Análisis Descriptivo y Predictivo para la Vigilancia de los Casos de Dengue Grave en la Ciudad de Cali

Andrés Mauricio Mena Ríos, Faber Esteban Hurtado Murillo, Jefferson Sánchez Andrade

Nota de Aceptación

Certificamos que el presente Trabajo de Grado Satisface, en alcances y calidad, todos los requisitos que demanda un Trabajo de Grado de Maestría.

David Arango Londoño

DAVID ARANGO LONDOÑO
Director

HERNÁN DARÍO VARGAS CARDONA
Jurado

H. FABIAN TOBAR TOSSE
Jurado

Aprobado en cumplimiento de los requisitos exigidos por la Pontificia Universidad Javeriana Cali, para optar el título de Magister en Ciencia de Datos.

Camilo Rocha

HERNAN CAMILO ROCHA NIÑO Ph. D.
Decano Facultad de Ingeniería y Ciencias

JUAN CARLOS MARTÍNEZ ARIAS
Director Posgrados de Ingeniería y Ciencias

Cali, 06 de julio de 2023

Acta de Correcciones al Documento de Trabajo de Grado

Santiago de Cali, 06 de julio de 2023

Autores: Andrés Mauricio Mena Ríos, Faber Esteban Hurtado Murillo, Jefferson Sánchez Andrade

Título del Trabajo de Grado: "Análisis Descriptivo y Predictivo para la Vigilancia de los Casos de Dengue Grave en la Ciudad de Cali"

Director: David Arango Londoño

Como indica el artículo 2.13 de las Directrices para Trabajo de Grado de Maestría, he verificado que el estudiante indicado arriba ha implementado todas las correcciones que los Jurados del Proyecto de Trabajo de Grado definieron que se efectuaran, como consta en el Acta de Evaluación correspondiente.

David Arango Londoño

_____

Firma del Director del Trabajo de Grado

Santiago de Cali, 06 de Julio de 2023

Ingeniero:
Juan Carlos Martínez Arias
Director Posgrados de Ingeniería
Facultad de Ingeniería y Ciencias
Pontificia Universidad Javeriana - Cali

Con el fin de cumplir con los requisitos exigidos por la Universidad para llevar a cabo el Trabajo de Grado y posteriormente optar por el título de Magíster en Ciencia de Datos, nos permitimos presentar a su consideración el proyecto de Trabajo de Grado denominado: Análisis Descriptivo y Predictivo para la Vigilancia de los Casos de Dengue Grave en la Ciudad de Cali, el cual será realizado por los estudiantes: Andrés Mauricio Mena Ríos con código 8973322, Faber Esteban Hurtado Murillo con código 0056314 y Jefferson Sánchez Andrade con código 8972756 perteneciente al énfasis en N/A, bajo la dirección del profesor David Arango Londoño.

El suscrito director del Trabajo de Grado autoriza para que se proceda a hacer la evaluación de este Proyecto ante el Tribunal que para el efecto se designe, toda vez que ha revisado cuidadosamente el documento y avala que ya se encuentra listo para ser presentado oficialmente.

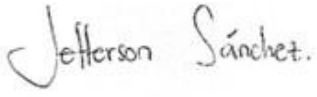
Atentamente,

Firma
Andrés Mauricio Mena Ríos

C.C. 1.088.355.333 de Pereira

Firma
Faber Esteban Hurtado Murillo

C.C. 1.116.433.397 de Zarzal



Firma
Jefferson Sánchez Andrade

C.C. 1.130.661.369 de Cali

Firma
David Arango Londoño

C.C. 1.130.586.950 de Cali

Maestría en Ciencia de Datos
Facultad de Ingeniería y Ciencias

FICHA RESUMEN
TRABAJO DE GRADO DE MAESTRÍA

TITULO: “Análisis Descriptivo y Predictivo para la Vigilancia de los Casos de Dengue Grave en la Ciudad de Cali”

1. ÉNFASIS: N/A
2. TIPO DE PROYECTO: Aplicado
3. ÁREA DE TRABAJO: Ciencia de Datos, *Machine Learning.*
4. ESTUDIANTES: Andrés Mauricio Mena Ríos; Faber Esteban Hurtado; Jefferson Sánchez Andrade.
5. CORREO ELECTRÓNICO: andres06@javerianacali.edu.co, fhurtado@javerianacali.edu.co, jsa198811@javerianacali.edu.co
6. DIRECCIÓN Y TELÉFONO:

   - Carrera 16 # 36-53, Dosquebradas, Risaralda, Colombia - 3058985087.
   - Calle 9 # 52-80, Cali, Colombia - 3185149704.
   - Carrera 85d # 54-60 Cali, Colombia - 3168696953.

7. DIRECTOR: David Arango Londoño.
8. VINCULACIÓN DEL DIRECTOR: Planta
9. CORREO ELECTRÓNICO DEL DIRECTOR: david.arango@javerianacali.edu.co
10. CO-DIRECTOR: No aplica.
11. GRUPO O EMPRESA QUE LO AVALA: No aplica.
12. OTROS GRUPOS O EMPRESAS: No aplica.
13. PALABRAS CLAVE: Modelo predictivo, *Machine Learning*, Dengue, Epidemiología, Cali Colombia.
14. ODS QUE APLICA EL PROYECTO (Agenda 2030): ODS 3: Salud y Bienestar, ODS 9: Industria, Innovación e Infraestructura y ODS 17: Alianzas para lograr los objetivos.
15. FECHA DE INICIO (Desarrollo del proyecto): 1/06/2022
16. RESUMEN: Este proyecto presenta un modelo predictivo para estimar la cantidad de casos de dengue grave en Cali. Incluye un análisis descriptivo de la enfermedad y la construcción de modelos predictivos basados en cuatro algoritmos de *Machine Learning* y el uso de fuentes de datos informales.

ANÁLISIS DESCRIPTIVO Y PREDICTIVO PARA LA VIGILANCIA DE LOS CASOS DE DENGUE GRAVE EN LA CIUDAD DE CALI

Andrés Mauricio Mena Ríos - Código 8973322
Faber Hurtado Murillo - Código 0056314
Jefferson Sánchez Andrade - Código 8972756

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

Director
MSc. David Arango Londoño

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, JULIO 04 DE 2023

CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

LISTA DE ANEXOS

INTRODUCCIÓN

Este proyecto de ciencia de datos desarrolla un modelo predictivo que permite estimar la cantidad de casos de dengue grave que ocurren en un determinado momento en la ciudad de Cali. Para eso, se realiza un análisis de la dinámica de la enfermedad, considerando aspectos como la temporalidad, la incidencia geográfica y algunas variables sociodemográficas; además, se construyen modelos predictivos basados en cuatro algoritmos de *Machine Learning*, el uso de fuentes de datos informales, y la incorporación de una variable novedosa como predictor.

La primera parte del proyecto se enfoca en análisis descriptivos del dengue grave en Cali, a partir del procesamiento de los registros históricos oficiales, con el propósito de comprender patrones y tendencias de la enfermedad e identificar factores relacionados con su incidencia. La segunda sección del proyecto gira alrededor de la determinación del mejor modelo para predecir la cantidad de casos de dengue en Cali, haciendo uso de una variedad de recursos de la ciencia de datos para la construcción, evaluación y análisis de los candidatos.

1. DEFINICIÓN DEL PROBLEMA

1.1. PLANTEAMIENTO DEL PROBLEMA

La enfermedad del dengue es un problema de salud pública vigente y alarmante. La Organización Mundial de la Salud (OMS) la considera la enfermedad transmitida por mosquitos más importante del mundo". Se estima que anualmente hay 400 millones de infecciones y 500 mil casos de dengue grave alrededor del mundo. Además, la incidencia del dengue ha aumentado en gran medida. La cantidad de casos de dengue notificados por la OMS se multiplicó por 8 en las últimas dos décadas, alcanzando un récord histórico en el año 2019, y ratificando que los casos mortales se concentran en el grupo etario más joven [1].

En Colombia, factores como las características geográficas y climatológicas, la presencia de diferentes serotipos del virus y la existencia del vector en la mayoría de los municipios propician la proliferación de los casos de dengue y dengue grave. A pesar de esto, se considera que la amenaza de la enfermedad no ha sido evaluada adecuadamente, y que hay deficiencias importantes en la vigilancia de la enfermedad a varios niveles. De hecho, la OMS ha admitido y documentado que el incremento en la carga de la enfermedad en Colombia durante los últimos años se debe en parte a un sistema de vigilancia pasivo e inoperante [1].

La situación en Cali es más inquietante. De acuerdo con el Instituto Nacional de Salud, Cali usualmente está en el top de entidades territoriales que aportan el Pareto de los casos de dengue a nivel nacional, y es común que se destaque por encima del percentil 75 en todas las clasificaciones de incidencia de dengue, incluyendo: dengue en general, dengue con signos de alarma y dengue grave [2]. Al cierre del 2020, la tasa de incidencia de dengue grave por cada 100.000 habitantes en el país fue de 1.8, mientras que en Cali alcanzó un récord de 8,3.

Respecto a este último indicador, la importancia del dengue grave radica en que es una complicación potencialmente mortal. Los pacientes que manifiestan determinados síntomas durante su fase crítica requieren observación médica estrecha e inmediata con manejo hospitalario, además de vigilancia durante la fase de convalecencia, para brindar la atención necesaria y evitar complicaciones mayores y el riesgo de muerte [3]. Estas condiciones hacen del dengue grave un objeto de interés particular, porque la cantidad de personas que desarrollan este tipo de complicaciones son un indicador clave para la gestión de las capacidades del sistema de salud local.

Considerando lo anterior, desde la perspectiva de las instituciones de salud, especialmente las que pertenecen a la red de salud pública, un desafío apremiante es afrontar mejor los aumentos súbitos de casos de dengue grave, debido a que, en el evento imprevisto de un brote epidémico, los procesos administrativos necesarios para asignar los recursos para atender la emergencia se ven sometidos a una gran presión, y con ello, se corre el riesgo de no ofrecer una respuesta efectiva. De este modo, el problema que se aborda en este proyecto se enfoca en el alto nivel de incertidumbre que experimenta el sistema local de atención y gestión de la salud respecto a la dinámica del dengue grave en la ciudad de Cali, y, en consecuencia, en la gestión del riesgo de operar desde un enfoque puramente reactivo, que podría comprometer la eficacia y eficiencia de los procesos de atención a las personas con dengue grave.

1.2. FORMULACIÓN DEL PROBLEMA

1. ¿Cómo se podría disminuir la incertidumbre sobre la dinámica del dengue grave en Cali para mejorar potencialmente las capacidades de vigilancia del sistema de salud pública municipal, de modo que pueda fortalecer sus estrategias de prevención y control?

2. ¿De qué manera se pueden determinar la cantidad de casos de dengue grave en Cali, de modo que la red de salud local pública pueda realizar una gestión oportuna y efectiva?

3. ¿Qué variables serían las más relevantes para un modelo de predicción de los casos de dengue grave en Cali?

2. OBJETIVOS DEL PROYECTO

2.1. OBJETIVO GENERAL

Diseñar un modelo predictivo de los casos de dengue grave en Cali para mejorar potencialmente las capacidades de control y vigilancia de la red de salud pública municipal, mediante el uso de técnicas de *Machine Learning*.

2.2. OBJETIVOS ESPECÍFICOS

1. Identificar las tendencias y patrones de los casos de dengue grave en Cali para desarrollar hipótesis que expliquen y ayuden a predecir su ocurrencia.
2. Observar los cambios en los patrones de los casos de dengue grave en Cali para determinar la existencia de un brote epidémico de dengue.
3. Explorar factores meteorológicos y sociodemográficos relacionados con los casos de dengue grave en Cali para detectar cambios en su cantidad y distribución.
4. Estimar un modelo de aprendizaje supervisado basado en factores meteorológicos y flujos de búsquedas en internet para predecir los casos de dengue grave en Cali.
5. Evaluar el desempeño del modelo de aprendizaje supervisado construido en el objetivo anterior.

3. MARCO TEÓRICO Y ANTECEDENTES

3.1. MARCO TEÓRICO

El marco teórico de este proyecto se divide en tres partes, así: primero, una referencia a la enfermedad del dengue; segundo, las técnicas empleadas; y tercero, un estándar empleado para el desarrollo del proyecto.

3.1.1. EL DENGUE

Para comenzar, para el desarrollo de este proyecto se ha elegido como objeto de estudio la enfermedad del dengue. Según la Organización Mundial de la Salud, el dengue es una infección vírica. El principal vector de la enfermedad es el mosquito *Aedes aegypti*. Existen cuatro serotipos del virus que causa la enfermedad, y en consecuencia es posible infectarse cuatro veces. Con todo, hasta ahora no existe un tratamiento específico para el dengue. [3]

Gutiérrez-Barbosa et al. [4] realizaron un análisis de los casos de dengue en Colombia desde 1970 hasta el 2020, basados en publicaciones científicas, para comparar el nivel de incidencia y letalidad nacional respecto al comportamiento en la región. El estudio incluye una discriminación a nivel de serotipos, determina los principales brotes históricos, y monitorea la cantidad de casos severos. Los resultados soportan la noción de que el dengue sigue siendo una enfermedad de interés en nuestro país, y que los esfuerzos por mejorar la respuesta del sistema de salud son pertinentes.

La vigilancia epidemiológica de las enfermedades de interés público en Colombia se encuentra a cargo del Instituto Nacional de Salud, quien realiza periódicamente el Boletín Epidemiológico Semanal, en el cual se analiza el comportamiento de diferentes enfermedades, entre ellas, el dengue, donde puede observarse el número de casos por entidad territorial, la gravedad de los síntomas, la incidencia de casos a lo largo del tiempo [2], entre otros análisis, cuyas fuentes de datos sirvieron de insumo en la elaboración del proyecto de grado.

