enero de 2000 a diciembre de 2019. Estos datos se basaron en datos horarios del ECMWF ERA5 a nivel de superficie con una resolución espacial de 0.1 [38].
- El Niño Southern Oscillation (ENSO) se analizó usando los indicadores Niño-1.2 and Niño-
- 3.4 del National Oceanic and Atmospheric Administration(NOAA) [40].
- Datos demográficos: proyecciones de población para cada municipio y año del Departamento Administrativo Nacional de Estadística (DANE) [41] .

Unidad de análisis:

- Unidad espacial: Cada uno de los 42 municipios en el valle del cauca
- Unidad temporal: Cada mes en el periodo comprendido entre enero 2000 a diciembre 2019

La recopilación de estos registros permitió conformar una base de datos de 10.080 filas, que corresponden a la información de los 42 municipios del Valle del Cauca en cada uno de los 12 meses durante 20 años, empezando en enero del año 2000 hasta diciembre del año 2019. Las variables del estudio incluyeron tanto categóricas como cuantitativas, agrupadas según sus características.

En total, se obtienen: una variable categórica nominal ('Municipio'), dos variables categóricas ordinales ('Año', 'Mes'), 16 variables cuantitativas continuas ('prec', 'tmax', 'tmin', 'hum', 'pob_urbana', 'acueducto', 'alcan', 'estrato_0', 'estrato_1', 'estrato_2', 'estrato_3', 'estrato_4', 'estrato_5', 'estrato_6', 'soi', 'nino12', 'tpromR', 'tmaxR') y tres variables cuantitativas discretas ('casos', 'Total_General', 'cabecera').

Variable	Descripción	Tipo
Año	Periodo (Año) en que se registra la información	Integer
Mes	Periodo (mes) en que se registra la información	Integer
Municipio	Municipio del Valle del cauca donde se registra la información	String
cod_mun	Código administrativo de municipio según DANE	Float
Prec	Precipitación acumulada mensual de acuerdo con imágenes satelitales	Float
Tmax	Promedio de temperatura máxima del mes de acuerdo con imágenes satelitales	Float
Tmin	Promedio de temperatura mínima del mes de acuerdo con imágenes satelitales.	Float
Hum	Promedio de humedad relativa mensual de aceurdo con imágenes satelitales	Float
mes_num	Mes representado en forma númerica (de 1 a 12)	Float
muni	Etiqueta numérica asignada a cada municipio	Float
casos	Número de casos de dengue presentados en el mes a nivel municipal	Integer
mes_final	Mes representado en forma númerica (de 1 a 12)	Float
DPMP	Código geográfico de municipio	Float
MPIO	Municipio del Valle del cauca donde se registra la información	Float
AÑO	Periodo (Año) en que se registra la información	Integer
Total_General	Número total de habitantes del municipio	Integer
cabecera	Número de habitantes en la cabecera (zona urbana) municipal	Float
pob_urbana	Porcentaje de habitantes en la cabecera (zona urbana) municipal	Float

acueducto	Porcentaje de habitantes con servicio de acueducto	Float
alcan	Porcentaje de habitantes con servicio de alcantarillado	Float
estrato_0	Porcentaje de viviendas de estrato cero en el Municipio	Float
estrato_1	Porcentaje de viviendas de estrato uno en el Municipio	Float
estrato_2	Porcentaje de viviendas de estrato dos en el Municipio	Float
estrato_3	Porcentaje de viviendas de estrato tres en el Municipio	Float
estrato_4	Porcentaje de viviendas de estrato cuatro en el Municipio	Float
estrato_5	Porcentaje de viviendas de estrato cinco en el Municipio	Float
estrato_6	Porcentaje de viviendas de estrato seis en el Municipio	Float
categoria	Etiqueta numérica asignada a cada municipio de acuerdo con la categoría (0-6)	Float
tmaxR	Rango de temperatura máxima en el mes según imágenes satelitales	Float
tprom	Promedio de temperatura del mes de acuerdo con imágenes satelitales	Float
tpromR	Rango de temperatura promedio en el mes según imágenes satelitales	Float
soi	Índice de oscilaciion del sur (índice que mide la Oscilación del Sur al correlacionar valores de presión atmosférica obtenidos en el Pacífico occidental con los del Pacífico central)	Float
nino12	Punto geográfico donde se registra la medición SOI	Float

Tabla 1: Descripción de variables

5. RESULTADOS

5.1 WEB SCRAPPING

Se propuso crear una variable que mostrara la cantidad de veces que las personas en el Valle del Cauca buscaron la palabra "dengue", con el objetivo de determinar si esta variable podía ser relevante en análisis futuros. Para esto, se utilizó el código de la Figura 2 que está diseñado para extraer y analizar datos sobre la frecuencia de búsquedas del término "dengue" en el Valle del Cauca, Colombia, utilizando la biblioteca pytrends, que interactúa con Google Trends. Este análisis cubrió el periodo desde el 1 de enero de 2004 hasta el 20 de mayo de 2024. La decisión de tomar datos desde el 1 de enero de 2004 para analizar la frecuencia de búsquedas del término "dengue" en el Valle del Cauca, utilizando Google Trends y la biblioteca pytrends, se debe a que Google Trends no ofrece datos anteriores a esa fecha.

Google comenzó a recopilar y hacer públicos los datos de tendencias de búsqueda a partir de 2004, que es cuando la herramienta fue lanzada. Por lo tanto, cualquier análisis que utilice Google Trends está limitado a este período inicial en adelante, lo que explica por qué el análisis se enfoca en los datos

desde 2004 hasta la fecha más reciente disponible (20 de mayo de 2024 en este caso).

```python
from pytrends.request import TrendReq

def obtener_busquedas_dengue_en_pradera_ultimo_anio():
    pytrends = TrendReq(hl='es-CO', tz=300)
    pytrends.build_payload(kw_list=['dengue'], geo='CO-VAC', timeframe='2004-01-01 2024-05
    data = pytrends.interest_over_time()

    busquedas_por_mes = data.resample('M').sum()

    print("Búsquedas de 'dengue' en EL VALLE DEL CAUCA por meses en el último año:")
    print(busquedas_por_mes)
    return busquedas_por_mes

if __name__ == "__main__":
    busquedas_por_mes = obtener_busquedas_dengue_en_pradera_ultimo_anio()
```

Figura 2: Código para creación de variable temporal de consultas de la palabra Dengue en el Valle del Cauca

- **IMPORTACIÓN DE LA BIBLIOTECA pytrends**

```python
from pytrends.request import TrendReq
```

Figura 3: Importacion de TrendReq

El código comienza importando TrendReq de la biblioteca Pytrends. Pytrends permite realizar consultas a Google Trends para obtener datos de interés sobre términos específicos.

- **DEFINICION DE LA FUNCION obtener_busquedas_dengue_en_pradera_ultimo_anio**

```python
def obtener_busquedas_dengue_en_pradera_ultimo_anio():
```

Figura 4: Creación de la función

Se definió una función que encapsula el proceso de configuración, obtención y análisis de datos de búsquedas del término "dengue".

- **CREACIÓN DE UNA INSTANCIA DE TrendReq**

```python
pytrends = TrendReq(hl='es-CO', tz=300)
```

Figura 5: Creación de Instancia TrendReq

Se creó una instancia de TrendReq con los parámetros hl='es-CO' (configuración de idioma para español de Colombia) y tz=300 (zona horaria correspondiente a Colombia).

- **CONFIGURACIÓN DEL Payload DE LA CONSULTA**

```
pytrends.build_payload(kw_list=['dengue'], geo='CO-VAC', timeframe='2004-01-01 2024-05-20')
```

Figura 6: Configuración del Payload

Se configuró el payload de la consulta con los siguientes parámetros:

kw_list=['dengue']: Lista de palabras clave a buscar, en este caso, "dengue".

geo='CO-VAC': Código geográfico para limitar la búsqueda al Valle del Cauca, Colombia.

timeframe='2004-01-01 2024-05-20': Intervalo de tiempo para los datos solicitados, desde el 1 de enero de 2004 hasta el 20 de mayo de 2024.

- **OBTENCIÓN DE LOS DATOS DE INTERÉS A LO LARGO DEL TIEMPO**

```
data = pytrends.interest_over_time()
```

Figura 7: Utilizar método interest_over_time

Se obtuvo los datos de interés a lo largo del tiempo utilizando el método interest_over_time. Estos datos contienen la frecuencia de búsqueda del término "dengue" durante el periodo especificado.

- **RE-MUESTREO DE LOS DATOS POR MES**

```
busquedas_por_mes = data.resample('M').sum()
```

Figura 8: Re-muestreo de los datos mes a mes

Los datos obtenidos se re-muestrearon por mes, utilizando el método resample('M') y sumando las búsquedas dentro de cada mes.

Como conclusión, el código presentado utilizó pytrends para extraer datos de Google Trends sobre la frecuencia de búsquedas del término "dengue" en el Valle del Cauca. Al re-muestrear los datos por mes, se proporcionó una visión clara de los patrones de búsqueda a lo largo de los últimos años. Esta información puede ser utilizada para analizar tendencias en el interés público sobre el dengue y apoyar estrategias de salud pública.

5.2 ANÁLISIS EXPLORATORIO DE DATOS

De acuerdo con el segundo objetivo del proyecto: *<<Realizar un análisis exploratorio detallado de los datos para identificar patrones, correlaciones y posibles variables relevantes que influyan en la incidencia de dengue.>>* se presenta un análisis descriptivo de los datos, a nivel univariado y multivariado.Se inició con un análisis de datos faltantes, analizando el porcentaje de registros ausentes por columna, permitiendo evidenciar que la base contiene completitud de datos en todas las variables y registros:

Variables	% Missing Values
Año	0.00000
Mes	0.00000
Municipio	0.00000
prec	0.00000
Tmax	0.00000
Tmin	0.00000
Hum	0.00000
casos	0.00000
Total_General	0.00000
cabecera	0.00000
pob_urbana	0.00000
acueducto	0.00000
alcan	0.00000
estrato_0	0.00000
estrato_1	0.00000
estrato_2	0.00000
estrato_3	0.00000
estrato_4	0.00000
estrato_5	0.00000
estrato_6	0.00000
soi	0.00000
nino12	0.00000
tmaxR	0.00000
tprom	0.00000
tpromR	0.0000

Tabla 2: Porcentaje de valores nulos por variable

A continuación, se analizó el comportamiento cada una de las variables, iniciando por el conteo de registros por municipio, donde se evidenció un total de 240 registros para cada uno, que corresponde a la toma de información en cada uno de los 12 meses durante 20 años. Seguido a esto, se evaluó el comportamiento de las variables climáticas partiendo de las estadísticas descriptivas presentadas en la

Tabla 3.

Variable	prec	tmax	tmin	hum	soi	nino12	tmaxR	tprom2	tpromR
Conteo	1080	1080	1080	1080	1080	1080	1080	1080	1080
Promedio	152,25	25,15	19,70	72,52	-0,05	23,21	3,47	21,54	1,75
Desv. Estandar	93,05	2,00	1,62	6,01	1,05	2,26	0,93	1,74	0,42
Mínimo	10,41	19,55	15,82	61,42	-2,99	19,20	0,83	17,01	0,52
Percentil 25	89,37	23,71	18,60	67,99	-0,80	21,27	2,83	20,38	1,47
Mediana	136,79	25,56	19,78	71,16	-0,17	23,02	3,45	21,73	1,73
Percentil 75	194,02	26,50	20,61	76,35	0,73	25,19	4,08	22,67	2,02
Máximo	954,16	30,31	26,39	87,15	3,02	28,29	6,91	27,41	3,75

Tabla 3: Estadísticas descriptivas para variables climáticas

También se presenta las descriptivas para las variables demográficas en la Tabla 4 y Tabla 5.

Variable	casos	Total_General	cabecera	pob_urbana	acueducto	alcan
Conteo	10080	10080	10080	10080	10080	10080
Promedio	17	102260	87023	0,61	0,90	0,78
Desv. Estándar	110	325121	317564	0,20	0,07	0,14
Mínimo	0	5258	2378	0,18	0,72	0,37
Percentil 25	0	15014	7327	0,45	0,85	0,70
Mediana	1	21913	10623	0,63	0,92	0,78
Percentil 75	6	54557	35463	0,77	0,95	0,89
Máximo	3118	2241491	2190363	0,98	0,99	0,98

Tabla 4: Estadísticas descriptivas para variables sociodemográficas

Variable	estrato_0	estrato_1	estrato_2	estrato_3	estrato_4	estrato_5	estrato_6
Conteo	10080	10080	10080	10080	10080	10080	10080
Promedio	0,00	0,40	0,46	0,11	0,01	0,01	0,00
Desv. Estándar	0,01	0,16	0,13	0,10	0,03	0,02	0,00
Mínimo	0,00	0,12	0,16	0,00	0,00	0,00	0,00
Percentil 25	0,00	0,30	0,36	0,04	0,00	0,00	0,00
Mediana	0,00	0,38	0,47	0,07	0,00	0,00	0,00
Percentil 75	0,00	0,49	0,56	0,16	0,01	0,00	0,00
Máximo	0,03	0,71	0,71	0,39	0,11	0,08	0,03

Tabla 5: Estadísticas descriptivas para variables sociodemográficas

Se acompaña el análisis descriptivo con el boxplot por cada una de las variables cuantitativas presentados en Figura 10: Porcentaje de población urbana por municipio, Figura 11: Comportamiento general para los estratos socioeconómicos, Figura 12: Comportamiento general para las variables población urbana, acueducto y alcantarillado, Figura 13: Porcentaje de población con acceso a acueducto por municipio y Figura 13. En ellas, es evidente que en más del 75% de los municipios, al menos el 50% de la población reside en el área urbana, además a lo menos el 80% y 60% de las localidades tienen acceso a acueducto y alcantarillado, respectivamente. Se aprecia además que la mayoría de los habitantes de los municipios del Valle del cauca están localizados en estratos 1, 2 y 3.

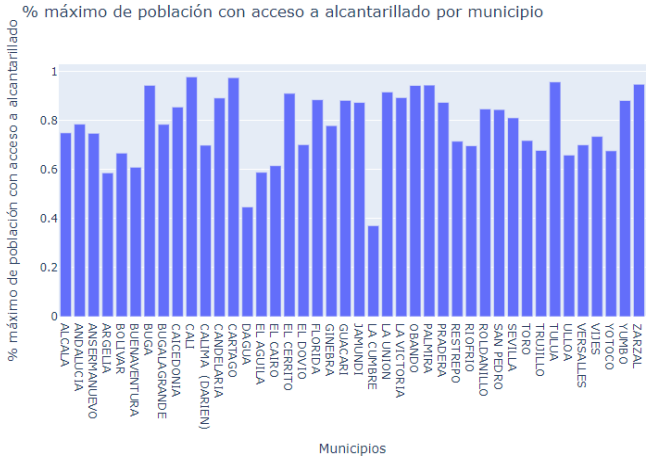

Figura 9: Porcentaje de población con acceso a alcantarillado por municipio

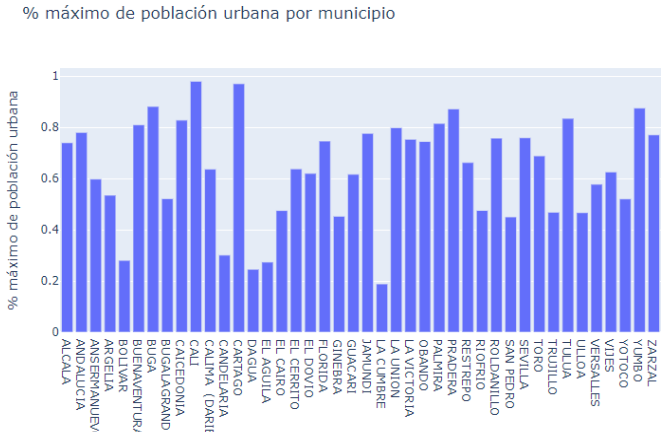

Figura 10: Porcentaje de población urbana por municipio

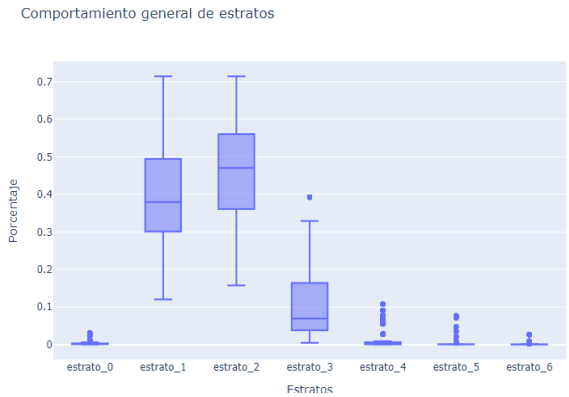

Figura 11: Comportamiento general para los estratos socioeconómicos

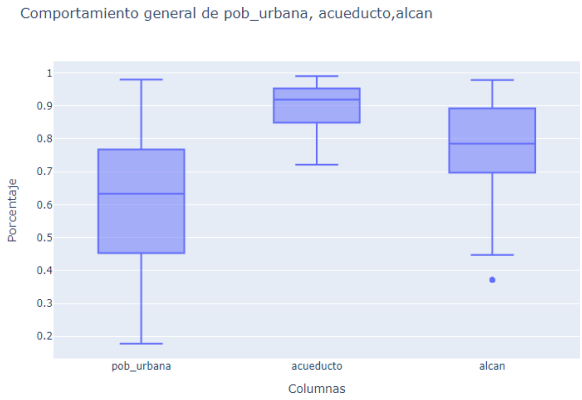

Figura 12: Comportamiento general para las variables población urbana, acueducto y alcantarillado

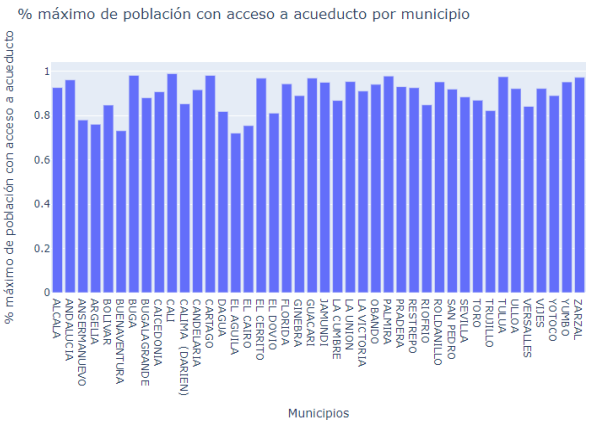

Figura 13: Porcentaje de población con acceso a acueducto por municipio

A continuación, en la Figura 15: Comportamiento espacio-temporal para la precipitación, Figura 16: Comportamiento espacio-temporal para la temperatura promedio y Figura 17: Comportamiento espacio-temporal para la temperatura mínima se expone el comportamiento de la cantidad de casos y las variables climáticas de manera espacio-temporal, lo que permite evidenciar patrones de comportamiento a través de los años observados respecto a cada variable, en cada municipio.

En Cali, la capital del Valle del Cauca es donde más casos de Dengue se reportan año a año, no obstante, es la ciudad con mayor población de todo el Departamento y se evidencian picos significativos entre los años 2008 a 2018. Cabe resaltar que el comportamiento del número de casos de dengue reportado en Cali es significativamente superior frente a los demás municipios del Valle del Cauca.

Figura 14: Comportamiento espacio-temporal para el número de casos de dengue

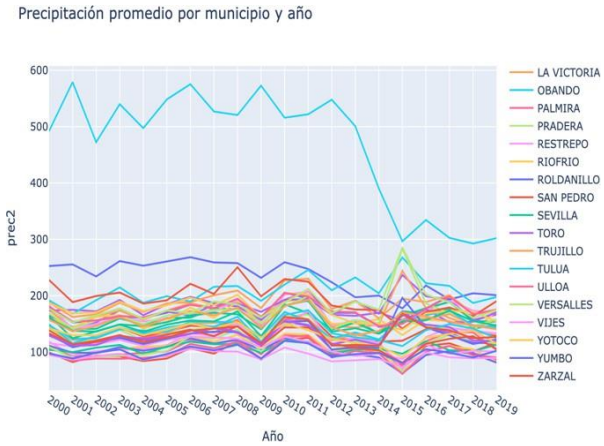

Figura 15: Comportamiento espacio-temporal para la precipitación

Respecto a las variables climáticas, se observa un comportamiento estable a través del tiempo, para cada municipio.

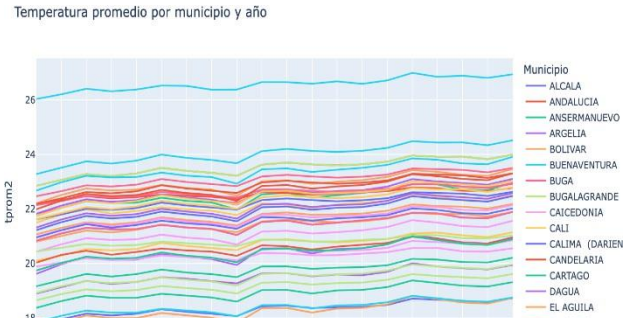

Figura 16: Comportamiento espacio-temporal para la temperatura promedio

La figura 18, muestra la matriz de correlación entre las variables objeto de estudio y permitió concluir que hay relaciones variadas entre las variables climáticas, sociodemográficas e infraestructura con la variable de casos. Aunque algunas correlaciones son fuertes, muchas de las relaciones con la variable de casos son débiles, indicando que factores adicionales pueden estar influyendo en la incidencia de casos de Dengue en los municipios del Valle del Cauca.

Al detallar la relación entre el número de casos de Dengue y las variables climáticas y socioeconómicas, se tuvo como resultado que: frente a la precipitación, existe una débil relación negativa, sugiriendo que mayores niveles de precipitación están ligeramente asociados con un menor número de casos; respecto a la temperatura máxima, hay una correlación de 0.19, indicando que temperaturas máximas más altas están ligeramente asociadas con un aumento de casos; respecto a la temperatura mínima, el resultado es similar a la temperatura máxima, es decir, la temperatura mínima también tiene una correlación importante con el número de casos, adicionalmente, el número de casos presenta una correlación considerable con la temperatura promedio; la relación con la humedad relativa sugiere que los niveles más altos de humedad están ligeramente asociados con una disminución en el número de casos; respecto al acceso a acueducto y alcantarillado, hay una relación positiva moderada, sugiriendo que

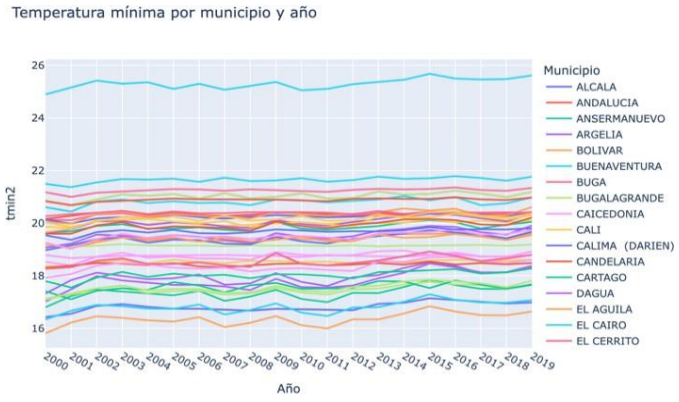

Figura 17: Comportamiento espacio-temporal para la temperatura mínima

una mayor cobertura de acueducto está asociada con un mayor número de casos, posiblemente reflejando una mejor detección en áreas con mejor infraestructura