3.1.2. TÉCNICAS EMPLEADAS

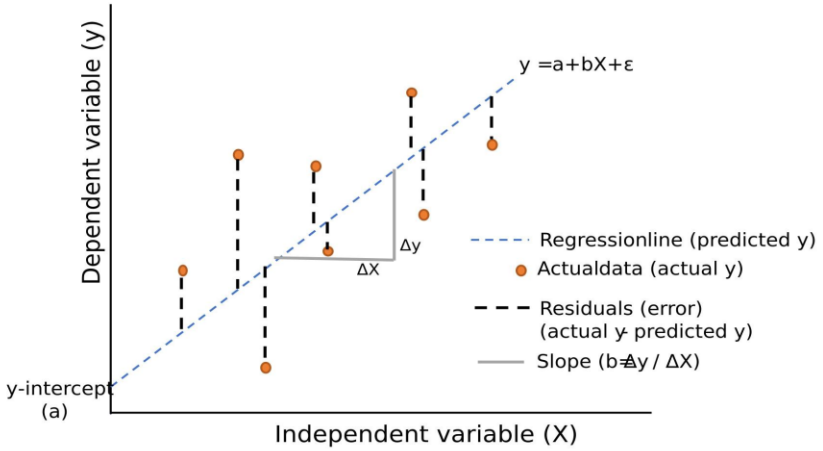
En segundo lugar, para el desarrollo de este proyecto se emplearon técnicas descriptivas y predictivas de la Ciencia de Datos, más precisamente, modelos supervisados de minería de datos, según Gordon S. et al. [5] ya que se conocía a-priori la variable objetivo, la cual está incluida en la base de datos histórica a ser procesada; a continuación se brinda una descripción de los modelos empleados, de acuerdo con James et al. [6]:

Regresión Lineal: La regresión lineal es una técnica estadística que busca establecer una relación entre una variable dependiente y una o más variables independientes. Es un método ampliamente utilizado para predecir o modelar el comportamiento de una variable objetivo basándose en variables predictoras.

El enfoque básico de la regresión lineal consiste en ajustar una línea recta a un conjunto de datos para representar la relación entre la variable dependiente (también conocida como variable respuesta) y las variables independientes (también conocidas como variables predictoras); el objetivo es encontrar los coeficientes de la línea recta que minimicen la diferencia entre los valores observados y los valores predichos. La Figura 1 presenta gráficamente los elementos de una regresión lineal.

Es importante resaltar que regresión lineal asume que la relación entre la variable dependiente y las variables independientes es lineal y que los errores de predicción siguen una distribución normal; el modelo resultante se puede utilizar para hacer predicciones sobre nuevos conjuntos de datos o para entender la influencia de las variables predictoras en la variable respuesta.

Figura 1. Esquema de una Regresión Lineal.



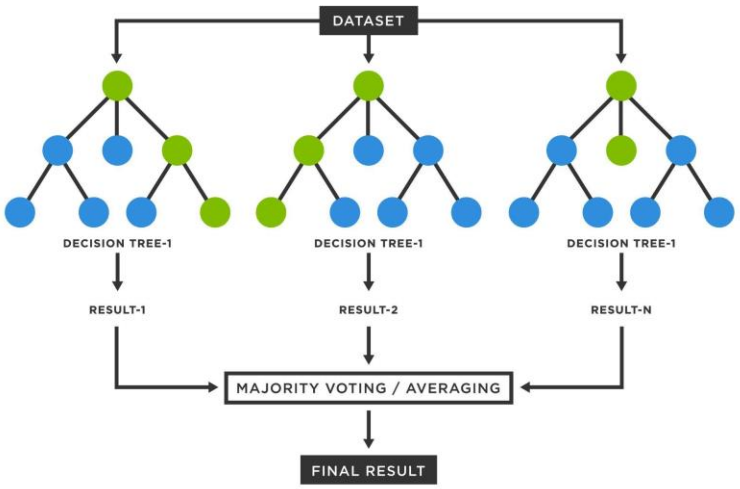
Fuente: Bedre [7].

Random Forest: Esta técnica combina múltiples árboles de decisión para realizar predicciones precisas y mejorar la capacidad de generalización del modelo. En *Random Forest*, se crea un conjunto (o "bosque") de árboles de decisión mediante un proceso llamado *bagging* (ensacado). Cada árbol se entrena con una muestra aleatoria del conjunto de datos de entrenamiento, y las predicciones finales se obtienen mediante la combinación de las predicciones de todos los árboles; una representación gráfica se puede visualizar la Figura 2.

La característica distintiva de *Random* Forest es que, en cada nodo de división de un árbol, solo se considera un subconjunto aleatorio de variables predictoras. Esto evita que un solo predictor dominante influya demasiado en las decisiones de división, lo que puede mejorar la robustez y la capacidad de generalización del modelo.

Random Forest es especialmente útil tanto en problemas de clasificación como regresión, donde se busca predecir la pertenencia a una clase o el valor numérico de una variable objetivo. Además de proporcionar predicciones precisas, *Random* Forest también puede evaluar la importancia relativa de las variables predictoras en el proceso de clasificación o regresión.

Figura 2. Representación gráfica del modelo *Random* Forest.

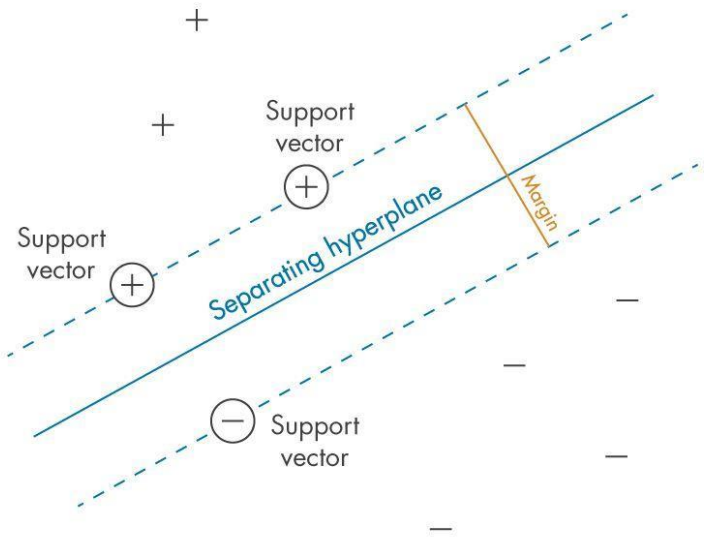


Fuente: Tibco [8].

***Support Vector Machine*:** El algoritmo de Máquina de Vectores de Soporte (SVM, por sus siglas en inglés) es un método de aprendizaje supervisado ampliamente utilizado en clasificación y regresión. Se aplica en diversas áreas, como el procesamiento de señales médicas, el procesamiento del lenguaje natural y el reconocimiento de imágenes y voz [9].

El objetivo principal del algoritmo SVM es encontrar un hiperplano que pueda separar de manera óptima dos clases diferentes de datos. La optimización implica encontrar el hiperplano con el margen más amplio entre las dos clases, representado por los signos positivo y negativo en la Figura 3. El margen se define como la distancia máxima entre el hiperplano y los puntos de datos más cercanos de cada clase. En problemas en los que la separación lineal no es posible, el algoritmo busca maximizar un margen flexible que permita un pequeño número de clasificaciones erróneas.

Figura 3. Definición del margen entre clases, el cual es optimizado con SVM.



Fuente: MathWorks [10].

Los vectores de soporte se refieren a un subconjunto de las observaciones de entrenamiento que determinan la ubicación del hiperplano de separación. El algoritmo SVM estándar se formula principalmente para problemas de clasificación binaria, aunque los problemas de clasificación multiclase se pueden reducir a una serie de problemas binarios.

***LightGBM*:** *Light Gradient Boosting Machine* es un algoritmo de aprendizaje automático de última generación, que busca implementar de manera eficiente y escalable el algoritmo de *boosting*, el cual es una técnica popular para crear modelos de aprendizaje automático con alta precisión.

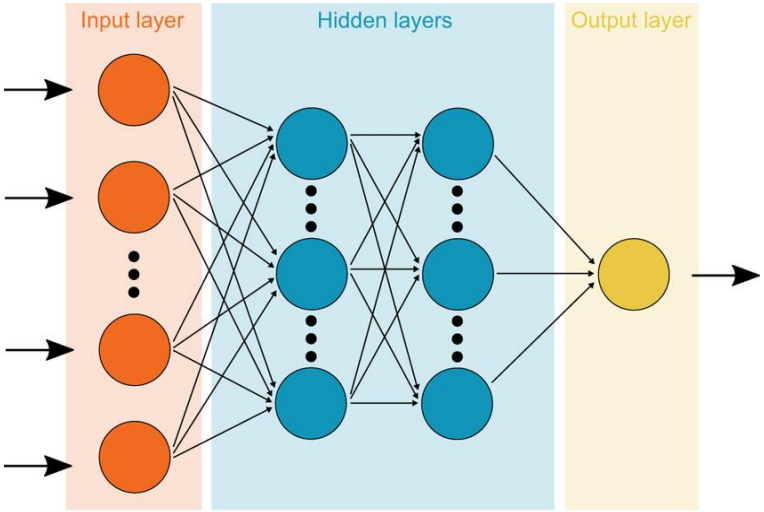
LightGBM utiliza un enfoque basado en árboles, donde se construye un conjunto de árboles de decisión secuencialmente, y cada árbol se ajusta a los errores residuales del modelo anterior;

siendo así, no se presenta una representación gráfica específica de este algoritmo, y es posible referirse nuevamente a la Figura 2 para evidenciarla. Sin embargo, a diferencia de otros algoritmos de *boosting*, *LightGBM* GBM utiliza la técnica de "*gradient-based one-side sampling*" para realizar divisiones en los nodos de los árboles de manera más eficiente. Esto permite un entrenamiento más rápido y una mayor escalabilidad, especialmente en conjuntos de datos grandes.

Una característica destacada de *LightGBM* es su capacidad para manejar variables categóricas de manera eficiente, sin necesidad de realizar codificaciones previas de las mismas. Esto es especialmente útil en problemas de aprendizaje automático donde las variables predictoras pueden ser de naturaleza categórica o nominal.

Red Neuronal Artificial: Una Red Neuronal Artificial (ANN) es un modelo de aprendizaje supervisado que se utiliza para la clasificación o regresión. La red está compuesta por múltiples capas de neuronas interconectadas que procesan los datos de entrada para producir una salida. Cada neurona en la red utiliza una función de activación para determinar su salida en función de las entradas recibidas y los pesos asociados. A manera de ilustración, la Figura 4 presenta un esquema de una red neuronal artificial con una capa de entrada, dos capas ocultas y una capa de salida.

Figura 4. Esquema de una Red Neuronal Artificial.



Fuente: Irrgang et al. [11].

3.1.3. CRISP-DM

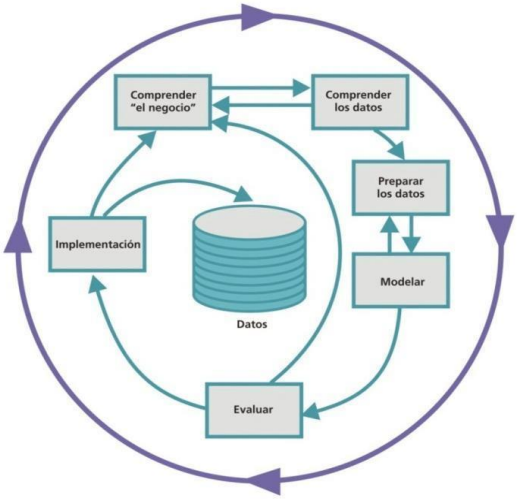
En tercer lugar, para el desarrollo de este proyecto se aplicó la metodología de desarrollo de proyectos de minería de datos denominada CRISP-DM (*Cross Industry Standard Process for Data Mining*), que se caracteriza por presentar una estrategia para el desarrollo eficaz de modelos analíticos a partir del entendimiento de las necesidades de los interesados.

En el marco de esta metodología se consideran las siguientes fases, que rigen el ciclo de vida del proyecto para dar contexto al alcance y aplicación del modelo de analítica:

- Fase I. Entendimiento del Negocio.
- Fase II. Entendimiento de la Data.
- Fase III. Preparación de los Datos.
- Fase IV. Modelado.
- Fase V. Evaluación.
- Fase VI. Despliegue.

CRISP – DM es una metodología abierta, que nace del esfuerzo de distintas organizaciones como como SPSS, NCR y DaimlerChrysler, que consta de un proceso iterativo de 6 fases de secuencia no rígida (ver Figura 5), donde es posible volver a la fase anterior para generar retroalimentación y perfeccionamiento en el proceso de analítica; en términos generales, es una metodología cíclica que permite generar resultados y oportunidades de mejora en cada una de sus fases para así generar un último resultado pueda ser implementado en un proyecto [12].

Figura 5. Fases del modelo CRISP – DM.



Fuente: Bellini Saibene et al. [13].

Tomando como referencia el trabajo de Espinoza Zúñiga [14], de manera específica cada una de las fases CRISP - DM hace alusión a lo siguiente:

Fase I. Entendimiento del Negocio.

Esta fase se centra en la definición de las necesidades de los interesados. Se trata de entender los objetivos y requerimientos del proyecto desde una perspectiva organizacional para convertir ese conocimiento en la definición de un problema de analítica de datos y una propuesta inicial de plan de trabajo sobre cómo abordarlo.

Fase II. Entendimiento de la Data.

Esta fase se enfoca en el estudio y comprensión de los datos. Se trata de obtener, examinar y familiarizarse con los datos, identificar problemas de calidad de datos y, eventualmente, identificar el potencial de ciertos subconjuntos de datos.

Fase III. Preparación de los Datos.

Esta fase trata fundamentalmente del análisis de los datos. Los datos deben pasar por un proceso de limpieza, depuración, selección y homogeneización para garantizar calidad, pertinencia y asertividad en los resultados de los modelos. El objetivo de esta fase es contar con una base operable. Aquí se incluyen todas las operaciones de transformación que tienen que ver con la adecuación de los datos a la tarea de analítica.

Fase IV. Modelado.

A partir de lo hecho en las fases anteriores, se determinan las opciones de analítica para aplicar y se evalúa su pertinencia conforme a los objetivos definidos. Se obtiene el modelo de analítica de datos que permite dar respuesta a las preguntas orientadoras de la organización.

Fase V. Evaluación.

En esta fase se evalúan los resultados obtenidos en función de los objetivos del negocio, es decir, se determina si los mismos pueden responder a los requerimientos de la organización.

Fase VI. Despliegue.

Esta fase consiste en la puesta en producción. Se trata de explotar la potencialidad de los modelos, integrarlos en los procesos de toma de decisión de la organización, difundir informes sobre el conocimiento extraído y estandarizar el proceso de analítica en los distintos procesos para ir generando potencialidades a partir de la utilización estratégica de la información.

3.2. ANTECEDENTES

Una fuente importante de consulta de trabajos previos fue el metaanálisis de Sylvestre et al [15]. Dicha investigación contiene referencias que coinciden con ciertos elementos distintivos de este proyecto, a saber: se centra en iniciativas que utilizan métodos de *Machine Learning* y hacen uso de datos tomados del mundo real, algunos de fuentes no tradicionales y no clínicas, para monitorear y predecir aspectos relacionados con el dengue. La investigación revisa un conjunto de publicaciones académicas fechadas entre el año 2000 y el 2020, e incluye 119 artículos y documentos de ponencias con iniciativas desarrolladas alrededor del mundo. Entre los *insights* más relevantes del metaanálisis para el propósito de este proyecto están:

1. Hay una gran cantidad de iniciativas que han demostrado que es posible robustecer los sistemas de vigilancia del dengue con la ayuda de algoritmos de *Machine Learning.*
2. Los árboles de decisión y las redes neuronales son las técnicas de mejor desempeño para predecir diferentes aspectos del dengue.
3. Desde una perspectiva tradicional, las tres variables climatológicas clave para predecir los casos de dengue a nivel local son: precipitación, temperatura y humedad.
4. El volumen de búsquedas en línea, entre otros flujos de datos no convencionales, son variables prometedoras para analizar la dinámica del dengue.

Respecto a la selección de variables de interés, en general, los sistemas de vigilancia epidemiológica formal recopilan información que incluye datos como el recuento de casos, los reportes serológicos de laboratorio y los análisis de la población de vectores, entre otros. Sin embargo, dichos sistemas de vigilancia tradicionales afrontan limitaciones comunes, como demoras en la consolidación y baja calidad de datos, subregistro, y altos costos de mejoramiento, entre otros. Para hacerle frente a dichas limitaciones, se han desarrollado iniciativas que buscan aprovechar fuentes novedosas de datos informales, que sean útiles para modelar el comportamiento de enfermedades infecciosas como el dengue. En este contexto, la evolución del internet y los motores de búsqueda han ofrecido una oportunidad, y han dado lugar a lo que algunos denominan “métodos de vigilancia basados en flujos de búsquedas”, partiendo de la idea de que las consultas en la web podrían ser generadas por personas que tienen, o conocen a alguien que tenga dengue, para saber un poco más sobre la enfermedad. Se ha evidenciado la utilidad y sensibilidad de los datos disponibles a través de estas aproximaciones para mejorar el reporte y la oportunidad del reporte de los brotes [16].

En el contexto nacional, A. A. M. González et al. [17] presentan un modelo bayesiano para el estudio del dengue en el departamento del Atlántico. Se trata de un trabajo relevante porque explora la relación entre los casos de dengue y factores fuera del sector salud, como algunos datos sociales, geográficos y económicos, para establecer áreas del departamento con mayor riesgo de la enfermedad. Esto es importante porque se corresponde con uno de los elementos del alcance del presente proyecto, aunque este último se enfoque en Cali. Considerando lo anterior, se aprovecharon los hallazgos de este trabajo previo para crear un inventario preliminar de variables sociodemográficas y económicas de interés, y también se exploraron algunos detalles útiles de la modelación, especialmente el uso de la tasa de mortalidad estandarizada para aislar el ruido de la variable dependiente, como insumo para el desarrollo del presente proyecto.

En el panorama local, L. Sepúlveda-Salcedo et al. [18] presentan un modelo matemático para la dinámica del dengue en Cali. Se trata de una adaptación del modelo de *Ross-Macdonald*, que está basado en ecuaciones diferenciales ordinarias, y que se utiliza para analizar los datos de un brote de dengue. El estudio incluye datos como: la población de mosquitos, la tasa de picadura, la proporción de humanos y mosquitos infectados, entre otros. En general, se trata de un trabajo que presenta información relevante para el presente proyecto, por ejemplo, identifica la fuente oficial de los datos, y presenta el comportamiento cualitativo de la cantidad de personas contagiadas con dengue cuando ocurre un brote epidémico. Sin embargo, su objetivo, las variables de interés y el método seleccionado evidencian grandes diferencias y, por tanto, dan lugar a una perspectiva de complementariedad para analizar el fenómeno en cuestión.

4. PREPARACIÓN DE DATOS

Esta sección consta esencialmente de dos partes: la primera se dedica a los análisis descriptivos necesarios para develar la dinámica del dengue grave en Cali; y la segunda parte se centra en el alistamiento de los insumos relevantes para la etapa de modelación.

4.1. PRIMERA PARTE - ANÁLISIS DESCRIPTIVO DE LA DINÁMICA DE DENGUE GRAVE EN CALI

Obtención de los datos iniciales

La Secretaría de Salud Pública de Cali suministró el conjunto de datos original, el cual consta de 573 registros y 8 variables, y se caracteriza en la Tabla 1.

Tabla 1. Caracterización general del conjunto de datos inicial.

Nombre	Variable	Descripción	Tipo de variable	Tipo de escala	Tipo
cod_eve	Código del evento	Identificador único de la enfermedad del dengue grave	Cualitativa	Nominal	float64
fec_not	Fecha de notificación	Fecha en la que se notificó la condición médica	Cuantitativa continua	Intervalo	datetime 64
edad_	Edad	Edad del paciente en el momento del diagnóstico	Cuantitativa continua	Razón	int64
sexo_	Sexo del paciente	Género del paciente	Cualitativa	Nominal	object
bar_ver_	Barrio o vereda de residencia	Ubicación geográfica de la residencia del paciente	Cualitativa	Nominal	object
dir_res_	Dirección de residencia	Dirección física completa donde vive o reside el paciente	Cualitativa	Nominal	object
nom_upgd	Nombre de la unidad primaria generadora de datos	Institución de salud que realizó la notificación	Cualitativa	Nominal	object
nreg	Número de registro	Identificador único asignado al paciente en la base de datos	Cuantitativa discreta	Nominal	int64

Fuente: Elaboración propia.

La base de datos inicial tiene tres variables con valores faltantes, así:

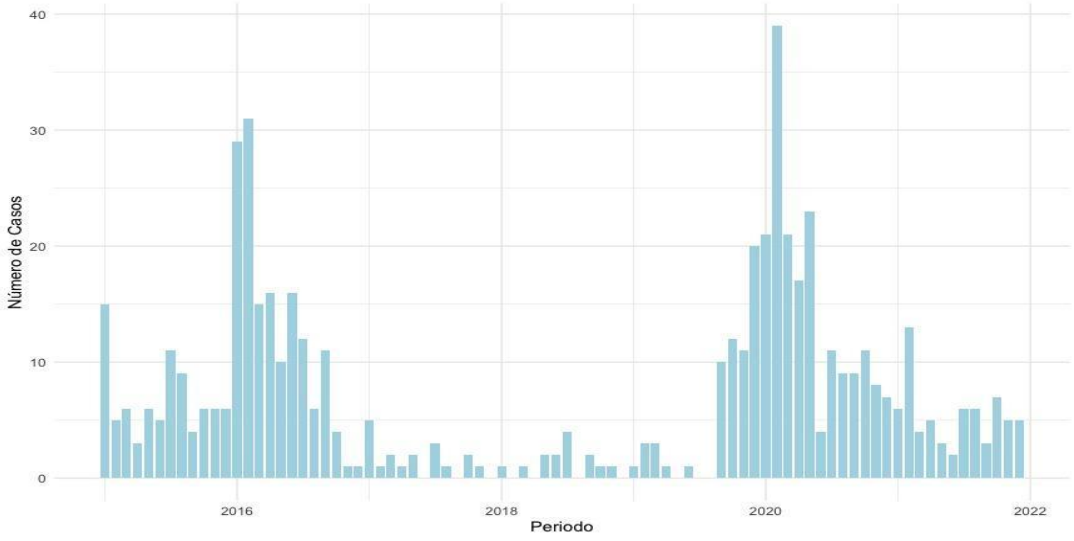
- “Barrio o vereda de residencia”, 126 registros sin datos.
- “Dirección de residencia” 37 registros sin datos.
- “Nombre de la unidad primaria generadora de datos”, 1 registro sin datos.

A partir de este punto, comenzó un proceso iterativo de exploración de la dinámica de la enfermedad en el contexto local. Ese proceso requirió buscar información adicional al conjunto de datos inicial para responder a las inquietudes que surgieron a lo largo del desarrollo del proyecto, como se evidencia a continuación.

Análisis de la tendencia temporal

Algunas de las primeras inquietudes apuntaban a la necesidad de realizar un análisis de la tendencia temporal de la enfermedad. Para comenzar, la Figura 6 muestra la evolución de la cantidad de casos mensuales de dengue grave en Cali durante el periodo de análisis. Ahí se evidencian algunos patrones importantes, como, por ejemplo, que la enfermedad está presente a lo largo de todo el año en la ciudad. También se observa que hubo dos grandes picos de contagio, uno en 2016 y otro en 2020, con una diferencia de cuatro años entre ellos. La cantidad de casos entre picos forman valles irregulares. Sin embargo, hasta este punto, no hay una tendencia general clara de aumento o disminución de los casos de dengue.

Figura 6. Diagrama de Barras - Cantidad mensual de casos de dengue grave en Cali en el periodo 2015 - 2021.

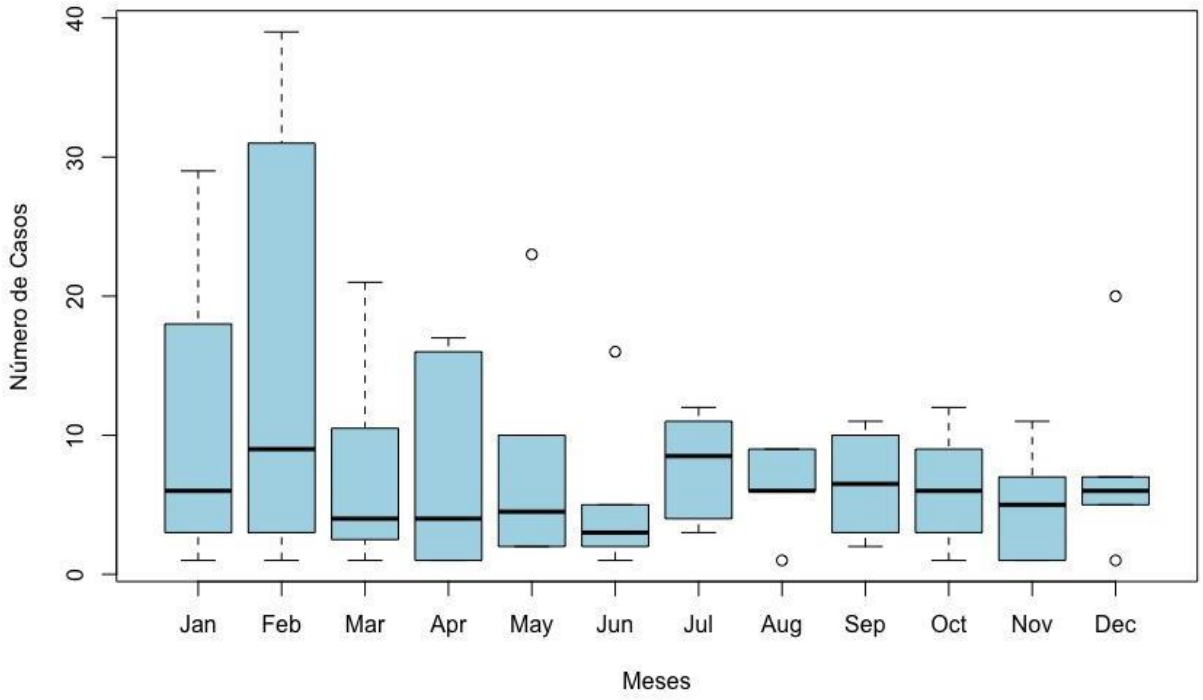


Fuente: Elaboración propia.

Análisis de casos mensualizados

Una aproximación diferente permitió identificar un patrón importante. Al agrupar los casos de dengue grave en función de los meses de ocurrencia, tal como se muestra en la Figura 7, se evidenció que, aunque los valores medios podrían sugerir homogeneidad, el rango mensual y las formas de las cajas revelan un patrón, según el cual, típicamente la mayor cantidad de casos de dengue grave se concentra en los primeros meses del año. Este es un hallazgo importante porque podría justificar el diseño de un protocolo que permita ofrecer una atención oportuna de las personas enfermas, especialmente durante los meses de mayor exigencia.

Figura 7. *Boxplot* - Cantidad mensual de casos de dengue grave en Cali en el periodo 2015 - 2021.



Fuente: Elaboración propia.