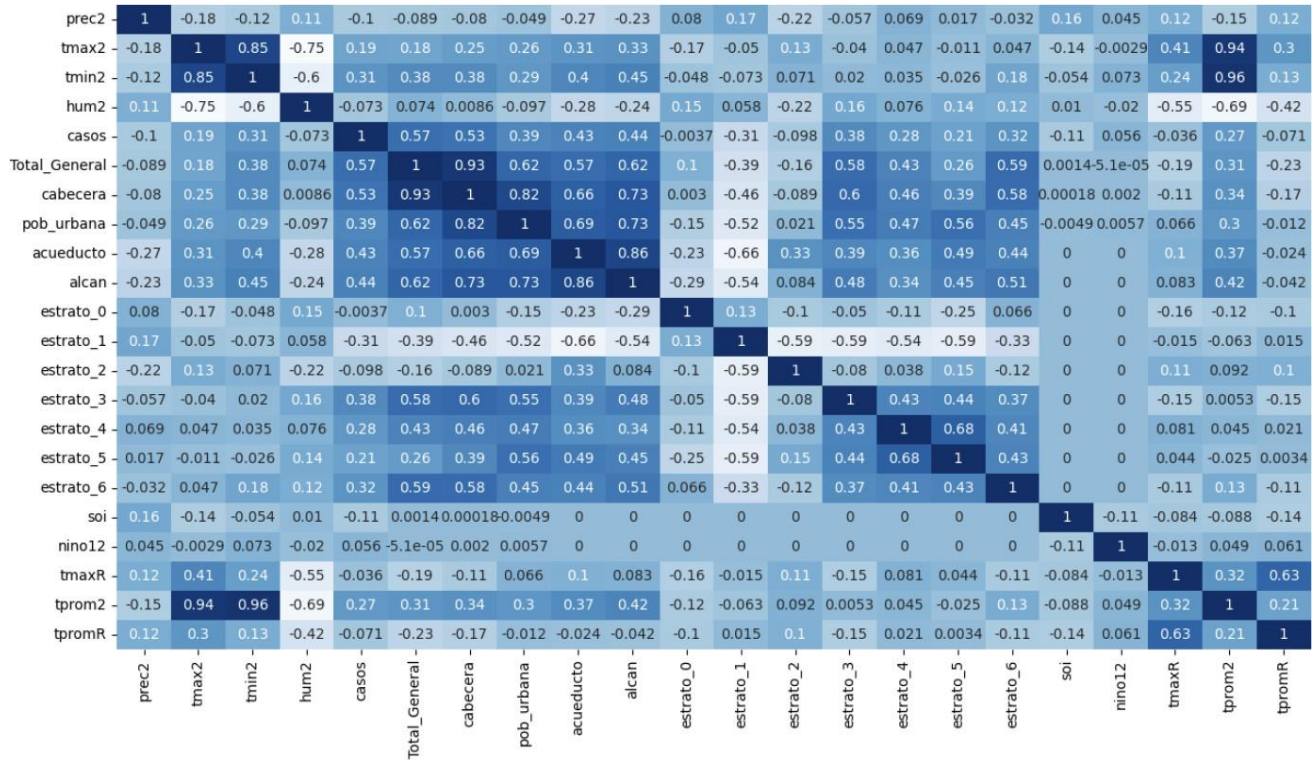

Figura 18: Análisis de correlación de variables

5.3 ANALISIS GEOGRÁFICO PARA LAS VARIABLES CLIMÁTICAS Y DEMOGRÁFICAS

Inicialmente, se presenta la ubicación geográfica de cada Municipio en el Valle del Cauca. Se puede identificar que al norte, se ubican: El Águila, Alcalá, Ansermanuevo, Argelia, El Cairo, La Unión, La Victoria, Ulloa, Bolívar, Cartago, El Dovio, Obando, Roldanillo, Toro, Versalles y Zarzal; en el centro están: Andalucía, Bugalagrande, El Cerrito, Ginebra, Guacarí, Buga, Riofrío, Trujillo, Calima – El Darién, Restrepo, San Pedro, Tuluá y Yotoco; al occidente se encuentra Buenaventura; al oriente Caicedonia y Sevilla; y al sur: Cali, Candelaria, Palmira, Dagua, Florida, Jamundí, La Cumbre, Pradera, Vijes y Yumbo.

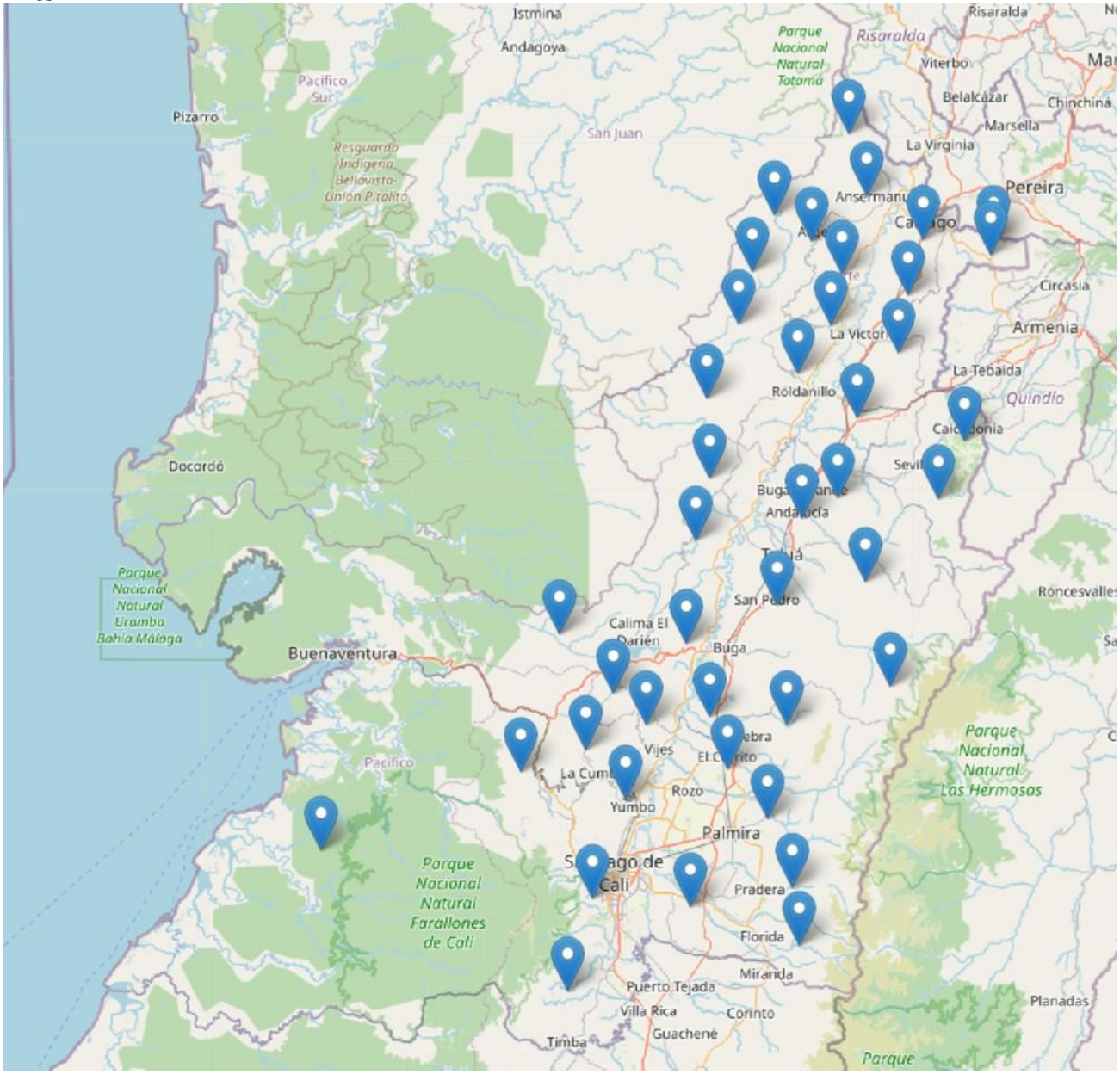

Figura 19: Ubicación geográfica de cada municipio sobre el mapa del Valle del Cauca

En la Figura 20: Población y promedio anual de casos de Dengue para cada municipio sobre el mapa geográfico del Valle del Cauca, se muestra que las áreas con mayor densidad están marcadas en colores más oscuros, indicando una mayor cantidad de personas viviendo en estas áreas, lo que permite concluir que la mayor concentración de población se está ubicada alrededor de Santiago de Cali y sus alrededores. Este patrón se mantiene en todos los años observados en este estudio. En cuanto al promedio de casos registrados en cada municipio, se evidencia un patrón consistente en la ciudad más densamente poblada y este se mantiene constante en el tiempo y lugar. Lo anterior, permite confirmar una consistencia en la distribución espacial a lo largo de los diferentes periodos para las dos variables analizadas; además, Santiago de Cali y sus alrededores se destacan como las áreas con mayor densidad de población y mayor

incidencia de dengue, lo que sugiere que las áreas urbanas densamente pobladas tienen un mayor riesgo de brotes de dengue por sus condiciones poblacionales. Para analizar este comportamiento geográfico, se decide agrupar en periodos de 5 años para efectos gráficos.

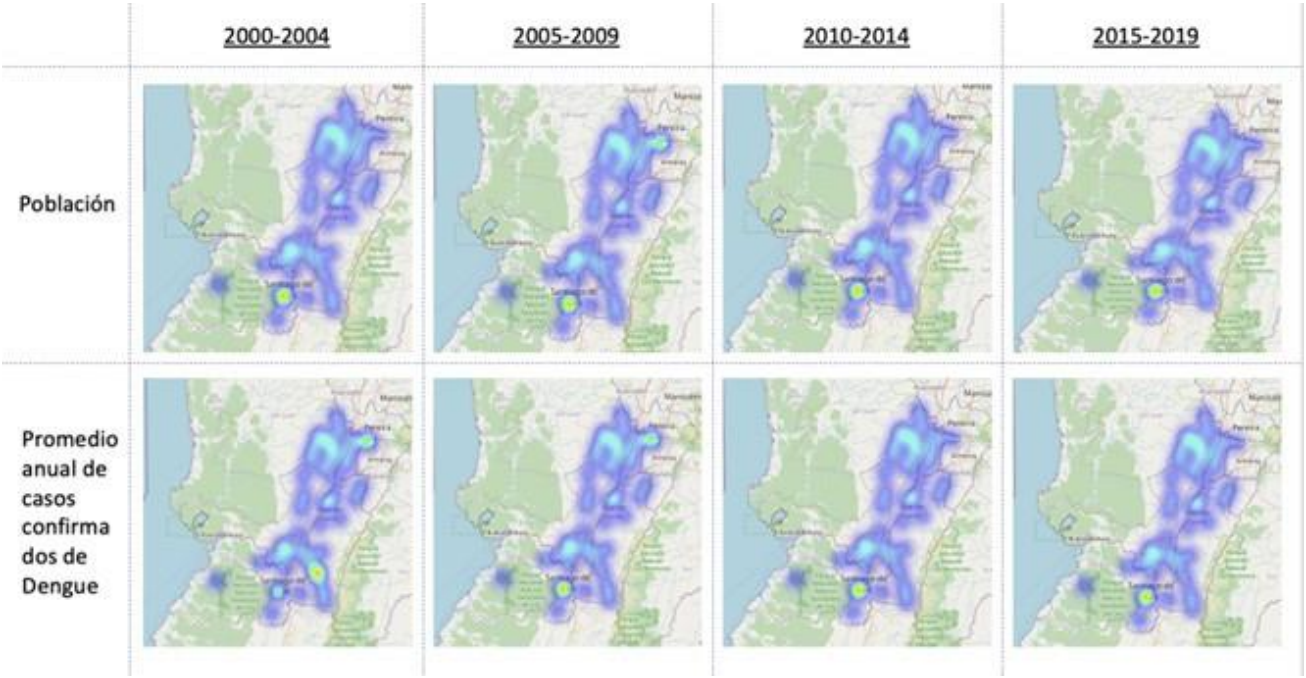

Figura 20: Población y promedio anual de casos de Dengue para cada municipio sobre el mapa geográfico del Valle del Cauca durante cada periodo de 5 años.

En la Figura 21 se observa que la temperatura anual promedio es similar para los municipios del Valle del Cauca, siendo especialmente mayor en los municipios del norte del Departamento, sin embargo, en el último periodo analizado, se evidencia un incremento en las temperaturas en todo el Valle. Esto mismo ocurre con la humedad relativa, es decir, en los municipios del norte del Valle, son superiores los niveles de humedad relativa, además hay un ascenso notorio en el primer y último periodo.

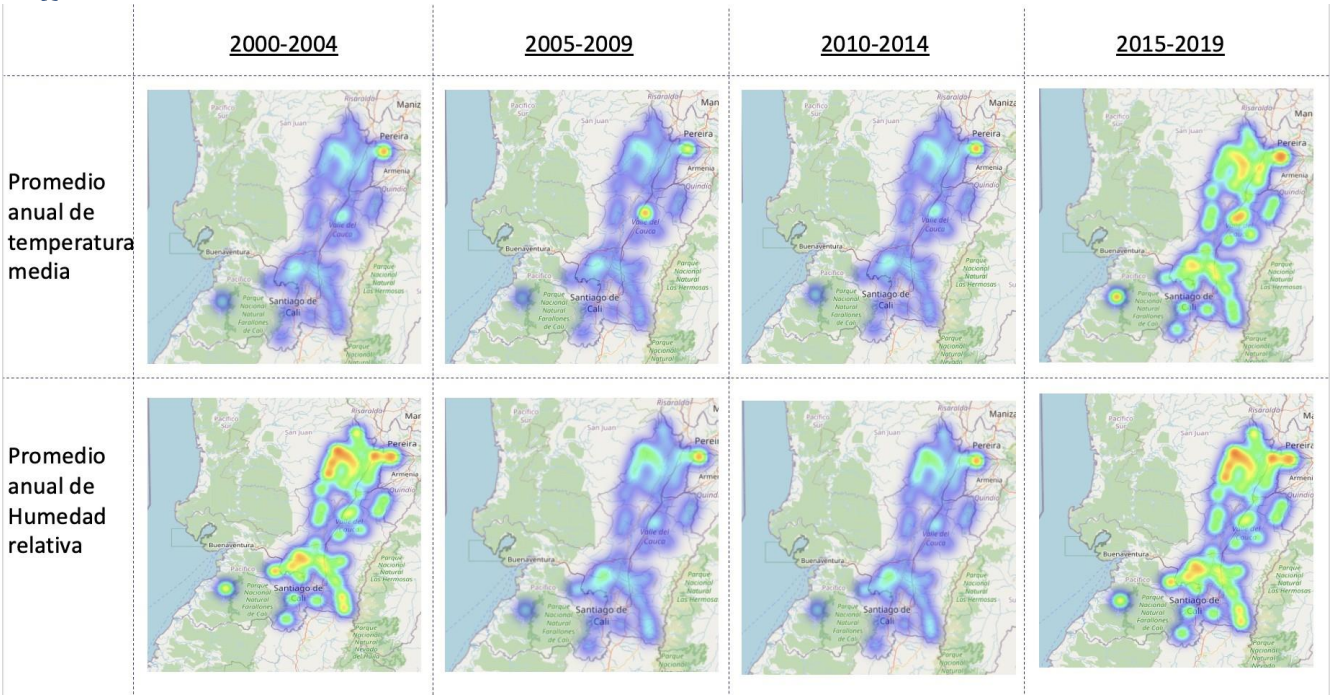

Figura 21: Temperatura promedio y Humedad Relativa para cada municipio sobre el mapa geográfico del Valle del Cauca

Por último, la Figura 22 permite apreciar que la zona con mayor frecuencia y niveles de precipitación es el occidente vallecaucano, específicamente en Buenaventura, que destaca por ser un municipio costero, con un clima cálido y tropical que favorece el incremento de la precipitación.

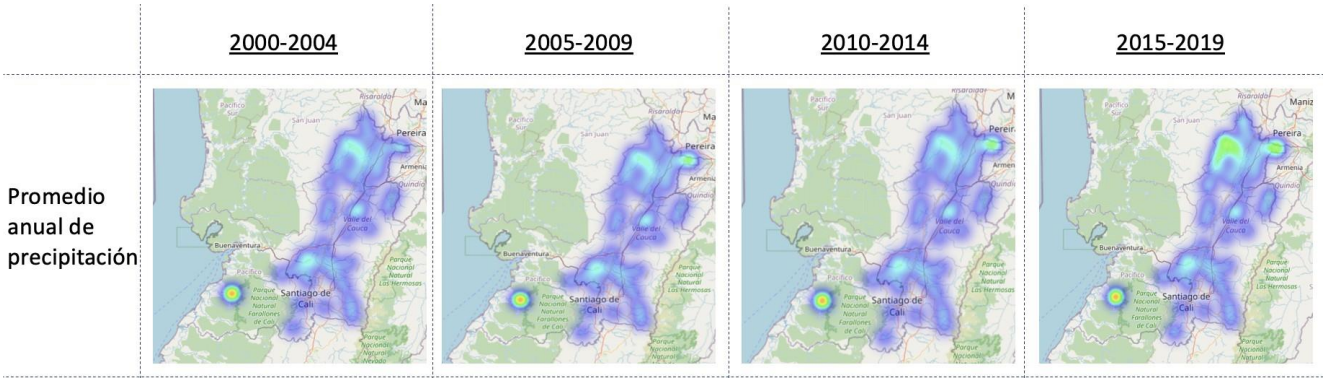

Figura 22: Precipitación para cada municipio sobre el mapa geográfico del Valle del Cauca

6. CONSTRUCCIÓN Y EVALUACIÓN DE MODELOS

6.1 SELECCIÓN DE VARIABLES

Según la información presentada en la Figura 18, se llevó a cabo una selección preliminar, conservando aquellas que muestran alguna relación con la variable "casos". Esto asegura la inclusión de variables explicativas que contribuyen a la predicción del número de casos reportados en los municipios del Valle del Cauca. Como resultado, se seleccionaron las variables 'prec2', 'tmin2', 'Total_General', 'acueducto', 'alcan' y 'estrato_3', las cuales presentan una correlación superior a 0.27.

6.2 SELECCIÓN DE MODELOS

Se tuvo en cuenta múltiples algoritmos de regresión para evaluar su rendimiento, considerando tanto modelos lineales como no lineales:

- **Regresión Lineal y variantes regularizadas:** Ridge y Lasso.
- **Árboles de decisión**: Decision Tree, Random Forest, y Gradient Boosting.
- **Modelos más complejos:** Support Vector Regressor (SVR) y el K-Nearest Neighbors (KNN).

Cada modelo fue entrenado utilizando el 70% de los datos escalados para entrenamiento y evaluado con el 30% restante como conjunto de prueba. La métrica utilizada para medir el desempeño fue el Mean Squared Error (MSE), que cuantifica el error cuadrático medio entre los valores reales y los predichos; un MSE más bajo indica un mejor ajuste del modelo.

Los resultados obtenidos en el entrenamiento y evaluación de los diferentes modelos de regresión permitieron comparar su desempeño mediante el MSE. A continuación, se presenta una interpretación de los valores obtenidos:

1. **Regresión Lineal (MSE: 0.1354):** Aunque tiene un MSE bajo, este modelo asume una relación lineal entre las variables predictoras y la variable objetivo. Esto puede ser una simplificación excesiva, ya que la propagación del dengue está influenciada por factores complejos y no necesariamente lineales. Sin embargo, su desempeño relativamente bueno sugiere que, para algunas regiones o condiciones específicas, la relación lineal podría reflejar adecuadamente la dinámica de los contagios.

   Los parámetros escogidos para este modelo son: {'copy_X': True, 'fit_intercept': True, 'n_jobs': None, 'positive': False}

2. **Ridge Regression (MSE: 0.0801):** Similar a la regresión lineal estándar, pero con una penalización aplicada a los coeficientes para evitar el sobreajuste. El rendimiento es prácticamente igual al de la regresión lineal, lo que indica que el modelo no está sufriendo de sobreajuste y que la regularización no mejora significativamente el rendimiento en este caso.

Los parámetros escogidos para este modelo son:
Hiperparámetros utilizados:
{'alpha': 1.0,
'copy_X': True,
'fit_intercept': True,
'max_iter': None,
'positive': False,
'random_state': None,
'solver': 'auto',
'tol': 0.0001}

3. **Lasso Regression (MSE: 0.1417):** Este modelo también aplica una penalización, pero a diferencia de Ridge, Lasso tiende a reducir algunos coeficientes a cero, realizando una selección automática de variables. El mayor error sugiere que esta simplificación adicional podría estar eliminando variables importantes para predecir la tasa de contagio de dengue.

   Los parámetros escogidos para este modelo son:
   {'alpha': 1.0,
   'copy_X': True,
   'fit_intercept': True,
   'max_iter': 1000,
   'positive': False,
   'precompute': False,
   'random_state': None,
   'selection': 'cyclic',
   'tol': 0.0001,
   'warm_start': False}

4. **Decision Tree (MSE: 0.1652):** Este modelo, aunque intuitivo y fácil de interpretar, tiende a sobreajustarse cuando no se limita su profundidad. El alto error cuadrático medio indica que el modelo está capturando demasiado ruido de los datos, lo que disminuye su capacidad para generalizar a nuevos casos.

   Los parámetros escogidos para este modelo son:
   {'ccp_alpha': 0.0,
   'criterion': 'squared_error',
   'max_depth': None,
   'max_features': None,
   'max_leaf_nodes': None,
   'min_impurity_decrease': 0.0,
   'min_samples_leaf': 1,
   'min_samples_split': 2,

'min_weight_fraction_leaf': 0.0,
'monotonic_cst': None,
'random_state': None,
'splitter': 'best'}

5. **Random Forest (MSE: 0.1220):** Este modelo combina múltiples árboles de decisión, lo que reduce la varianza y mejora la generalización. Su MSE ligeramente superior sugiere que, aunque tiene un rendimiento excelente, podría no captar con el mismo detalle algunos

   Los parámetros escogidos para este modelo son:

   {'bootstrap': True,
   'ccp_alpha': 0.0,
   'criterion': 'squared_error',
   'max_depth': None,
   'max_features': 1.0,
   'max_leaf_nodes': None,
   'max_samples': None,
   'min_impurity_decrease': 0.0,
   'min_samples_leaf': 1,
   'min_samples_split': 2,
   'min_weight_fraction_leaf': 0.0,
   'monotonic_cst': None,
   'n_estimators': 100,
   'n_jobs': None,
   'oob_score': False,
   'random_state': None,
   'verbose': 0,
   'warm_start': False}

6. **Gradient Boosting (MSE: 0.1213):** Este modelo presentó el menor error cuadrático medio, lo que indica que es el más eficiente evaluado para predecir la tasa de contagio de dengue. Gradient Boosting crea una secuencia de árboles de decisión, donde cada uno corrige los errores del anterior, y es conocido por su capacidad para manejar relaciones no lineales y complejas en los datos. El bajo MSE indica que este modelo es capaz de captar mejor las características subyacentes que afectan la propagación del dengue en los municipios analizados.

   Los parámetros escogidos para este modelo son:
   {'alpha': 0.9,
   'ccp_alpha': 0.0,

'criterion': 'friedman_mse',
'init': None,
'learning_rate': 0.1,
'loss': 'squared_error',
'max_depth': 3,
'max_features': None,
'max_leaf_nodes': None,
'min_impurity_decrease': 0.0,
'min_samples_leaf': 1,
'min_samples_split': 2,
'min_weight_fraction_leaf': 0.0,
'n_estimators': 100,
'n_iter_no_change': None,
'random_state': None,
'subsample': 1.0,
'tol': 0.0001,
'validation_fraction': 0.1,
'verbose': 0,
'warm_start': False}

7. **SVR (MSE: 0.1349**): El modelo de Soporte Vectorial para regresión (SVR) también mostró un buen rendimiento, aunque ligeramente menos preciso que los modelos basados en árboles. Este modelo es útil para datos con relaciones complejas, pero su mayor error indica que podría no estar capturando todas las interacciones entre las variables predictoras de manera tan eficiente como Gradient Boosting o Random Forest.

   Los parámetros escogidos para este modelo son:
   {'C': 1.0,
   'cache_size': 200,
   'coef0': 0.0,
   'degree': 3,
   'epsilon': 0.1,
   'gamma': 'scale',
   'kernel': 'rbf',
   'max_iter': -1,
   'shrinking': True,
   'tol': 0.001,
   'verbose': False}

8. **KNN (MSE: 0.1510)**: El modelo K-Nearest Neighbors tiene un error considerablemente mayor. Este modelo no hace supuestos sobre la forma de la relación entre las variables, pero parece que no está capturando de manera efectiva las relaciones espaciales y temporales complejas

que influyen en la propagación del dengue.