Prueba de estacionalidad

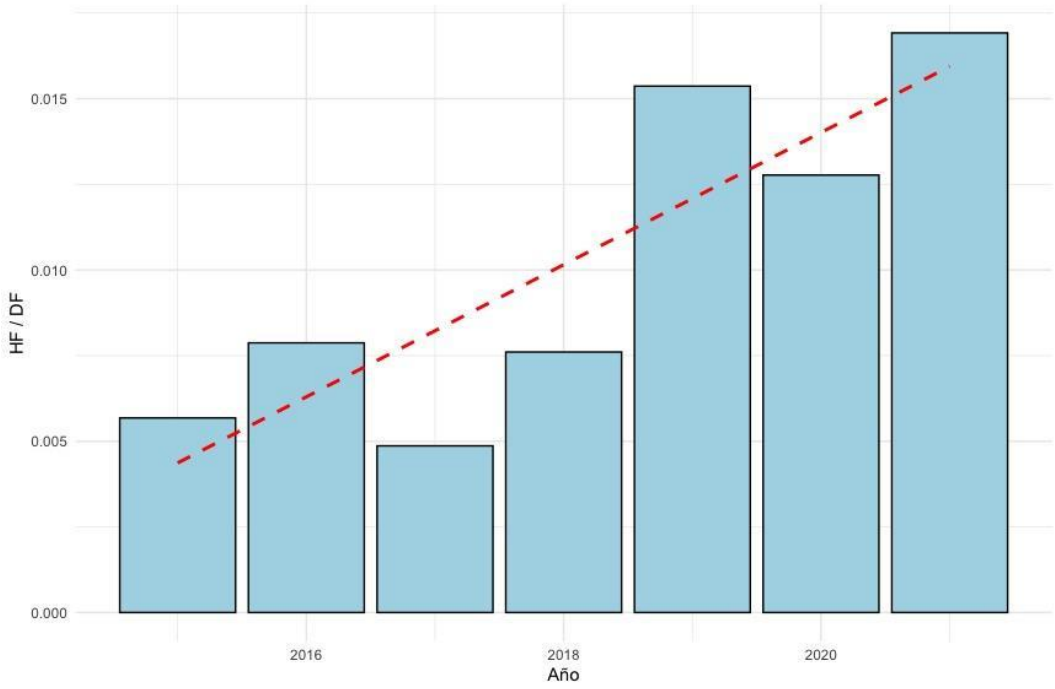
Un paso adicional en el análisis temporal fue utilizar la prueba de *Dickey-Fuller* Aumentada (ADF) para evaluar si las propiedades estadísticas de la variable "cantidad de casos de dengue grave" permanecen constantes a lo largo del tiempo. La hipótesis nula de esta prueba estadística es que

existe una raíz unitaria, lo que significa que la serie no es estacionaria; mientras que la hipótesis alterna básica es que la serie es estacionaria [19]. Una vez ejecutada la prueba ADF, se obtuvo un p-valor de 0.0557 y, así las cosas, no se tiene suficiente evidencia estadística para decir que se trata de una serie estacional. A partir de esto, se concluyó que se trata de una variable que se comporta distinto a lo largo del tiempo, lo cual refleja una alta complejidad para la tarea de identificar sus patrones subyacentes.

Dengue grave versus dengue

Un último paso en la búsqueda de tendencias temporales de la enfermedad consistió en comparar la cantidad anual de casos de dengue y dengue grave, según se presenta en la Figura 8. Esta comparación permitió evidenciar una tendencia inquietante: la proporción de los casos graves respecto al total de casos de dengue tiende a incrementarse.

Figura 8. Diagrama de barras - Proporción de los casos de dengue grave sobre los casos de dengue en Cali (HF/DF).



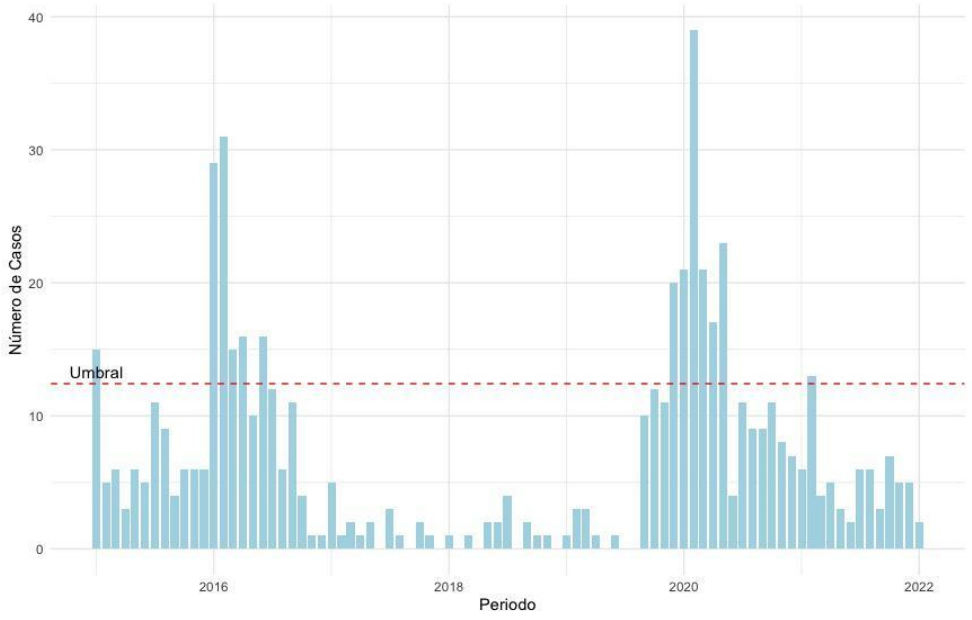
Fuente: Elaboración propia.

Determinación de un brote epidémico de dengue grave

En función de los *insights* conseguidos hasta ahora, y recurriendo a la literatura, se realizó una propuesta para determinar la existencia de un brote de dengue grave. Esta tarea típicamente implica analizar la incidencia de la enfermedad en una zona específica y establecer si esta supera un umbral definido a partir de registros históricos. Una práctica común consiste en calcular ese umbral basado en la media y la desviación estándar de los datos. Sin embargo, algunos autores afirman que ese método no es el más aconsejable y, en su lugar, proponen el uso de recursos asociados a la teoría de valores extremos [20]. Siguiendo ese tipo de recomendaciones, la propuesta específica en este proyecto consiste en determinar la existencia de un brote de dengue grave en Cali al comparar la cantidad de infecciones de un mes contra un valor de referencia calculado a partir del uso de la distribución generalizada de Pareto, utilizando con todos los registros históricos.

Considerando los picos históricos y la conveniencia de detectar un brote en forma temprana y con menos casos, se utilizó el percentil 85 para definir el umbral en cuestión. El resto de los parámetros necesarios se calcularon a través del ajuste automático a los datos disponibles. Todo esto condujo a definir que la cantidad que sirve de umbral para establecer la existencia de un brote son 13 casos de dengue grave en Cali.

Figura 9. Diagrama de Barras - Umbral para detectar un brote de dengue grave en Cali.



Fuente: Elaboración propia.

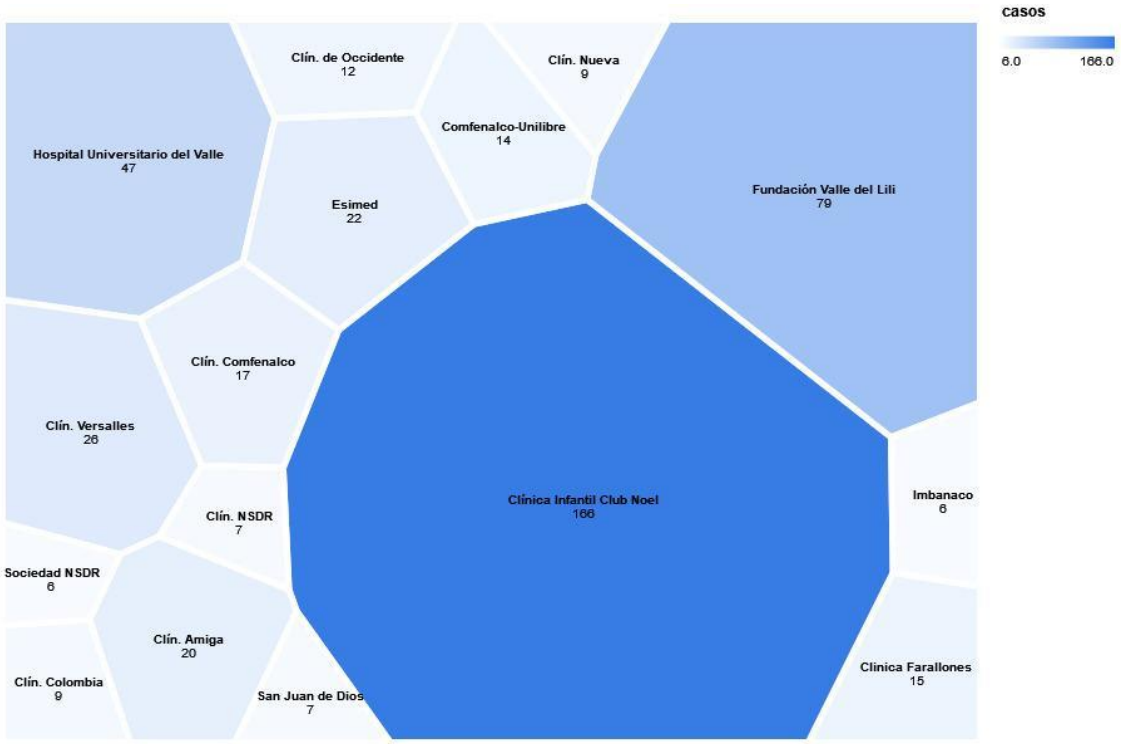
Exploración de las otras variables del conjunto de datos inicial

Una vez satisfechas las cuestiones iniciales con la exploración y análisis de los casos de dengue grave en forma independiente, se procedió a comenzar a vincular las demás variables del conjunto de datos inicial. La primera variable en ese proceso expansivo fue el atributo que identifica la institución que atendió y reportó cada uno de los casos de la enfermedad, según se aborda a continuación.

Instituciones de salud

Según los registros oficiales, 71 instituciones de salud reportaron casos de dengue grave en Cali entre 2015 y 2021. Así que se realizó un análisis para identificar el top de instituciones con mayor cantidad de casos. La Figura 10 presenta las entidades Pareto, y permite advertir tres protagonistas: la Clínica Infantil Club Noel, la Fundación Valle del Lili y el Hospital Universitario del Valle. Estas tres instituciones atendieron y reportaron el 51% de los casos de la enfermedad.

Figura 10. Diagrama de *Voronoi* - Pareto de las instituciones de salud según la cantidad de casos de dengue grave reportados en Cali en el periodo 2015-2021.



Fuente: Elaboración propia.

Lo anterior es un hallazgo importante porque se puede mejorar la gestión de la enfermedad al considerar la concentración del volumen de procesos de atención. A manera de ilustración, se podría evaluar la posibilidad de priorizar la asignación de recursos, la logística de alistamiento y los acuerdos interinstitucionales que soporten un manejo efectivo de los casos de dengue grave, especialmente en el marco de un brote epidémico.

Estrategia de análisis de variables sociodemográficas

El análisis de factores sociodemográficos, y su posible relación con los casos de dengue grave en Cali, se realizó de la siguiente manera:

a) Explorar los datos demográficos incluidos en la base de datos inicial.
b) Realizar un proceso de geolocalización de los registros individuales.
c) Realizar un análisis en busca de patrones temporo-espaciales de los casos por comuna.

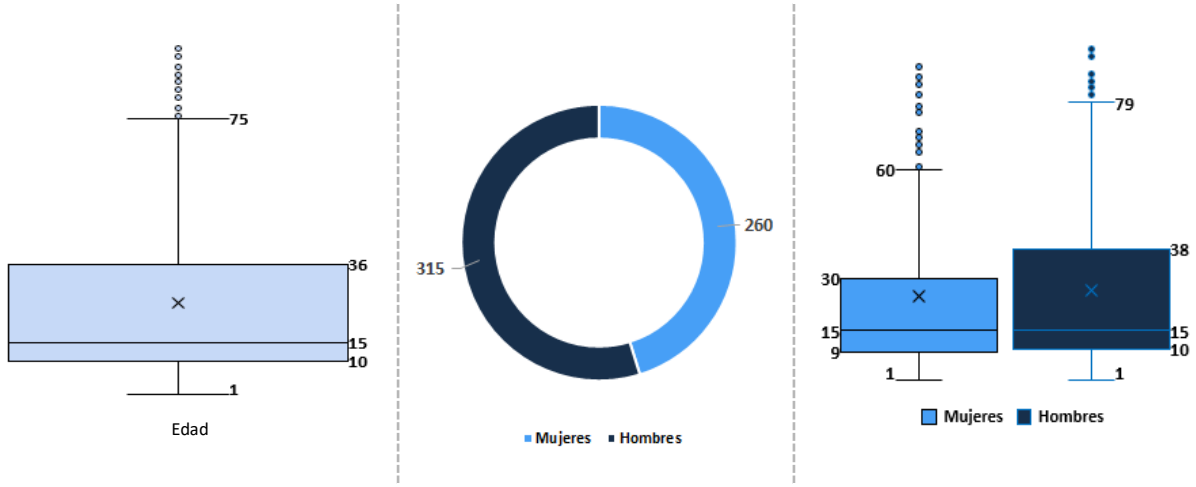
El desarrollo de estos pasos y sus resultados se presentan a continuación.

Variables sociodemográficas disponibles

Retomando la base de datos inicial, se realizó un análisis que generó las primeras impresiones sobre la importancia de las variables sociodemográficas en la dinámica del dengue grave en Cali. Inicialmente, se exploraron las variables disponibles: edad y sexo, como se observa en la Figura 11. Algunos hallazgos iniciales fueron:

- Hay una gran concentración de casos de la enfermedad entre las personas más jóvenes. Esto se evidencia al considerar que, aunque el rango de edad de las personas afectadas es amplio, alcanzando los 75 años, el 50% de los registros corresponde a personas de hasta 15 años.
- En forma independiente, el sexo del paciente no parece un factor relevante, porque la proporción entre hombres y mujeres no resulta tan desigual.
- Al cruzar las variables edad y sexo, la principal novedad es que el rango de edad de las mujeres afectadas es más pequeño, sobre todo en la mitad superior de los registros (Q3 y Q4).

Figura 11. Caracterización inicial de los casos de dengue grave en función de variables sociodemográficas.



Fuente: Elaboración propia.