Los parámetros escogidos para este modelo son:
{'algorithm': 'auto',
'leaf_size': 30,
'metric': 'minkowski',
'metric_params': None,
'n_jobs': None,
'n_neighbors': 5,
'p': 2,
'weights': 'uniform'}

En resumen, los mejores resultados se obtuvieron con Gradient Boosting y Random Forest, gracias a su capacidad para manejar relaciones no lineales y complejas. Sin embargo, modelos más simples como la Regresión Lineal o Ridge también mostraron un rendimiento decente, aunque están limitados por la sencillez de sus supuestos.

Es importante destacar que los modelos de regresión empleados, como la regresión lineal, Ridge, Lasso y los árboles de decisión, no consideraron explícitamente la estructura de los datos tipo panel. Este tipo de datos se caracteriza por observar múltiples entidades (en este caso, municipios) a lo largo del tiempo, lo que introduce dependencias temporales y espaciales entre las observaciones. Los modelos tradicionales de regresión asumen que las observaciones son independientes entre sí, lo cual puede llevar a una subestimación o sobreestimación del error si no se considera la estructura inherente de los datos.

Por ejemplo, la tasa de contagio en un municipio en un periodo específico podría estar influenciada no solo por las variables climáticas y sociodemográficas de ese momento, sino también por eventos ocurridos en periodos anteriores (dependencia temporal) o por lo que sucede en municipios cercanos (dependencia espacial). Modelos más adecuados para este tipo de datos serían los modelos de regresión de datos de panel, que permiten manejar estas dependencias al incluir efectos fijos o aleatorios que capturan la heterogeneidad entre las entidades (municipios) y las interacciones a lo largo del tiempo.

Además, existen enfoques avanzados, como modelos de series de tiempo o modelos espacio-temporales, que podrían proporcionar una mejor representación de la dinámica del dengue al considerar relaciones temporales y geográficas. Estos modelos son capaces de manejar la correlación temporal entre observaciones sucesivas y la posible influencia espacial entre municipios vecinos, aspectos cruciales en la propagación de enfermedades contagiosas como el dengue.

Asimismo, al evaluar estos modelos de manera convencional, una métrica común es el coeficiente de determinación, o R-squared (R^2), que mide la proporción de la varianza de la variable dependiente explicada por las variables independientes en un modelo de regresión. Un valor de R^2 cercano a 1 indica que el modelo captura gran parte de la variabilidad en los datos, mientras que un valor cercano a 0

sugiere que el modelo explica poco de esta variabilidad. Sin embargo, al trabajar con datos de panel o modelos espacio-temporales, el R^2 debe interpretarse con precaución, pues estos modelos pueden incluir efectos de dependencia que no siempre son reflejados en esta métrica tradicional.

En los resultados obtenidos:

- **Regresión Lineal:** R^2 de 0.043, lo que indica que este modelo lineal solo explica alrededor del 4.3% de la variación en los datos. Aunque es un valor positivo, sugiere que el modelo no es muy efectivo para capturar la relación entre las variables predictoras y la tasa de contagio de dengue. Esto puede deberse a que la relación entre las variables no es estrictamente lineal.

- **Ridge Regression:** Presentó un R^2 de 0.0431, lo que indica que también explica aproximadamente el 4.3% de la variación. Su rendimiento es casi idéntico al de la regresión lineal, lo que refuerza la idea de que un modelo lineal, incluso con penalización, no es adecuado para describir con precisión la complejidad de este conjunto de datos.

- **Lasso Regression:** Obtuvo un R^2 negativo (-0.0009), lo que significa que el modelo tiene un rendimiento inferior al de una línea horizontal que predice el valor promedio de la variable dependiente. Esto indica que Lasso no es adecuado para estos datos.

- **Decision Tree:** Presentó un R^2 de -0.167, lo que indica que este modelo no solo falló en capturar la relación entre las variables, sino que sus predicciones fueron significativamente peores que una predicción trivial basada en el promedio de la tasa de contagio de dengue. Un valor negativo de R^2 sugiere que el modelo sobre ajustó los datos de entrenamiento, pero no generalizó bien en los datos de prueba.

- **Random Forest:** Obtuvo el mayor R^2 con un valor de 0.152, lo que sugiere que este modelo es el que mejor captura la relación entre las variables predictoras y la tasa de contagio de dengue. Aunque este valor no es extremadamente alto, refleja que el modelo explica el 15.2% de la variación en los datos, lo cual es aceptable dado que se está trabajando con datos complejos y no lineales.

- **Gradient Boosting:** Presentó un R^2 de 0.143, muy cercano al de Random Forest, lo que confirma que este también es un buen modelo para explicar la variabilidad en la tasa de contagio.

- **SVR:** Obtuvo un R^2 de 0.0472, lo que indica que aproximadamente el 4.72% de la variación en la tasa de contagio de dengue es explicada por las variables independientes en el modelo. Aunque el valor es positivo, sugiere que SVR no captura efectivamente la relación entre las variables en este conjunto de datos, lo que podría indicar que la relación entre las variables no es lineal o que hay factores adicionales no considerados que podrían influir en la tasa de contagio.

- **KNN (K-Nearest Neighbors):** También obtuvo un valor de R^2 negativo, -0.066, lo que indica que este modelo no captura adecuadamente la relación entre las variables y tiene un rendimiento inferior al promedio.

En conclusión, se evidenció que los modelos Random Forest y Gradient Boosting no solo obtuvieron el mejor MSE, sino que también mostraron R^2 relativamente más altos en comparación con otros modelos, lo que confirma su capacidad para captar las relaciones complejas y no lineales en los datos. Estos resultados resaltan la importancia de considerar el contexto y la estructura de los datos al seleccionar modelos, así como la necesidad de explorar enfoques más sofisticados para abordar problemas relacionados con datos de panel.

6.3 MODELO LSTM PARA DATOS TIPO PANEL

El modelo LSTM está diseñado para manejar dependencias en los datos secuenciales, lo que permite capturar dependencias temporales y mantener la estructura espacial en bases de datos longitudinales o tipo panel, como es el caso de la base original. Teniendo en cuenta este contexto, se realiza el entrenamiento de una red neuronal recurrente LSTM.

En la base de datos el factor temporal se representa con la variable 'Periodo' y el factor espacial con la variable 'Municipio' como se presenta a continuación:

	Municipio	prec2	tmin2	Total_General	acueducto	alcan	estrato_3	casos	Periodo
176	BUGA	134.89339	18.80076	123824	0.98192	0.94366	0.16852	14	200001
140	BUGA	110.17971	18.75231	123824	0.98192	0.94366	0.16852	10	200002
324	BUGA	147.59850	18.91008	123824	0.98192	0.94366	0.16852	18	200003
38	BUGA	171.14922	19.13162	123824	0.98192	0.94366	0.16852	27	200004
359	BUGA	173.04630	19.18594	123824	0.98192	0.94366	0.16852	16	200005

Figura 23: Estructura de la base de datos

La variable de respuesta corresponde al logaritmo de la tasa de contagio por cada mil habitantes, para cada municipio y para cada periodo, de tal modo que:

$$Tasa\_contagio = \frac{Casos}{Total_{General}} * 1000$$

$$\log\_Tasa\_contagio = \log_{10}[\text{"}Tasa\_contagio\text{"}]$$

Para este modelo, se usaron como variable de respuesta “log_Tasa_contagio” y variables predictoras 'prec2', 'tmin2', 'acueducto', 'alcan', 'estrato_3'. Debido a que la escala de las variables es diferente, inicialmente se escalaron los datos para mantener la comparabilidad. Por la estructura temporal de la base, se generan las secuencias de aprendizaje, manteniendo 12 meses de secuencia durante el proceso.

Parámetros:

```
        model = Sequential([
  LSTM(units=100, return_sequences=True, input_shape=(X_train.shape[1], X_train.shape[2]),
kernel_regularizer='l2'),
  Dropout(0.5),  # Aumento de dropout
  BatchNormalization(),

  LSTM(units=100, return_sequences=True, kernel_regularizer='l2'),
  Dropout(0.5),
  BatchNormalization(),

  LSTM(units=50, return_sequences=False, kernel_regularizer='l2'),
  Dropout(0.5),

  Dense(units=100, activation='relu'),  # Más neuronas en la capa densa
  Dropout(0.5),
  Dense(1)
])
```

Parámetros para compilación:

```
optimizer = Adam(learning_rate=0.0005)
model.compile(optimizer=optimizer, loss='mean_squared_error', metrics=['mae'])
early_stopping = EarlyStopping(monitor='val_loss', patience=20, restore_best_weights=True)
```

Resultados:

	R^2	MAE	Loss promedio
Train	0.2439	0.0514	0.0062
Test	0.2184	0.0478	0.0058

Tabla 6: Resultados modelo LSTM para datos tipo panel

Las métricas de desempeño del modelo de predicción resultaron bajas, lo que puede atribuirse a la presencia significativa de ceros en los datos. Esta alta proporción de cercanos a cero dificulta que el modelo capture patrones consistentes, afectando su capacidad predictiva. La variabilidad en los datos y la predominancia de ceros reduce la precisión y el ajuste del modelo. Gráficamente, se observa que

el modelo no capta adecuadamente el comportamiento de la tasa de contagio, ya que los valores predichos son, generalmente, inferiores a los reales. Se destaca que el modelo no presenta sobreajuste, ya que las métricas de entrenamiento se mantienen estables en el conjunto de prueba. Las técnicas de regularización L2, Dropout y Early Stopping contribuyen a evitar el sobreajuste en el modelo.

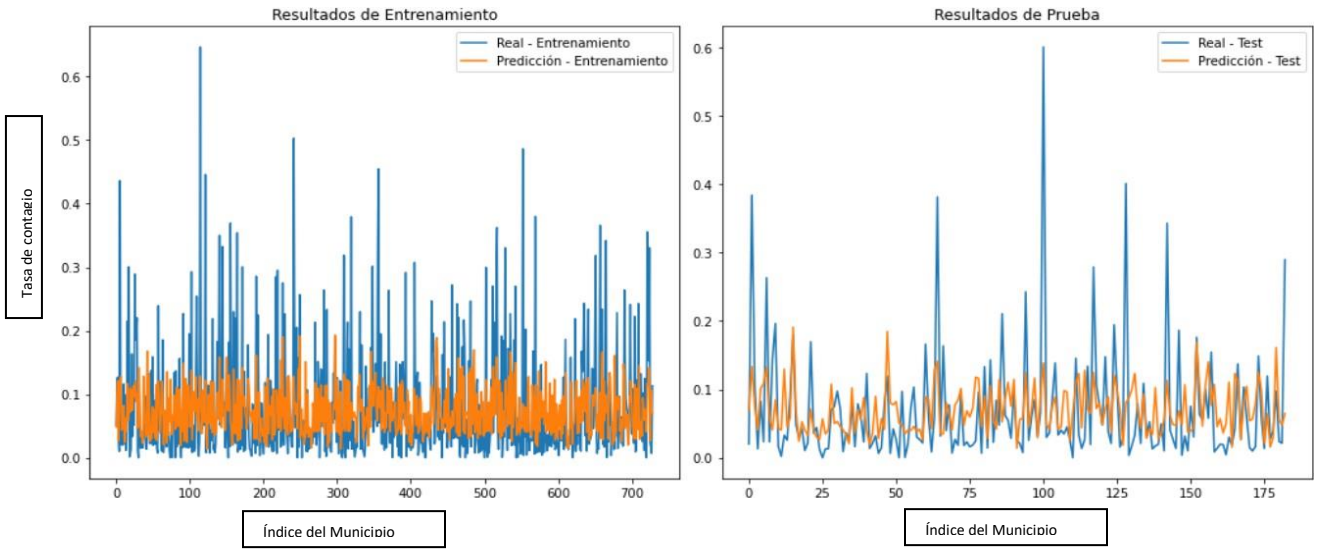

Figura 24: Comportamiento del modelo LSTM para la predicción de la tasa de contagios

Con base en el análisis de la base de datos, se observó que el comportamiento de los casos reportados por municipio en el Valle del Cauca presenta una heterogeneidad significativa. Esta variabilidad implica que algunos municipios tienen dinámicas particulares en la incidencia de casos, posiblemente influenciadas por factores demográficos, socioeconómicos, climáticos u otros aspectos contextuales.
Por lo tanto, se decidió ajustar el enfoque del modelo. En lugar de intentar predecir el número de casos considerando todos los municipios como un único conjunto homogéneo, se optó por agrupar los municipios según patrones similares en la cantidad de casos reportados. Esta agrupación permite abordar las diferencias intrínsecas en el comportamiento de los casos entre municipios. Una vez definidos estos grupos, se construyen modelos específicos para cada grupo.

Además, dado que la incidencia de los casos está influenciada por factores temporales (como variaciones estacionales o tendencias anuales), se incorpora el factor tiempo en el modelo. Este enfoque combinado busca mejorar la precisión del modelo al reconocer y capturar tanto las diferencias espaciales (entre municipios o clústeres) como las temporales, reflejando de manera más adecuada la realidad observada en los datos.

6.4 MODELO LSTM PARA DATOS TIPO PANEL CON DIVISIÓN DEL CONJUNTO DE DATOS POR CONGLOMERADOS PARA MUNICIPIOS

Se tuvo en cuenta dos factores fundamentales para dividir el conjunto de datos: inicialmente, realiza

un análisis de conglomerados a través de un K-Means usando como información principalmente el número de casos y los municipios, para determinar las diferencias estadísticas importantes y cuantos grupos se deben tener en cuenta. Adicionalmente, se investigó sobre el nivel de complejidad de los servicios prestados en la red de salud que presentan los municipios en el Valle del Cauca.

Análisis de conglomerados

Se realiza un análisis de conglomerados K-Means para identificar los grupos de municipios con comportamiento diferencial en el número de casos reportados, para esto, se define a priori utilizar dos clústeres obteniendo como resultado que el primer clúster está conformado por los municipios de Cali, Buga, Tuluá, Cartago y Palmira, y el segundo grupo por los restantes 37 municipios del Valle del Cauca.

Se observa que estos 5 municipios tienen un mayor número de casos reportados versus los demás municipios del Departamento.

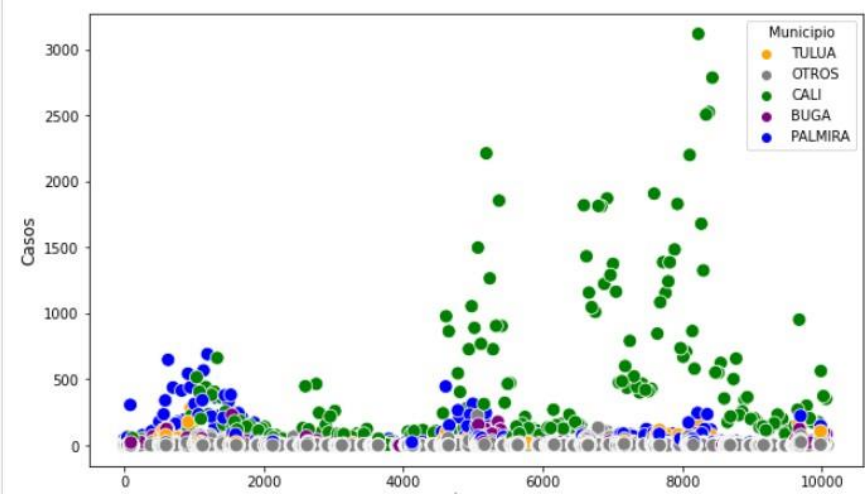

Figura 25: Casos reportados en municipios con hospitales Nivel 3 y municipios sin hospitales Nivel 3

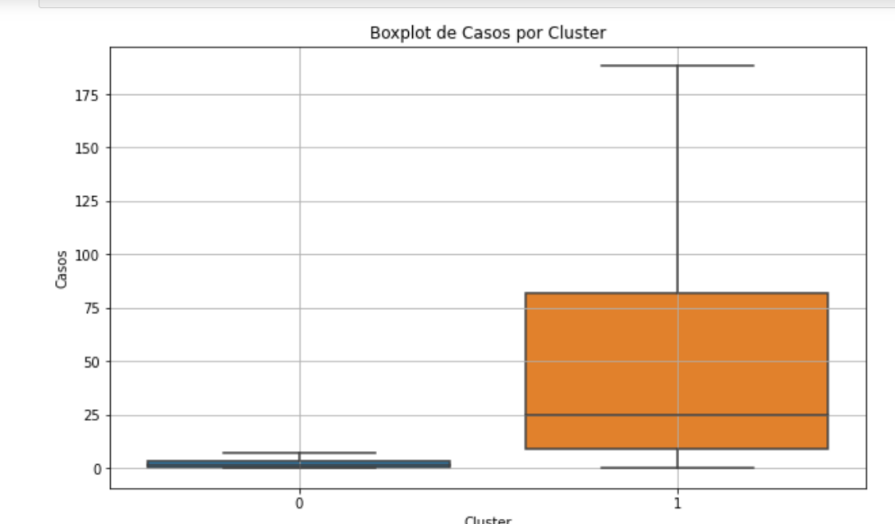

Figura 26: Boxplot para número de casos por Clúster sin observar datos atípicos

Estructura servicios de salud de los municipios en el Valle del Cauca Niveles de complejidad en los hospitales públicos y privados.

En Colombia, el sistema de salud clasifica los hospitales en niveles de complejidad de acuerdo con los servicios que ofrecen y los recursos tecnológicos y humanos disponibles. Estos niveles se dividen principalmente en tres categorías [1] [42] :

- **Primer nivel de atención:**
  Los hospitales de primer nivel ofrecen servicios básicos de atención primaria. Estos incluyen consultas médicas generales, servicios de enfermería, vacunación, odontología básica, control prenatal, programas de promoción y prevención, y tratamiento de enfermedades comunes. Están orientados a la atención de la mayoría de las necesidades de salud de la población, sin la necesidad de tecnologías o especialistas avanzados.

- **Segundo nivel de atención:**
  Los hospitales de segundo nivel tienen mayor complejidad que los de primer nivel. Ofrecen atención especializada en áreas como pediatría, ginecología, medicina interna y cirugía general. También cuentan con laboratorios más avanzados, servicios de imagenología y hospitalización. Estos centros se encargan de la atención de problemas de salud que no pueden ser resueltos en los hospitales de primer nivel.

- **Tercer nivel de atención:**
  Los hospitales de tercer nivel son los más complejos dentro del sistema de salud colombiano. Ofrecen atención altamente especializada en áreas como cardiología, oncología, neurocirugía y unidades de cuidados intensivos (UCI). Estos hospitales están equipados con tecnología de punta y profesionales especializados, lo que les permite tratar condiciones de salud críticas y realizar intervenciones quirúrgicas complejas.

De acuerdo con la Red Colombiana contra el Ataque Cerebral (RecaVar) [2] [43], en el Valle del Cauca hay únicamente 13 hospitales de tercer nivel, de los cuales 10 se encuentran en Cali, uno en Guadalajara de Buga, uno en Palmira y uno en Tuluá. Estos centros hospitalarios atienden casos de salud complejos, incluidos los derivados de complicaciones por dengue, que provienen de otros municipios del departamento e incluso de otras regiones del suroccidente colombiano. Como resultado, muchos de los casos remitidos se registran directamente en estos cuatro municipios principales.

En la base de datos, el comportamiento de estos 4 municipios es diferencial respecto al número de casos de dengue reportados, como se observa a continuación:

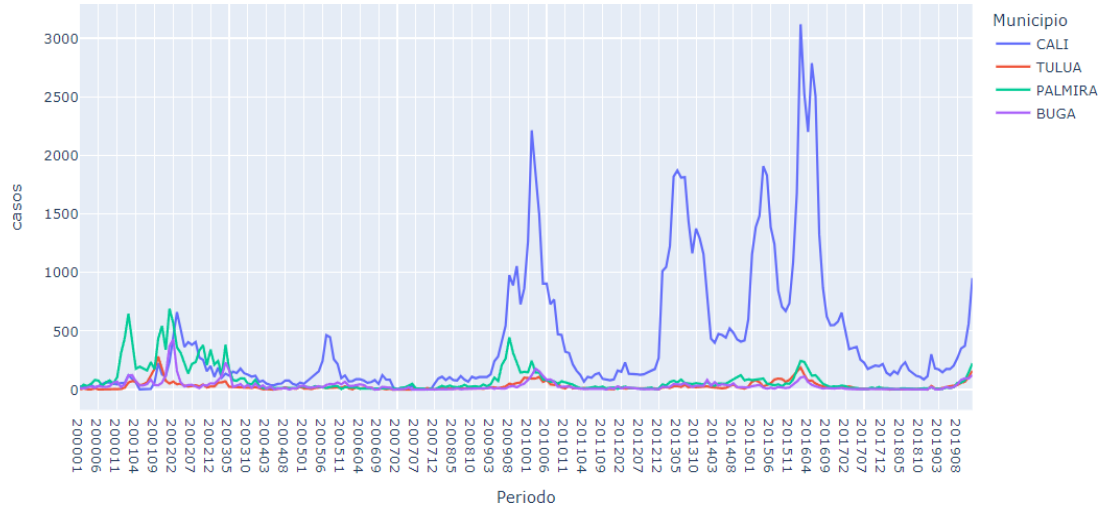

Figura 27: Comportamiento del número de casos para municipios con hospitales de Nivel 3

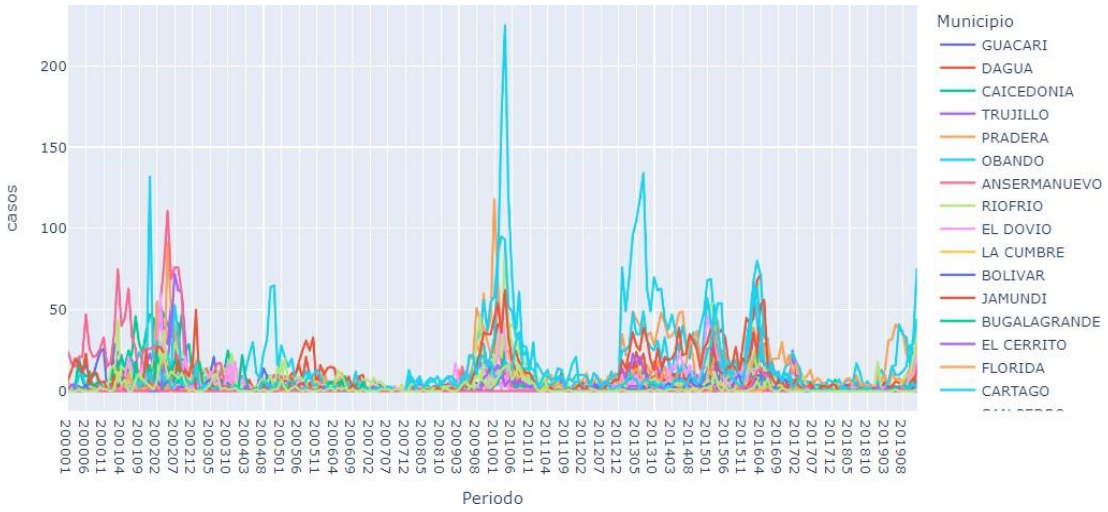

Figura 28: Comportamiento del número de casos para municipios sin hospitales de Nivel 3

Debido a la diferencia en el comportamiento de los casos reportados de dengue, los clúster encontrados en el análisis de conglomerados y considerando la dinámica de atención en los centros de salud y hospitales, se decide dividir la base de datos original en dos conjuntos. El primer conjunto se obtendrá aplicando un filtro por municipio, conservando únicamente los casos de "CALI", "PALMIRA", "BUGA" y "TULUÁ". El segundo conjunto incluirá los datos de los demás municipios, excluyendo estos cuatro del filtrado.