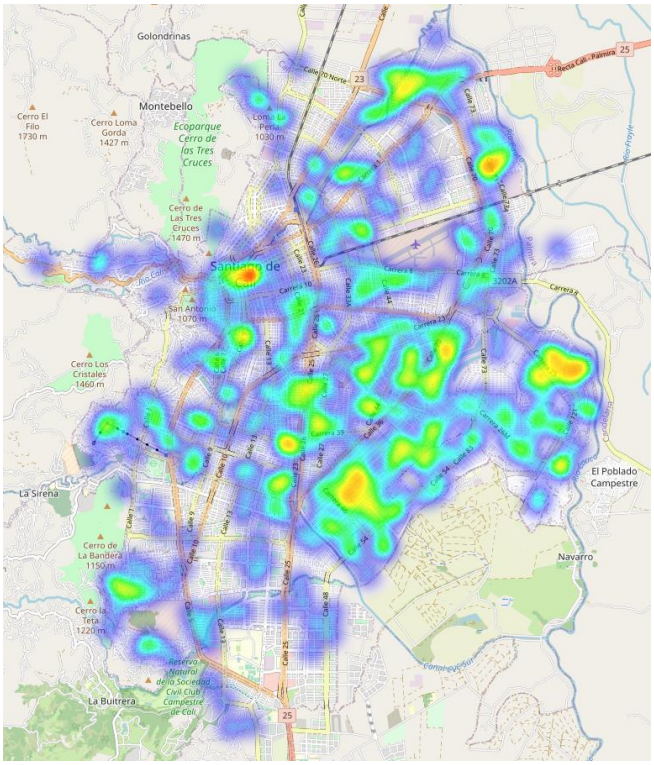
Geolocalización de los datos individuales

Habiendo agotado las posibilidades de exploración directa de variables sociodemográficas de la base de datos inicial, se procedió a geolocalizar los registros individuales, a fin de obtener insumos para realizar análisis complementarios. Esto requirió, por un lado, la estandarización de las direcciones de residencia; y, por otro lado, la imputación de datos faltantes, cuando hubiera lugar a ello, valiéndose de la variable "barrio o vereda de residencia". Una vez completada dicha rutina de estandarización, se utilizó el API de geolocalización de Google para obtener dimensiones geográficas: latitud, longitud, barrio y comuna para cada registro.

Distribución geográfica de los casos individuales

Con la nueva información obtenida, se procedió a identificar la distribución geográfica de los casos individuales de dengue grave en Cali. Para tales efectos, se elaboró un mapa de calor, en el que los colores rojo y naranja representan los puntos de mayor incidencia de la enfermedad, y los colores azul y verde indican una baja incidencia, según se muestra en la Figura 12. A partir de esto, se identificaron sectores de interés especial, porque tuvieron una alta concentración de casos de dengue grave durante el periodo de análisis: uno de estos, ubicados en la zona centro de la ciudad; y otros, en la zona oriente.

Figura 12. Mapa de calor de los casos de dengue grave en Cali entre 2015 - 2021.



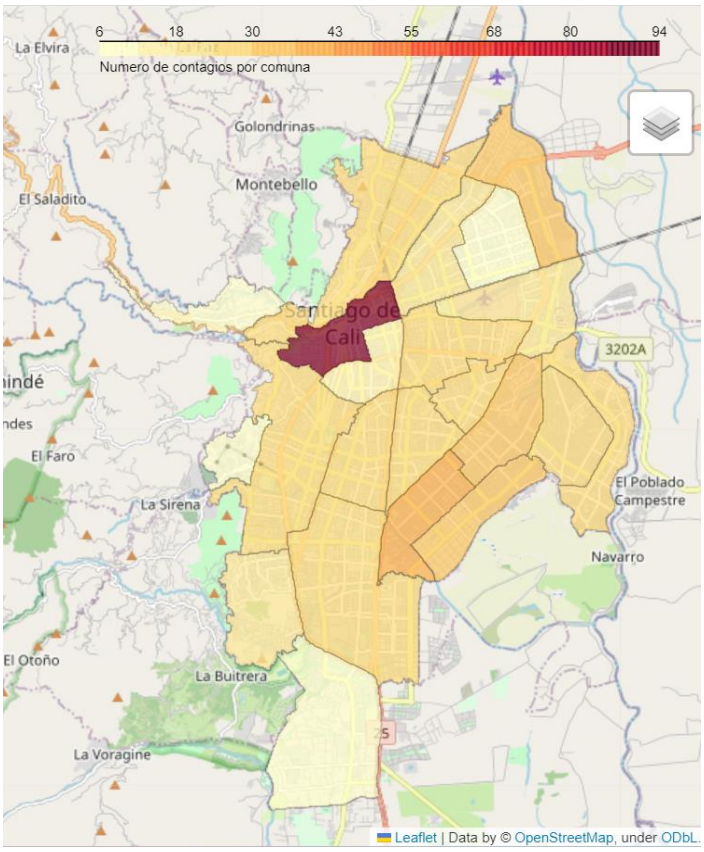
Fuente: Elaboración propia.

Dengue grave por comunas

La idea de que ciertos niveles de agregación de los datos podrían hacer emerger patrones y tendencias específicos, fue la motivación que impulsó el análisis de la cantidad de casos de dengue grave por comuna. A partir de esto, el planteamiento fue realizar una revisión en dos niveles: el primero, para determinar si había comunas que se destacaran por la concentración de casos acumulados de dengue grave en todo el periodo de análisis; el segundo, para establecer si hubo algún patrón destacado a lo largo de los siete años comprendidos entre 2015 y 2021.

Cuando se logró agregar los datos por comuna, se analizaron los resultados acumulados en un mapa de Cali, en el cual la mayor concentración de casos de la enfermedad aparece en color rojo oscuro, como se presenta en la Figura 13. Un hallazgo sorpresivo fue encontrar que la comuna 3 sobresale en solitario entre todas las demás, porque alcanzó los 93 casos en el periodo de análisis, cuando la segunda comuna sólo llegó a 36 casos en el mismo lapso.

Figura 13. Mapa coroplético - Cantidad acumulada de casos de dengue grave por comuna de Cali en el periodo 2015 - 2021.



Fuente: Elaboración propia.

Siguiendo el planteamiento de González et. al. [17], se decidió aislar el ruido de la variable "cantidad de casos de dengue grave", realizando un ajuste que corresponde al cálculo de la tasa de incidencia de dengue grave por cada 100.000 habitantes, en función de la población total de la comuna de procedencia de las personas con la enfermedad.

Dicho análisis condujo a otro hallazgo relevante: la tasa de incidencia anual del dengue grave de la comuna 3 fue consistentemente la más alta de la ciudad en cada uno de los siete años que comprende el estudio, con valores que oscilaron entre 5,21 y 13,84 veces más altos que el promedio de las comunas de Cali. La Tabla 2 presenta los detalles.

Tabla 2. Tasa de incidencia del dengue grave por cada 100.000 habitantes por comuna y año.

Comuna	2015	2016	2017	2018	2019	2020	2021
1	1,13	5,01	0,00	0,00	1,67	1,65	-
2	3,49	6,18	0,00	0,00	0,85	8,45	5,04
3	38,79	95,79	8,93	9,10	15,05	71,71	20,80
4	3,75	7,26	1,82	1,82	5,41	8,95	3,56
5	0,00	2,72	0,00	0,00	1,76	0,88	0,87
6	2,11	3,91	1,32	1,32	1,31	9,78	1,95
7	1,40	7,12	0,00	2,87	1,43	8,49	4,22
8	2,93	5,10	1,02	0,00	7,07	10,03	-
9	2,22	13,16	2,70	0,00	5,45	21,65	-
10	4,51	4,87	0,98	0,97	4,83	6,72	2,86
11	2,79	6,53	0,00	0,00	1,82	9,04	1,80
12	4,49	10,27	1,47	1,46	4,34	10,06	2,86
13	2,81	4,11	0,00	0,00	0,70	11,09	4,14
14	1,74	1,95	1,30	0,00	1,93	7,02	4,44
15	4,39	5,64	0,81	0,81	3,20	7,14	2,37
16	5,60	6,22	1,04	1,03	3,07	16,24	4,04
17	3,58	4,54	0,00	0,00	4,19	2,38	0,59
18	3,80	4,54	0,90	0,00	5,26	5,22	-
19	3,54	4,65	0,93	1,83	0,91	9,03	1,80
20	5,77	0,00	1,73	0,00	0,00	3,42	5,10
21	6,23	4,92	0,00	0,00	2,29	3,79	4,53
22	0,00	23,36	0,00	0,00	0,00	3,85	-
Promedio	**3,95**	**6,65**	**0,81**	**0,66**	**2,89**	**8,51**	**2,81**

Fuente: Elaboración propia.

Relación entre el estrato moda de las comunas y casos de dengue grave

Una vez establecida la pertinencia de agregar los casos de dengue grave por comuna, surgió la inquietud de si hubiese alguna relación significativa entre la cantidad de infecciones de una comuna y su estrato socioeconómico moda. Por tratarse de dos variables categóricas, se realizó una prueba Chi-Cuadrado de independencia para analizar su relación.

Se formularon las hipótesis de la prueba, así:

H0 = Las variables son independientes.

H1 = Las variables no son independientes.

Luego, se realizó la construcción de la tabla de contingencias y se calcularon las frecuencias esperadas para cada combinación de factores, como se observa en la Tabla 3.

Tabla 3. Frecuencias observadas y esperadas de los casos de dengue grave por estrato moda y comuna.

Frecuencias observadas:

Comuna	Estrato						Total
	1	2	3	4	5	6	
1	6						**6**
2					28		**28**
3			93				**93**
4		18					**18**
5			7				**7**
6		34					**34**
7			18				**18**
8			26				**26**
9			17				**17**
10			27				**27**
11			24				**24**
12			24				**24**
13		34					**34**
14	29						**29**
15	32						**32**
16		37					**37**
17					24		**24**
18			23				**23**
19				25			**25**
20	10						**10**
21	27						**27**
22						6	**6**
Total	**104**	**123**	**259**	**25**	**52**	**6**	**569**

Frecuencias esperadas:

Comuna	Estrato						Total
	1	2	3	4	5	6	
1	1,1	1,3	2,73	0,26	0,55	0,06	**6**
2	5,12	6,05	12,7	1,23	2,56	0,3	**28**
3	17	20,1	42,3	4,09	8,5	0,98	**93**
4	3,29	3,89	8,19	0,79	1,64	0,19	**18**
5	1,28	1,51	3,19	0,31	0,64	0,07	**7**
6	6,21	7,35	15,5	1,49	3,11	0,36	**34**
7	3,29	3,89	8,19	0,79	1,64	0,19	**18**
8	4,75	5,62	11,8	1,14	2,38	0,27	**26**
9	3,11	3,67	7,74	0,75	1,55	0,18	**17**
10	4,93	5,84	12,3	1,19	2,47	0,28	**27**
11	4,39	5,19	10,9	1,05	2,19	0,25	**24**
12	4,39	5,19	10,9	1,05	2,19	0,25	**24**
13	6,21	7,35	15,5	1,49	3,11	0,36	**34**
14	5,3	6,27	13,2	1,27	2,65	0,31	**29**
15	5,85	6,92	14,6	1,41	2,92	0,34	**32**
16	6,76	8	16,8	1,63	3,38	0,39	**37**
17	4,39	5,19	10,9	1,05	2,19	0,25	**24**
18	4,2	4,97	10,5	1,01	2,1	0,24	**23**
19	4,57	5,4	11,4	1,1	2,28	0,26	**25**
20	1,83	2,16	4,55	0,44	0,91	0,11	**10**
21	4,93	5,84	12,3	1,19	2,47	0,28	**27**
22	1,1	1,3	2,73	0,26	0,55	0,06	**6**
Total	**104**	**123**	**259**	**25**	**52**	**6**	**569**

Fuente: Elaboración propia.

El procedimiento continuó con la determinación de los valores de Chi-Cuadrado Calculado y Chi-Cuadrado Crítico con un nivel de significancia del 95%. Los resultados fueron 2.845 y 130, respectivamente. Como el valor de Chi-Cuadrado Calculado es mayor que el Chi-Cuadrado Crítico, se rechaza la hipótesis nula (H0) y, por tanto, se concluyó que existe una relación entre el estrato moda de las comunas y la cantidad de casos de dengue grave por comuna.

Resumen de los resultados del análisis descriptivo

La primera parte de la preparación de los datos proporcionó un listado de *insights* clave sobre la dinámica del dengue grave en Cali, los cuales incluyen en forma breve:

- La mayor incidencia del dengue grave ocurre en los primeros meses del año.
- Se trata de un fenómeno complejo, que no se comporta como una serie estacional.
- La proporción de casos de dengue grave sobre casos de dengue tiende a crecer.
- Se podría hablar de un brote de dengue grave a partir del umbral de los 13 casos.
- Tres instituciones de salud atendieron poco más de la mitad de los casos históricos.
- Los casos de dengue grave se concentran en las personas más jóvenes.
- La comuna 3 es el principal aportante de casos de dengue grave cada año.
- El estrato moda de las comunas está relacionado con la incidencia del dengue grave.

4.2. SEGUNDA PARTE: PREPARACIÓN DE LOS INSUMOS PARA LA MODELACIÓN

La segunda parte de la preparación de datos incluye principalmente: la selección preliminar de las variables de interés, el análisis exploratorio de datos, el uso del análisis de rezagos y la definición final del conjunto de datos para la etapa de modelación.

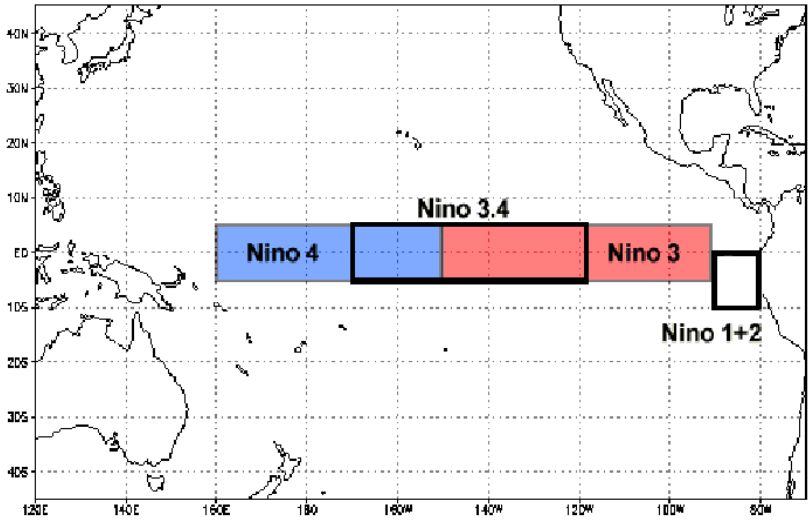
Selección de variables de interés

Esta selección tenía el desafío de identificar las variables a partir de las cuales se podría entrenar un modelo de *Machine Learning* que fuera capaz de predecir los casos de dengue grave en Cali, en un mes determinado. A continuación, se presentan las variables, un argumento de escogencia y la fuente de los datos de cada una.

1. Variables meteorológicas clásicas: precipitación, humedad y temperatura. Esta selección se basa en los hallazgos de trabajos previos, los cuales señalan consistentemente su relevancia para predecir el dengue [15]. Un detalle adicional que vale la pena mencionar, es que se encontraron diferentes opciones para cada variable, por ejemplo, temperatura sobre la superficie de la tierra o temperatura a dos metros de la superficie de la tierra, entre otras. De modo que esta selección requirió evaluar las opciones de cada variable, y elegir la mejor, en función de su relación con la cantidad de casos de dengue grave. La fuente datos es un repositorio de acceso libre publicado por la NASA (*National Aeronautics and Space Administration)* [21], en el que se dispone de los registros del satélite MERRA-2 (*Modern-Era Retrospective Analysis for Research and Applications*), los cuales son administrado por *Goddard Earth Sciences e Information Services Cente*r.

2. Índice del Niño-3.4. La escogencia de esta variable se basa en la relación que se ha evidenciado en la literatura entre el fenómeno de El Niño y los brotes de dengue en diferentes países tropicales, incluyendo Colombia [22]. El índice de Niño-3.4 corresponde al registro mensual de anomalías de la temperatura superficial del Océano Pacífico, la cual se mide en la región que se indica en la Figura 14. La fuente de datos son los reportes del Centro de Predicción del Clima, que es una entidad asociada a la Administración Nacional Atmosférica y Oceánica del Departamento de Comercio de los Estados Unidos [23].

Figura 14. Región de El Niño 3.4 sobre el Pacífico ecuatorial.



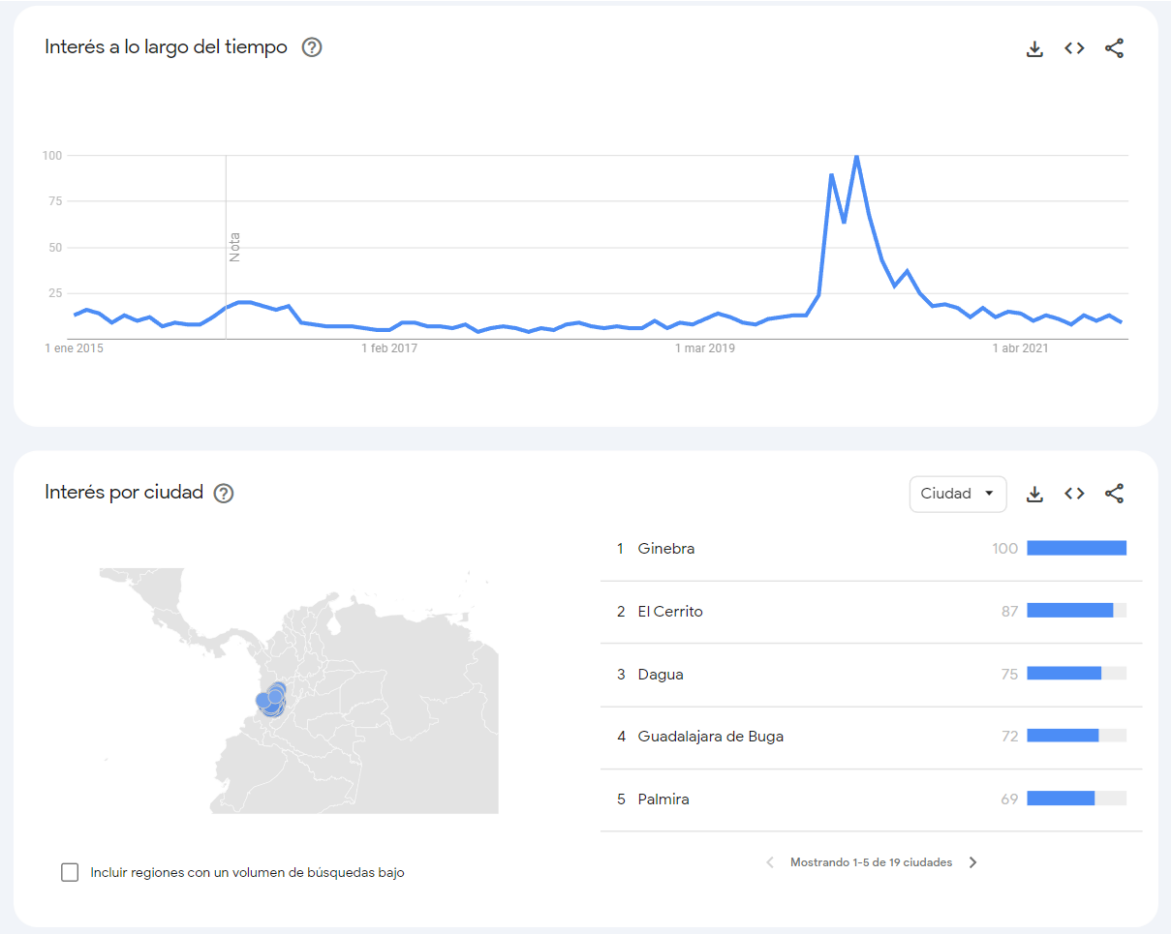
Fuente: Sánchez [24].

3. Índice de búsquedas web. La selección de esta variable novedosa es parte de las recomendaciones de trabajos previos en los que se reconoce su potencial para complementar los análisis predictivos relacionados con los brotes de dengue [15]. Se trata de una medición del interés en un término de búsqueda en Google a lo largo del tiempo, y se basa en una escala relativa, de tal forma que entre más alto el valor, más frecuentemente se han realizado consultas de ese término [25]. La fuente de la información es *Google Trends*.

   Es importante mencionar que *Google Trends* es una herramienta que nos permite conocer la frecuencia con la que los usuarios realizan diversas consultas en Google; por tanto, puede ser usada, por ejemplo, para comparar la popularidad de diferentes términos, seguir las tendencias a lo largo del tiempo y ver en qué regiones se realizan estas búsquedas.

   Para el desarrollo del presente proyecto, se recopilaron los índices de búsqueda web de acuerdo con *Google Trends* para el término "dengue" en el Valle del Cauca, cuyo comportamiento se puede observar en la Figura 15. Cabe resaltar que, a nivel local, no se encontró ningún antecedente de uso de esta variable; con todo, la idea subyacente que justifica su inclusión es que estos flujos de datos están basados en el comportamiento de las personas, y, por tanto, podrían reflejar patrones y tendencias en tiempo real.

Figura 15. Búsquedas del término "Dengue" en el Valle del Cauca en el periodo 2015 - 2021.



Fuente*: Google Trends* [25].

4. Casos de dengue grave. Una de las principales ideas detrás de esta escogencia es que la enfermedad del dengue tiene diferentes manifestaciones, que van desde los casos asintomáticos hasta los casos de dengue grave [3]. Estos últimos implican un alto riesgo para la vida de las personas, y típicamente demandan más atención y comprometen más recursos de la red de salud de la ciudad. La fuente de datos de esta variable fue la Secretaría de Salud Pública de Cali.

En suma, se seleccionaron seis variables de interés, incluyendo la variable a predecir, para incorporarlas en el proceso de análisis para determinar la conveniencia de usarlas dentro de la etapa de modelamiento. Las seis variables son cuantitativas y se dispuso de ellas en forma de datos estructurados.

Primeras transformaciones

El proceso de recolección de los datos permitió identificar que los registros de todas las variables no tenían el mismo nivel de granularidad, y, por tanto, evidenció la necesidad de realizar las primeras transformaciones. Por ejemplo, la cantidad de casos de dengue grave tenían registros diarios, pero el índice del Niño-3.4 tenían registros mensuales.

Se probó con distintas alternativas en busca del nivel de agregación más conveniente, valorando las implicaciones en términos de la relación entre las variables de interés llamadas a ser predictoras y la cantidad de casos de dengue grave. Algunas de las opciones de agregación evaluadas incluyeron: semanas calendario, semanas epidemiológicas, periodos epidemiológicos, y meses. El resultado de los análisis fue que la opción más conveniente era agregar los datos diarios a registros mensuales.

En el caso de las variables meteorológicas clásicas, se realizó la agregación utilizando el valor promedio de los datos para representar su comportamiento general.

Análisis exploratorio de datos - Primera parte

A partir de lo anterior, se consolidó un conjunto de datos de trabajo que permitió avanzar con los siguientes pasos del proceso de preparación, el primero de los cuales fue un análisis exploratorio para tener una mejor comprensión de las variables seleccionadas. Este análisis se realizó en dos partes, separadas por una aplicación de transformaciones necesarias.
Para comenzar, en términos de dimensionalidad, el conjunto de datos está compuesto por seis variables, y 84 registros, sin duplicados ni valores faltantes. Los registros corresponden a los valores mensuales de cada variable, entre enero del 2015 y diciembre del 2021. La exploración univariada inició por la caracterización general de las variables de interés, tal como se presenta en la Tabla 4.

Tabla 4. Caracterización general de las variables de interés.

Variable	Descripción	Unidad de medida	Tipo de variable	Tipo de escala	Tipo
humidity	Promedio mensual de la humedad relativa a 2 metros de la superficie de la tierra.	%	Cuantitativa continua	Intervalo	float64
rainfall	Promedio mensual de la precipitación en la superficie de la tierra.	mm	Cuantitativa continua	Razón	float64
tmax	Promedio mensual de la temperatura máxima del aire a 2 metros de la superficie.	°C	Cuantitativa continua	Intervalo	float64
searches	Índice mensual de búsquedas web en Google.	%	Cuantitativa continua	Razón	int64
nino	Índice del Niño 3.4.	°C	Cuantitativa continua	Intervalo	float64
c220	Cantidad mensual de casos de dengue grave.	Casos	Cuantitativa discreta	Razón	int64

Fuente: Elaboración propia.

Luego, se realizó una descripción de cada una de las variables del conjunto de datos, mediante las estadísticas que resumen la tendencia central y dispersión y la presencia de valores faltantes, como se presenta en la Tabla 5.

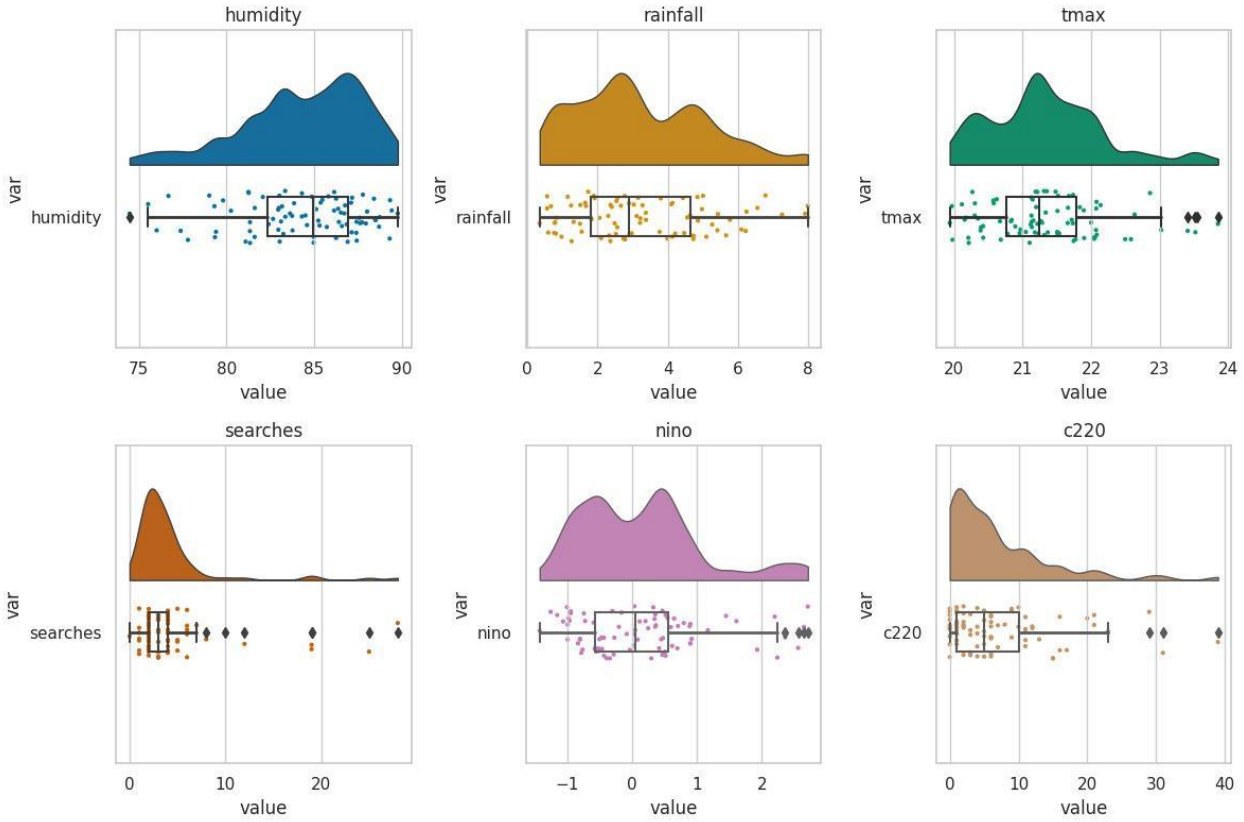
Tabla 5. Estadísticas descriptivas de las variables de interés.