6.5 MODELO LSTM PARA LOS CUATRO MUNICIPIOS CON HOSPITALES CON NIVEL DE ATENCIÓN 3

Para este modelo, se usaron como variable de respuesta “casos” y variables predictoras 'prec2', 'tmin2', 'Total_General', 'acueducto', 'alcan', 'estrato_3'. Debido a que la escala de las variables es diferente, inicialmente se escalaron los datos para mantener la comparabilidad. Por la estructura temporal de la base, se generan las secuencias de aprendizaje, manteniendo 10 meses de secuencia durante el proceso.

Parámetros:

```
model = Sequential([
LSTM(units=100, return_sequences=True, input_shape=(X_train.shape[1], X_train.shape[2]), kernel_regularizer='l2'),
Dropout(0.2), # Aumento de dropout
BatchNormalization(),
LSTM(units=100, return_sequences=True, kernel_regularizer='l2'),
Dropout(0.2),
BatchNormalization(),
LSTM(units=50, return_sequences=False, kernel_regularizer='l2'),
Dropout(0.2),
Dense(units=100, activation='relu'), # Más neuronas en la capa densa
Dropout(0.2),
Dense(1)
])
```

Parámetros para compilación:

```
optimizer = Adam(learning_rate=0.0005)
model.compile(optimizer=optimizer, loss='mean_squared_error', metrics=['mae'])
early_stopping = EarlyStopping(monitor='val_loss', patience=20, restore_best_weights=True)
```

Resultados:

	R^2	MAE	Loss promedio
Train	0.8788	0.1949	0.2352
Test	0.8227	0.2260	0.2694

Tabla 7: Resultados para modelo LSTM para Municipios con hospitales Nivel de atención 3

Los resultados del ajuste del modelo muestran métricas aceptables tanto en el conjunto de entrenamiento como en el de prueba. En relación con el R^2, se establece que el 88% de la variabilidad de los valores reales puede ser explicada por las predicciones del modelo. Además, el error promedio y la tasa de pérdida son bajos, lo que respalda el buen ajuste del modelo.

Gráficamente, se observa que el modelo capta adecuadamente el comportamiento del número de

casos reportados en los municipios con hospitales de nivel 3. También se destaca que el modelo no presenta sobreajuste, ya que las métricas de entrenamiento se mantienen estables en el conjunto de prueba. Las técnicas de regularización L2, Dropout y Early Stopping contribuyen a evitar el sobreajuste en el modelo.

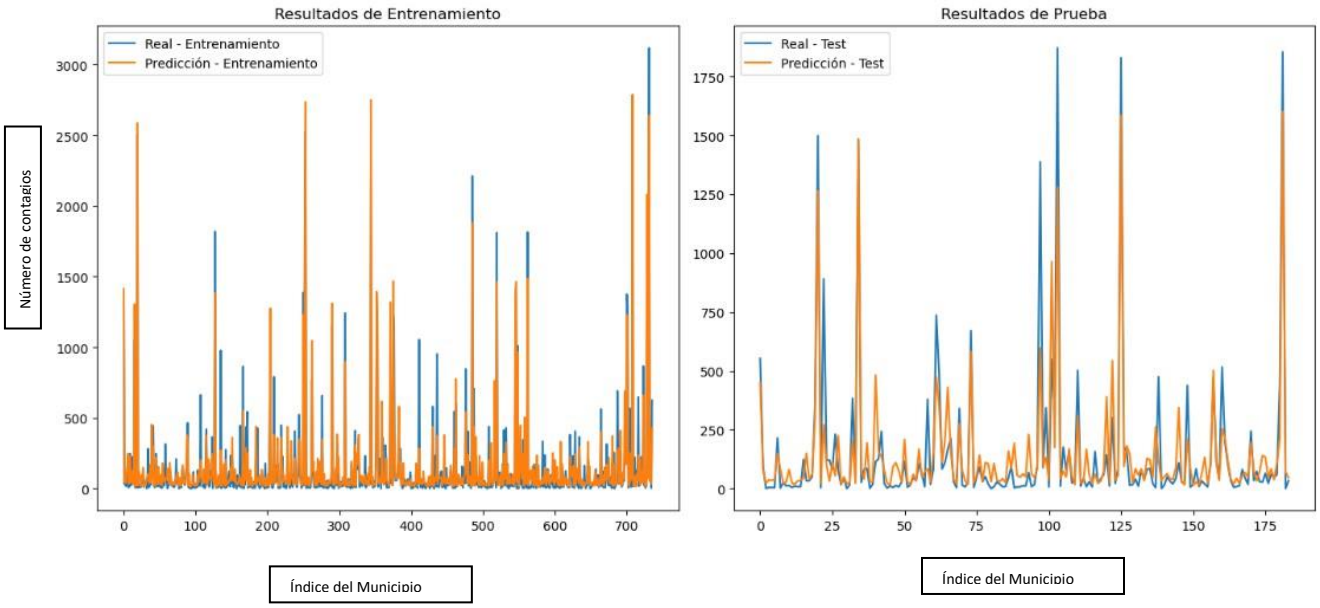

Figura 29: Comparación entre valores reales y predichos para train y test

Cálculo de la tasa

Este modelo nos brinda información sobre el número de contagios por municipio para cada periodo en el Valle del Cauca, partiendo de información climática y sociodemográfica; no obstante, nuestro objetivo plantea el análisis para la tasa de contagios, por lo que podemos utilizar las predicciones del número de contagios y dividirla por el número de habitantes en el municipio, tal y como se calculó la tasa anteriormente. De tal modo qué:

$$\boldsymbol{Tasa\_contagio\_estimada} = \frac{\boldsymbol{Casos}}{\boldsymbol{Tota\_General}} * \mathbf{1000}$$

6.6 MODELO LSTM PARA LOS MUNICIPIOS SIN HOSPITALES CON NIVEL DE ATENCIÓN 3

Para este modelo, se usaron como variable de respuesta “casos” y variables predictoras 'prec2', 'tmin2', 'Total_General', 'acueducto', 'alcan', 'estrato_3' y se excluyeron los municipios “CALI” “BUGA”, “TULUA” y “PALMIRA”. Por la estructura temporal de la base, se generan las secuencias de aprendizaje, manteniendo 10 meses de secuencia durante el proceso.

Parámetros:

```
model = Sequential([
LSTM(units=100, return_sequences=True, input_shape=(X_train.shape[1], X_train.shape[2]), kernel_regularizer='l2'),
   Dropout(0.2),
   BatchNormalization(),

   LSTM(units=100, return_sequences=True, kernel_regularizer='l2'),
   Dropout(0.2),
   BatchNormalization(),

   LSTM(units=50, return_sequences=False, kernel_regularizer='l2'),
   Dropout(0.2),

   Dense(units=100, activation='relu'),  # Más neuronas en la capa densa
   Dropout(0.2),
   Dense(1)
])
```

Parámetros para compilación:

```
optimizer = Adam(learning_rate=0.0005)
model.compile(optimizer=optimizer, loss='mean_squared_error', metrics=['mae'])
early_stopping = EarlyStopping(monitor='val_loss', patience=20, restore_best_weights=True)
```

Resultados:

	R^2	MAE	Loss promedio
Train	0.4576	0.3688	0.6136
Test	0.4610	0.3959	0.6810

Tabla 8: Resultados para modelo LSTM para hospitales sin nivel de atención 3

Los resultados del ajuste del modelo muestran métricas poco aceptables tanto en el conjunto de

entrenamiento como en el de prueba. Este modelo presenta gran cantidad de ceros en la variable de respuesta, es decir, hay municipio en los que, por periodos de tiempo prolongados, no reportan casos de dengue; esto conlleva a que el modelo pierda eficiencia en la predicción y se refleja en las métricas. Gráficamente, se observa que las predicciones del modelo tienen un comportamiento inferior versus lo valores reales.

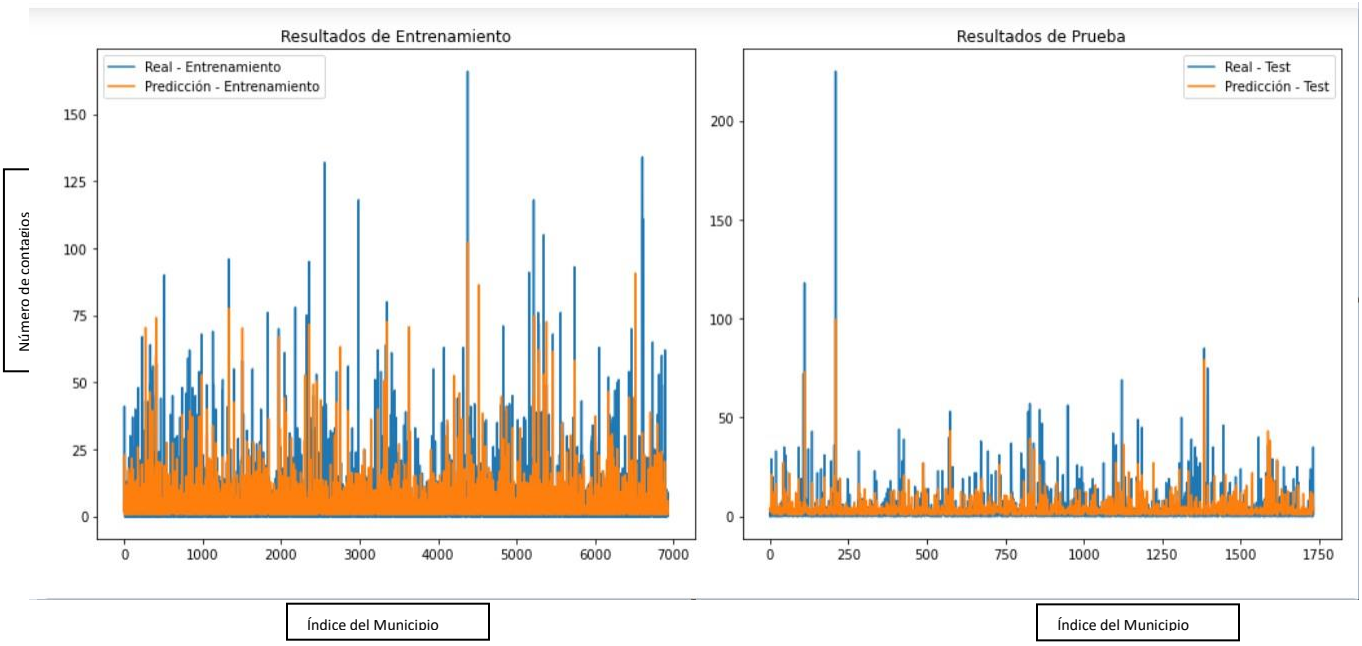

Figura 30: Comparación entre valores reales y predichos para train y test para modelo de otros municipios

6.7 INTERFAZ GRÁFICA PARA USO DEL MODELO EN EL CONTEXTO REAL

Se ha desarrollado una interfaz gráfica de usuario (GUI) para facilitar la predicción del número de casos en los cuatro municipios con hospitales nivel 3 en el Valle del Cauca, utilizando el modelo de red neuronal previamente entrenado. Esta interfaz fue implementada en Python usando la biblioteca tkinter, y la lógica de predicción se basa en un modelo de predicción de series temporales. A continuación, se describen los principales componentes de la implementación:

Para cargar el modelo de predicción, se utiliza la función load_model de TensorFlow para cargar un modelo previamente entrenado y guardado en el archivo Mejor_modelo_Cali_87.h5. Este modelo fue entrenado para predecir el número de casos con base en seis variables de entrada: prec2, tmin2, acueducto, alcantarillado, Total_General y el estrato_3.

```
# Cargar el modelo previamente entrenado
model = load_model('Mejor_modelo_Cali_87.h5')
```

Figura 31: Código para cargue de modelo

Para que los datos de entrada coincidan con los utilizados en el entrenamiento del modelo, se utilizan los escaladores StandardScaler de la biblioteca scikit-learn. Se ajustan estos escaladores con los datos originales cargados desde el archivo datos.csv. Esto asegura que los valores de entrada y salida se normalicen correctamente antes de ser procesados por el modelo.