Descripción	humidity	rainfall	tmax	searches	nino	c220
count	84,00	84,00	84,00	84,00	84,00	84,00
mean	84,27	3,24	21,35	4,17	0,14	6,82
std	3,50	1,86	0,86	4,68	0,95	7,60
min	74,47	0,35	19,93	0,00	-1,42	0,00
25%	82,32	1,80	20,75	2,00	-0,56	1,00
50%	84,92	2,90	21,25	3,00	0,05	5,00
75%	86,96	4,65	21,79	4,00	0,56	10,00
max	89,77	7,99	23,85	28,00	2,71	39,00

Fuente: Elaboración propia.

A continuación, se complementa el análisis de distribución de datos, utilizando gráficos de densidad y *Boxplots* como se observa en la Figura 16.

Figura 16. Gráficos de Densidad y *Boxplot* para la distribución de datos.



Fuente: Elaboración propia.

En este punto del análisis, se destacan los siguientes hallazgos relevantes:

- Hay diferencias importantes en la escala de los datos.
- Todas las variables tienen distribuciones sesgadas.
- La variable “Niño Oceánico 3.4” (nino) toma valores negativos.
- La variable “casos de dengue grave” (c220) tiene 11,9% de sus registros con valor cero.
- Exceptuando c220 y searches, la mayoría de los registros de las variables son valores únicos.

Transformaciones complementarias

Algunos de los hallazgos de la exploración univariada, y la expectativa de uso de algunos métodos de *Machine Learning* y determinadas métricas de evaluación, condujeron a utilizar dos procedimientos de transformación adicionales.

El primer procedimiento consistió en escalar los valores para estandarizar los rangos y, así, mejorar la estabilidad numérica e interpretabilidad de ciertos cálculos necesarios [26], además de evitar el riesgo de sesgo en los algoritmos más susceptibles a este tipo de fenómenos, al cambiar la representación de los datos sin perder información [27].

En este punto, vale la pena destacar que, como parte del proceso de desarrollo iterativo, la primera aproximación para elegir el método de escalamiento a emplear condujo a un hallazgo empírico relevante, el cual fue confirmado de manera posterior mediante una revisión de literatura: los métodos de escalamiento usados tienen efectos sobre el desempeño de los modelos [28].

El segundo procedimiento que se utilizó fue aplicar una transformación *Yeo-Johnson* para simplificar la distribución de los datos [29], pues aun cuando algunos algoritmos no exijan cumplir los supuestos de homogeneidad, aditividad, y normalidad de los errores, la calidad y precisión de las estimaciones podrían verse beneficiadas de este tipo de transformaciones.

Análisis exploratorio de datos - Segunda parte

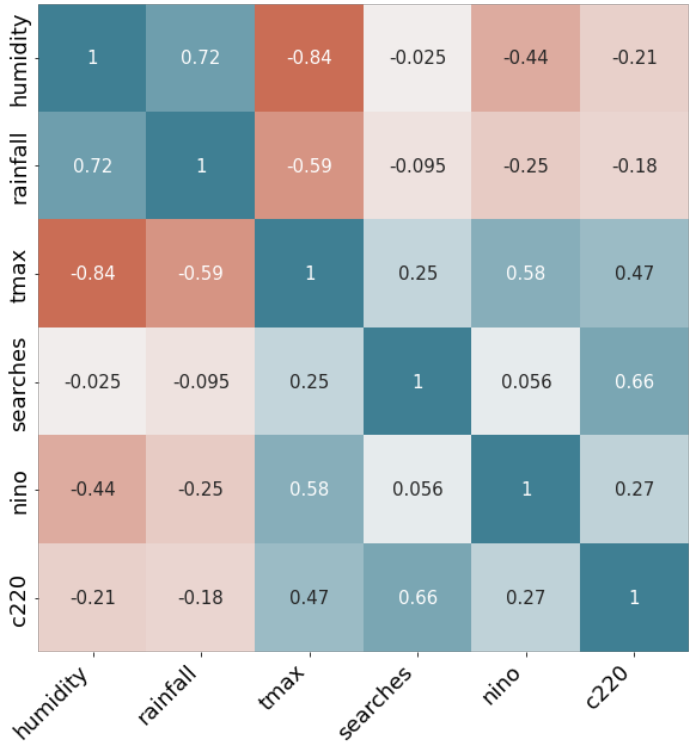
La segunda parte del análisis exploratorio consistió en estudiar la relación entre las variables transformadas, primero a partir de los Coeficientes de Correlación de *Pearson*; y luego, a partir de un análisis del factor de inflación de la varianza.

En primer lugar, la inspección de la relación de las variables a través del Coeficiente de Correlación de Pearson tuvo dos resultados importantes. Por un lado, las variables que tienen la relación lineal más fuerte con c220 son: el índice de búsquedas web y la temperatura máxima, tal como se evidencia en la Figura 17. Esto es relevante porque centra la atención en la más novedosa de las variables seleccionadas. Con todo, este resultado no va en detrimento de la importancia de las otras variables, las cuales, como se indicó anteriormente, tienen un lugar destacado en la literatura. Además, al tratarse de un fenómeno complejo, no se espera que todas sus relaciones sean de tipo lineal.

Por otro lado, se comprobó que algunas de las variables llamadas a ser predictoras están fuertemente relacionadas entre sí, como se aprecia en la Figura 17. Por ejemplo, las variables

humedad y precipitaciones tienen una correlación positiva fuerte; y las variables humedad y temperatura máxima tienen una correlación negativa fuerte. Este hallazgo es relevante porque eventualmente se configuraría un fenómeno de multicolinealidad. Por lo tanto, se evidenció la necesidad de revisar, por un lado, la selección final de variables, y por el otro, la selección de algoritmos, de modo que sean lo menos sensible posible a ese tipo de problemas.

Figura 17. Mapa de calor de la matriz de correlaciones.

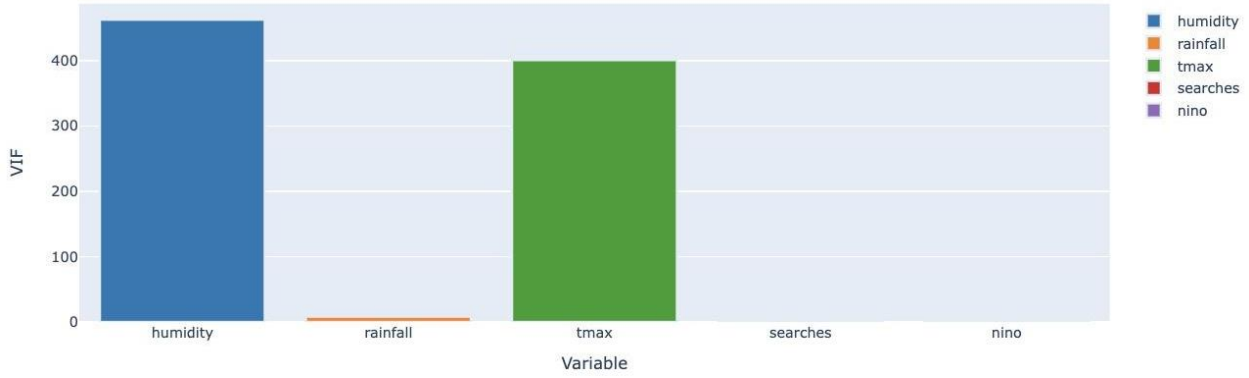


Fuente: Elaboración propia.

En segundo lugar, se realizó un análisis de inflación de la varianza (VIF) para complementar el diagnóstico realizado y analizar alguna alternativa de solución.

Al aplicar el procedimiento necesario, se evidenció que las variables “humedad” y “temperatura máxima” tienen valores del factor de inflación de la varianza que son excesivamente altos en comparación con las otras variables, como se observa en la Figura 18. Esto confirmó la conveniencia de evaluar las variables para eliminar alguna de ellas en busca de reducir la multicolinealidad.

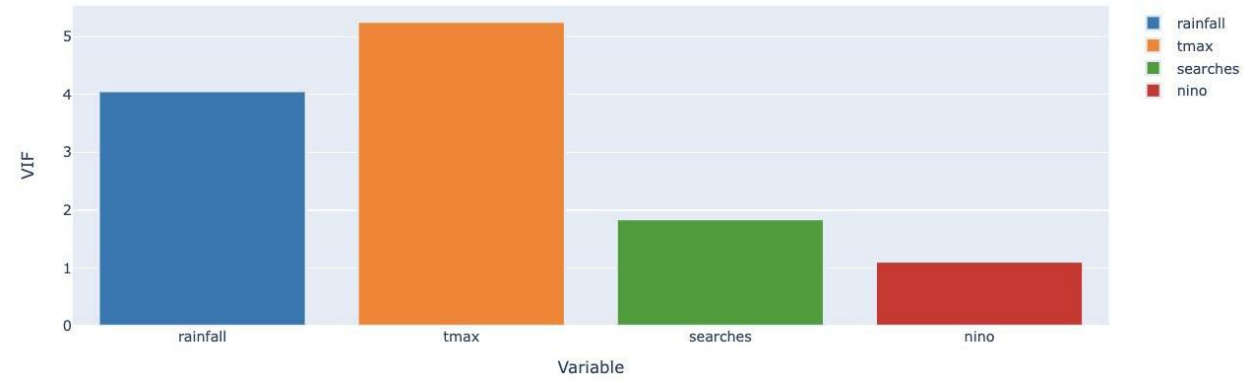
Figura 18. Diagrama de barras - Factor de inflación de la varianza de todas las variables.



Fuente: Elaboración propia.

A continuación, se realizó un experimento excluyendo la variable "humedad", para aplicar nuevamente el procedimiento de análisis del VIF. Los resultados de las variables incluidas fueron valores mucho más homogéneos, como se observa en la Figura 19.

Figura 19. Diagrama de barras - Factor de inflación de la varianza sin la variable humedad.



Fuente: Elaboración propia.

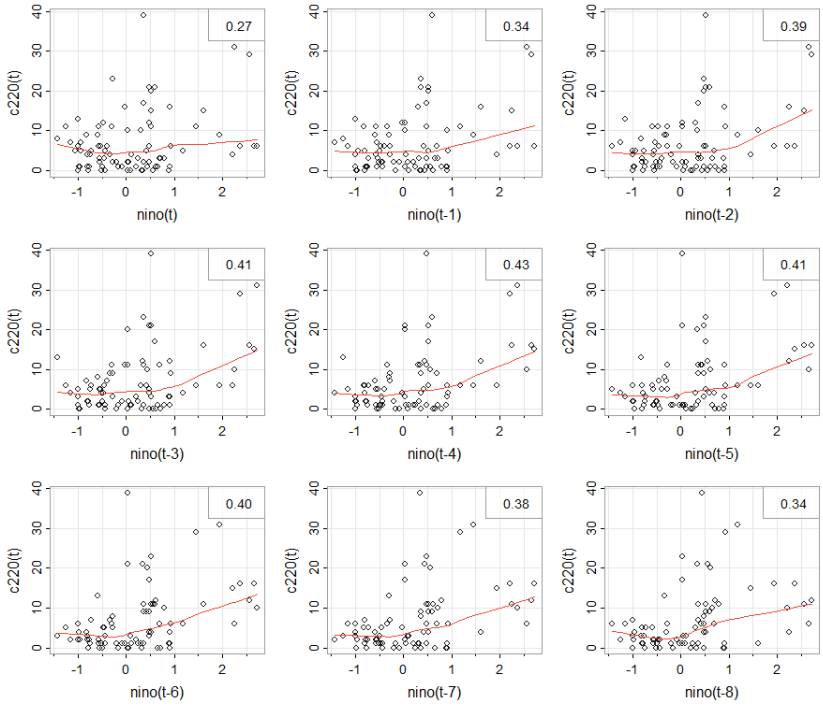
Por tanto, se decidió eliminar la variable "humedad" del conjunto de datos, para evitar posibles sesgos y errores en los análisis posteriores. Esta depuración condujo a tener un conjunto de datos final que representa un insumo de mejor calidad para la etapa de modelación.

Análisis de rezagos

Considerando la propuesta de algunos trabajos previos, se realizó un análisis de rezagos de las variables llamadas a ser predictoras, como aplicación de la ingeniería de variables. La intención fue identificar los efectos de generar desplazamientos temporales de cada una de las variables predictoras, sobre su relación con la variable c220, para confirmar o ajustar las expectativas relacionadas a su nivel de relevancia, desde la perspectiva de las relaciones lineales.

A partir de lo anterior, se fijó un rango de 8 meses de desplazamientos, y se inspeccionó el Coeficiente de Correlación de Pearson en cada desplazamiento, como se puede ver en la Figura 20. Las gráficas de cada una de las variables restantes están disponibles en el Anexo 1.

Figura 20. Diagramas de Dispersión - *Lags*: Índice del Niño 3.4 VS c220.



Fuente: Elaboración propia.

El resultado clave de este análisis es que validó la hipótesis de que se podrían evidenciar relaciones lineales más fuertes entre los registros pasados de las variables llamadas a ser predictoras, y el número de casos de dengue grave de un determinado periodo. Los resultados están resumidos en la Tabla 6, que incluye la configuración de la mejor relación entre variables predictivas y variable de respuesta.

Tabla 6. Relación entre variables predictoras y variable de respuesta - Análisis de Rezagos.

Variable	Coef. Correlación Original	Rezago Más Conveniente	Coef. Correlación con Rezago Más Conveniente	Resultado
humidity	-0,21	6	-0,43	Mejoró
rainfall	-0,18	7	-0,28	Mejoró
tmax	0,47	1	0,48	Mejoró
searches	0,66	0	0,66	Se mantuvo
nino	0,27	4	0,43	Mejoró

Fuente: Elaboración propia.