La Interfaz Gráfica de Usuario (GUI), se crea con tkinter, donde se definen etiquetas y campos de entrada para que el usuario introduzca los valores de las seis variables. Cada uno de estos valores es recibido y almacenado para su posterior uso en el proceso de predicción.

Una vez que el usuario introduce los valores, estos se procesan y se escalan para ajustarse a los datos de entrenamiento. La función de predicción toma estos valores escalados, los pasa al modelo de red neuronal y obtiene una predicción escalada. Posteriormente, el resultado es desescalado para obtener el valor real y se muestra al usuario en un cuadro de mensaje.

La GUI tiene un botón que, al ser presionado, ejecuta la predicción y muestra el resultado mediante un cuadro de diálogo emergente. En caso de error, se muestra un mensaje de advertencia al usuario.

```
# Botón para hacer la predicciónS
predict_button = tk.Button(root, text="Hacer Predicción", command=hacer_prediccion)
predict_button.pack()
```

Figura 32: Código de creación de botón en la GUI para ejecutar el modelo

A continuación, se presenta una demostración visual de la interfaz gráfica de usuario (GUI) que fue desarrollada para realizar predicciones de casos utilizando un modelo de red neuronal previamente entrenado. Esta interfaz permite que el usuario ingrese los valores correspondientes a las variables de entrada y obtenga el número estimado de casos.

La interfaz fue diseñada para ser simple y funcional, ofreciendo una experiencia de usuario intuitiva para quienes deseen realizar predicciones sin necesidad de interactuar directamente con el código subyacente.

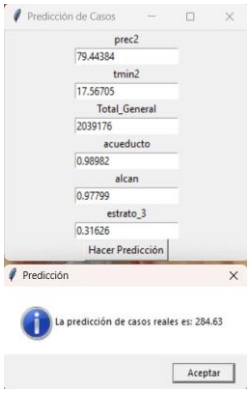

Figura 33: Visualización de interfaz gráfica funcional

La interfaz gráfica incluye campos de texto para ingresar valores específicos de las variables relevantes (prec2, tmin2, acueducto, alcantarillado, Total_General y el estrato_3) y un botón que permite generar la predicción de casos. Los resultados son mostrados en una ventana emergente al realizar la predicción.

7. CONCLUSIONES Y TRABAJOS FUTUROS

7.1 CONCLUSIONES

La incorporación de 22 variables climáticas, sociales y demográficas provenientes de diversas fuentes como SIVIGILA, Copernicus Climate Change Service y el DANE, está en línea con los estudios que enfatizan la importancia de integrar factores externos para predecir casos de dengue. Tal como se menciona en el estado del arte, modelos recientes han mostrado que la inclusión de variables climáticas (temperatura, precipitaciones, humedad) mejora significativamente la precisión predictiva, como lo demuestra el uso de Random Forest en Costa Rica y LSTM en Brasil.

La decisión de dividir los municipios en grupos de acuerdo con sus características refleja una estrategia para abordar la heterogeneidad observada en el comportamiento de los casos. Esta estrategia tiene antecedentes en estudios que también segmentaron los datos por regiones o características específicas para mejorar la precisión de los modelos, como el análisis espacio-temporal de los casos en departamentos colombianos y el enfoque basado en clústeres sugerido en la revisión sistemática de métodos predictivos.

En el análisis exploratorio de los datos se evidencia que existe consistencia y constancia en el comportamiento de las variables climáticas en la mayoría de los municipios del Valle del Cauca, además, es clara la diferencia en número de habitantes y número de casos reportados entre municipios. Por otra parte, en la variable casos existe gran cantidad de ceros, demostrando que en los municipios pequeños no se reportan tantos casos, lo que plantea la hipótesis que en estos municipios no se tiene una vigilancia rigurosa de la enfermedad debido a las limitaciones administrativas que se puedan presentar o se reportan en la ciudad capital, Cali, debido la complejidad de los servicios de salud.

Las variables *'prec', 'tmin2', 'Total_General', 'acueducto', 'alcan' y 'estrato_3'* presentan una correlación superior a 0.27 con la variable “casos” y fueron usadas para el entrenamiento del modelo, permitiendo recopilar información relevante del comportamiento de esta enfermedad.

Se evaluaron distintos modelos de regresión, haciendo principal énfasis en las redes neuronales secuenciales LSTM que permiten mantener el comportamiento de datos tipo panel que es original de la base de datos. Fue necesaria la división del conjunto de datos de acuerdo con los resultados obtenidos en la clusterización y a la configuración de la infraestructura de salud en el Valle del Cauca, permitiendo obtener un modelo con métricas estables y aceptables para el conglomerado de

municipios comprendidos por “CALI”, “BUGA” “TULUA” y “PALMIRA”. Para este modelo se creó una interfaz gráfica que permita obtener la predicción del modelo con los datos que el usuario le suministre.

Para los demás municipios, se evidencia la presencia de varios meses en los que no se presentan casos reportados de dengue, esto produce a que el comportamiento de esta enfermedad sea más difícil de capturar por un modelo de regresión. A esto se le conoce como "inflación de ceros" y conlleva que los modelos de regresión tradicionales o modelos como el LSTM, no capturen adecuadamente la naturaleza discreta y sesgada de los datos. La inflación de ceros puede distorsionar los resultados al subestimar o sobreestimar la relación entre las variables predictoras y la respuesta. Este fenómeno se presentó tanto para el número de casos como para la tasa de contagios por municipio en el Valle del Cauca.

La implementación de una interfaz gráfica que permite al usuario realizar predicciones personalizadas refleja un enfoque orientado a la aplicación práctica de los resultados, algo que no siempre se enfatiza en la literatura, pero que aumenta la utilidad del modelo en contextos reales. Esto se alinea con el interés creciente en crear herramientas accesibles para la gestión de brotes, como se sugiere en estudios recientes.

7.2 TRABAJOS FUTUROS

Para futuros trabajos, se propone incluir más características climáticas y socioeconómicas que permitan predecir de una mejor manera los casos de dengue en cada uno de los municipios del Valle del Cauca, manteniendo la dependencia temporal y espacial.

Se propone evaluar otros modelos que permitan comprender la estructura de la base de datos y la cantidad importante de ceros por la naturaleza propia del caso de uso, lo que permite predecir el número de casos y tasas de contagio en municipios con poca población y limitaciones en la red de salud.

Se plantea evaluar la inclusión del web scraping en los modelos a nivel de municipio, ya que actualmente solo es posible detectar las búsquedas a nivel departamental.

8. REFERENCIAS BIBLIOGRÁFICAS

[1] J. S. C. Camargo, *Análisis epidemiológico en la reactivación del COVID-19 año 2021 mediante modelos de machine learning*. Bogotá, 2023.

[2] J. A. C. Gallego, *Influencia de variables sociales, económicas y espaciales en enfermedades transmitidas por vectores usando algoritmos y técnicas de machine learning*. Pereira, 2018.

[3] G. L. E. Maquen-Niño and otros, "Una revisión sistemática de modelos de clasificación de dengue utilizando machine learning," 2023.

[4] U. de Vigilancia de la Salud, *Condiciones del cambio climático que favorecen al aumento del dengue*. Tegucigalpa, Honduras, 2019.

[5] O. M. de la Salud, "Dengue."

[6] M. de Salud y Protección Social, "Guía de manejo clínico del dengue en Colombia," 2023, *Bogotá, Colombia*.

[7] D. J. Gubler, "Dengue and dengue hemorrhagic fever," *Clin Microbiol Rev*, 1998.

[8] O. M. de la Salud, "Dengue and severe dengue," 2019.

[9] P. U. C. de Chile, "Epidemia, pandemia o endemia."

[10] M. G. Guzmán and E. Harris, "Dengue," *The Lancet*, vol. 385, no. 9966, pp. 453–465, 2015, doi: 10.1016/S0140-6736(14)60572-9.

[11] B. Schölkopf and A. Smola, *Learning with Kernels*. MIT Press, 2001.

[12] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput*, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.

[13] Y. Feng, J. Min, X. Yang, C. Wang, L. Zhang, and Y. Liu, "Deep learning-based framework for dengue outbreaks prediction," *IEEE Access*, vol. 8, pp. 92985–92995, 2020, doi: 10.1109/ACCESS.2020.2993727.

[14] A. Montgomery, "Introduction to Linear Models," in *Introduction to Linear Models*, Cambridge University Press, 2015, pp. 1–50.

[15] G. J. et al., *An Introduction to Statistical Learning*. Springer, 2021.

[16] G. Casella and R. Berger, *Statistical Inference*. Duxbury, 2002.

[17] T. H. et al., *The Elements of Statistical Learning*. Springer, 2009.

[18] L. B. et al., *Classification and Regression Trees*. Wadsworth, 1984.

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

[20] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann Stat*, vol. 29, no. 5, pp. 1189–1232, 2001, doi: 10.1214/aos/1013203451.

[21] R. Caruana and A. Niculescu-Mizil, "An Empirical Comparison of Supervised Learning Algorithms," in *Proceedings of the 23rd International Conference on Machine Learning*, 2006.

[22] 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.

[23] V. N. Vapnik, *The nature of statistical learning theory*. Springer, 1995. doi: 10.1007/978-1-4757-2440-0.

[24] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *Am Stat*, vol. 46, no. 3, pp. 175–185, 1992, doi: 10.1080/00031305.1992.10475879.

[25] B. Baltagi, *Econometric Analysis of Panel Data*, 5th ed. Wiley, 2013.

[26] H. Hsiao, *Analysis of Panel Data*. Cambridge University Press, 2003.

[27] J. Wooldridge, *Introductory Econometrics: A Modern Approach*, 6th ed. South-Western College, 2016.

[28] P. Arellano, *Panel Data Econometrics*. Oxford University Press, 2003.
[29] A. Cameron and P. Trivedi, *Microeconometrics: Methods and Applications*. Cambridge University Press, 2005.
[30] K. Clark and T. Linzer, “Should I use fixed or random effects?,” *Political Sci Res Methods*, vol. 3, no. 2, pp. 399–408, 2015.
[31] N. R. Draper and H. Smith, *Applied Regression Analysis*, 3rd ed. Wiley, 1998.
[32] J. Rodríguez and C. Correa, “Predicción temporal de la epidemia de dengue en Colombia: Dinámica probabilista de la epidemia,” 2009.
[33] J. O. Roldán-Vargas and others, “Probabilistic spatial-temporal prediction of total and severe epidemic of dengue in Colombia,” *Revista de Salud Pública*, vol. 20, pp. 352–358, 2018.
[34] G. L. E. Maquen-Niño and others, “A systematic review of dengue classification models using machine learning,” *RISTI - Revista Ibérica de Sistemas e Tecnologias de Informação*, pp. 5–27, 2023.
[35] V. B. P. A, “Uso del aprendizaje automatizado y de variables climáticas como herramienta para la predicción del riesgo de dengue en Costa Rica,” 2020.
[36] E. Muñoz, G. Poveda, M. P. Arbeláez, and I. D. Vélez, “Spatiotemporal dynamics of dengue in Colombia in relation to the combined effects of local climate and ENSO,” *Acta Trop*, vol. 224, p. 106136, 2021.
[37] E. Mussumeci and F. C. Coelho, “Large-scale multivariate forecasting models for Dengue: LSTM versus random forest regression,” *Spat Spatiotemporal Epidemiol*, vol. 35, p. 100372, 2020, doi: 10.1016/j.sste.2020.100372.
[38] H. Boogaard, J. Schubert, A. De Wit, J. Lazebnik, R. Hutjes, and G. der Grijn, “Agrometeorological indicators from 1979 to present derived from reanalysis,” 2020.
[39] I. N. de Salud, “Casos confirmados de Dengue.”
[40] D. A. N. de Estadística, “Censo nacional de población y vivienda, Colombia 2018,” 2018.
[41] D. A. N. de Estadística (DANE), “Metodología de la estratificación socioeconómica urbana para servicios públicos domiciliarios,” 2015, *Santa Fe de Bogotá, Colombia*.
[42] M. de Salud de Colombia, “Normatividad sobre infraestructura física hospitalaria.”
[43] R. C. contra el Ataque Cerebral - RecaVar, “Hospitales de tercer nivel en Colombia.”