El último paso del análisis de rezagos fue consolidar un conjunto de datos complementario, a partir de una intersección temporal que reconozca la menor cantidad de registros de las variables rezagadas. Este conjunto de datos complementario está diseñado para comprobar si tiene la capacidad de mejorar el desempeño del modelo de *Machine Learning*.

5. MODELADO

Esta sección se enfoca en el desarrollo y la evaluación del modelo predictivo que permitió completar los objetivos del proyecto. Para esto, se presentan los detalles relacionados con la elección de las métricas de desempeño, la selección de algoritmos de *Machine Learning,* la definición de la estrategia de modelación y validación elegida para determinar el mejor modelo, y la evaluación de resultados de los modelos candidatos.

Elección de métricas

Con la intención de realizar una evaluación integral del desempeño de los modelos candidatos, se decidió utilizar un conjunto de métricas que se destacan en términos de interpretabilidad, complementariedad y robustez ante valores extremos, estas son: el error cuadrático medio (RMSE), el error absoluto medio (MAE) y el coeficiente de determinación (R2).

Una diferencia importante entre el MAE y el RMSE es que el primero se comporta de forma lineal, pues tratar el tamaño de los errores de forma indistinta, lo que lo hace relativamente intuitivo; mientras que el RMSE es más exigente, porque castiga más los errores grandes que los pequeños, y, por tanto, es un poco más complejo [30]. Por su parte, el coeficiente de determinación (R2) proporciona una noción de la bondad de ajuste y, en consecuencia, la probabilidad de que el modelo pueda predecir bien los registros no observados [10]. Con todo, las tres métricas ayudaron a tener una idea de la calidad de las predicciones de los modelos.

Elección de algoritmos

La elección de algoritmos se basó en los siguientes aspectos clave: las lecciones aprendidas de trabajos anteriores que tenían retos similares, la incorporación de un algoritmo clásico que permita comparar la evolución de las capacidades predictivas, la reciente popularidad de algunos algoritmos en las comunidades de ciencia de datos, y la capacidad de ciertos algoritmos para manejar bien relaciones no lineales.

A continuación, se presentan los algoritmos elegidos y alguna idea que soportó su escogencia:

- *Random Forest (RF).* Este algoritmo se basa en árboles de decisión, lo cual corresponde con las alternativas de mejor desempeño en proyectos previos.

- *LightGBM.* Este algoritmo, relativamente nuevo, se ha destacado por su velocidad y eficacia de entrenamiento. Esto resultaba conveniente para el enfoque iterativo previsto.

- *Support Vector Machine (SVM)*. A este algoritmo clásico se le reconoce su buen desempeño al tratar con conjuntos de datos relativamente pequeños, como en este proyecto.

- Redes Neuronales Artificiales (*ANN*). Se seleccionó este algoritmo de mayor complejidad, entre otras cosas, por el prestigio de su capacidad de generalización y por poder aprender patrones subyacentes complejos.

Estrategia de modelación y validación

Para soportar la determinación objetiva del mejor modelo, se definió una estrategia de modelación y validación basada en la división del conjunto de datos en tres particiones: entrenamiento, prueba y validación.

La implementación de dicha estrategia se basó en la técnica de validación cruzada *k-fold,* y el uso de la siguiente rutina:

a) Dividir el conjunto original de datos en dos partes: un subconjunto del 90% de los registros para los procesos de entrenamiento-prueba; y otro del 10%, para la validación.
b) Dividir el subconjunto de datos de entrenamiento-prueba en once grupos iguales. Cada grupo a su vez está dividido en 90% de registros para entrenamiento y 10% para prueba, habilitando una corrida de entrenamiento-prueba por cada grupo.
c) Calcular las métricas con los resultados del uso de los datos de prueba en cada corrida.
d) Realizar ajustes a los modelos en función del valor promedio de cada métrica.
e) Determinar el modelo de mejor desempeño.
f) Utilizar el modelo ganador con los datos de validación para confirmar su desempeño.

A manera de práctica complementaria para evaluar el desempeño de los modelos, se utilizó una regresión lineal como un tipo de “modelo ingenuo”, a fin de establecer las líneas base de comparación del desempeño de los modelos candidatos. Así, reconociendo las características particulares del conjunto de datos, se determinó con mayor propiedad la dimensión de la mejora obtenida por el uso de los otros modelos.

Otra dinámica utilizada en el proceso de desarrollo iterativo que vale la pena resaltar es que se evaluó cada uno de los modelos a través del coeficiente de determinación en el conjunto de datos de entrenamiento y en el conjunto de datos de prueba, para monitorear el sobreajuste de los modelos. Si el coeficiente era significativamente más alto en los datos de entrenamiento que en los de prueba, se consideraba que era necesario realizar modificaciones al modelo para corregir el problema de sobreajuste.

Definición de candidatos

Se decidió que los modelos candidatos corresponderían a la aplicación de los algoritmos seleccionados sobre los dos conjuntos de datos creados: los datos en estado original, y los datos que tienen rezagos temporales. En consecuencia, se evaluaron en total diez modelos candidatos, incluyendo los modelos ingenuos.

A continuación se presenta la configuración elegida para el uso de cada algoritmo avanzado. La definición de los hiperparámetros se basó en un proceso de experimentación iterativa ejecutado para optimizar el rendimiento de los modelos candidatos.

Tabla 7. Configuración de los algoritmos.

Algoritmo	Hiperparámetros	Descripción
RF	n_estimators: 100 criterion: 'squared_error' max_depth: None max_features: 'auto' oob_score: True n_jobs: -1 random_state: 0	Se utilizaron 100 árboles. Se empleó el error cuadrático medio (MSE) como medida de calidad de las divisiones, y se permitió que los árboles crecieran hasta su máxima profundidad. Además, se consideraron todas las características en cada división. Se calculó el error fuera de la bolsa para evaluar el rendimiento y se aprovechó el uso de todos los núcleos del procesador. Por último, se estableció una semilla aleatoria para garantizar la reproducibilidad de los resultados.
LightGBM	boosting_type: 'gbdt' objective: 'regression' random_state: 0 colsample_bytree: 1 learning_rate: 0.1 n_estimators: 100	Se utilizó el boosting type 'gbdt', que es una implementación de Gradient Boosting, con el objetivo de realizar una regresión. Además, se estableció una semilla aleatoria para garantizar la reproducibilidad de los resultados. Se utilizó un colsample_bytree de 1, lo que significa que se consideran todas las características en cada árbol. Se seleccionó una tasa de aprendizaje de 0.1 y se entrenaron 100 estimadores en el modelo.
SVM	C: 1 gamma: 'scale' kernel: 'rbf' epsilon: 0.1	Se eligió C=1 para obtener un enfoque equilibrado entre el ajuste a los datos de entrenamiento y la capacidad de generalización. También se utilizó la configuración ‘scale’ para gamma a fin que se adaptara automáticamente según la escala de los datos de entrada. Se utilizó el kernel ‘rbf’ pensando en su capacidad para modelar relaciones no lineales en la regresión. Finalmente, se estableció que el nivel de tolerancia en los errores de regresión debería ser 0.1.

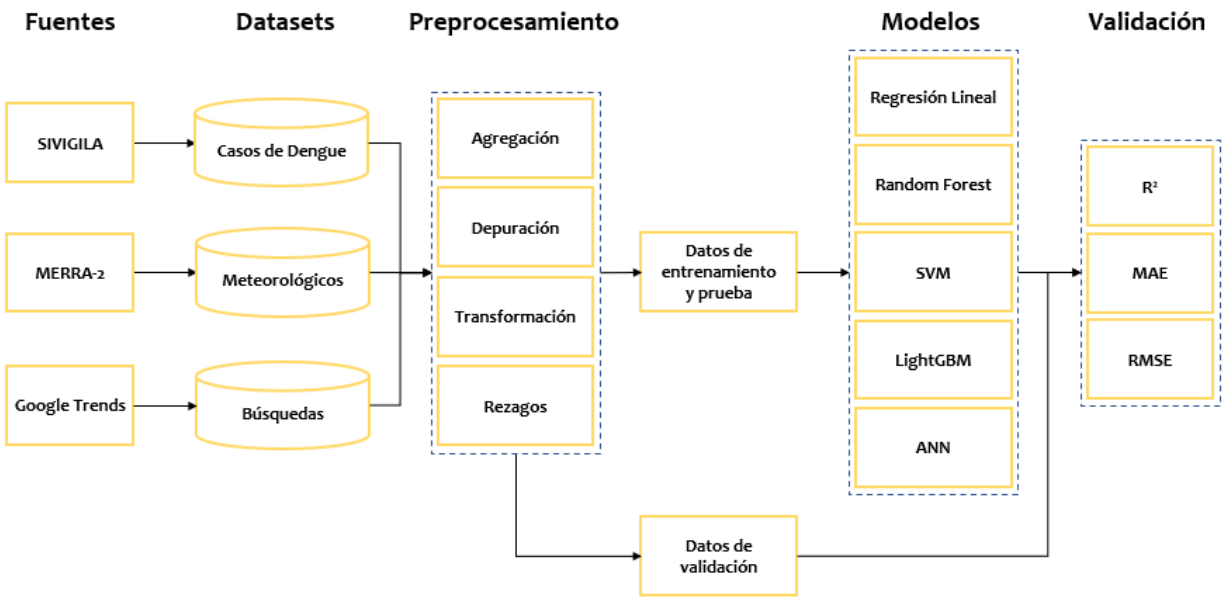
ANN	units: 32 activation: 'relu' input_shape: (7,) dropout(0.05) units: 64 activation: 'linear' dropout(0.1) units: 64 activation: 'relu' dropout(0.1) units: 64 activation: 'linear' dropout(0.1) units: 1 activation: 'linear' dropout(0.05) loss: 'mae' optimizer: 'adam'	La red neuronal artificial tiene las siguientes características: La red se compone de varias capas densas o fully connected. La primera capa tiene 32 unidades y utiliza la función de activación ReLU. Además, recibe datos de entrada con una forma de (7,). Se añadió una capa de Dropout con una tasa de 0.05 después de la primera capa para ayudar a regularizar la red y evitar el sobreajuste. La segunda capa tiene 64 unidades y utiliza la función de activación lineal. Se añadió otra capa de Dropout con una tasa de 0.1 después de la segunda capa. La tercera capa tiene 64 unidades y utiliza la función de activación ReLU. Se añadió otra capa de Dropout con una tasa de 0.1 después de la tercera capa. La cuarta capa tiene 64 unidades y utiliza la función de activación lineal. Se añadió otra capa de Dropout con una tasa de 0.1 después de la cuarta capa. La última capa tiene una unidad y utiliza la función de activación lineal. Se añadió una última capa de Dropout con una tasa de 0.05. La función de pérdida utilizada para entrenar el modelo es el error medio absoluto (MAE). Se seleccionó el optimizador 'adam' para ajustar los pesos de la red durante el entrenamiento.

Fuente: Elaboración propia.

Flujo de trabajo general

Llegados a este punto, se consideró conveniente contar con un resumen del flujo de trabajo general del proyecto, involucrando todas las etapas indicadas hasta ahora, como aparece en la Figura 21.

Figura 21. Resumen del flujo de trabajo general.



Fuente: Elaboración propia.

En lo que respecta al entorno de trabajo, se utilizó la plataforma *Google Colab* para la implementación del proyecto, por contar con un entorno integrado que permite ejecutar código en línea y facilitar el trabajo colaborativo. Se enriqueció el proceso de escritura del código con algunos recursos disponibles en dicha plataforma, incluyendo bibliotecas preinstaladas, lo cual simplificó el proceso de configuración del entorno de trabajo.

Comparación de los resultados

Al comparar los resultados obtenidos para cada uno de los modelos, se evidenció que el candidato de mejor desempeño es el modelo número 10, el cual corresponde al algoritmo Redes Neuronales aplicado al conjunto de datos de variables rezagadas, según se evidencia en la Tabla 7. Se trata del mejor modelo porque es la opción que minimiza los errores y maximiza la bondad de ajuste.

Tabla 8. Comparación del desempeño de los modelos candidatos.

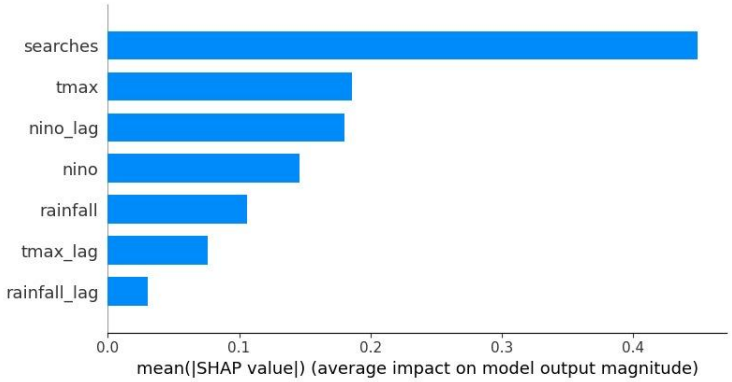
Modelo	*Dataset*	Algoritmo	R2 *train*	R2 test	MAE	RMSE
1	Original	Regresión Lineal	0.44	0.00	0.68	0.82
2	Original	RF	0.91	0.06	0.65	0.77
3	Original	SVM	0.75	0.07	0.71	0.81
4	Original	*LightGBM*	0.57	0.11	0.65	0.77
5	Original	ANN	0.98	0.67	0.38	0.47
6	Rezagado	Regresión Lineal	0.49	0.07	0.65	0.79
7	Rezagado	RF	0.92	0.19	0.58	0.73
8	Rezagado	SVM	0.78	0.19	0.60	0.73
9	Rezagado	*LightGBM*	0.69	0.23	0.58	0.71
10	Rezagado	ANN	0.99	0.72	0.33	0.45

También vale la pena resaltar que se confirmó la hipótesis de que el conjunto de datos complementario, el que incluye las variables que tienen rezagos temporales, efectivamente mejoró el desempeño de todos los algoritmos utilizados.

Evaluación de la importancia de las variables predictoras

Una vez identificado el modelo ganador, se analizó la relevancia de las variables predictoras. Para tales efectos, se utilizó el *Kernel Explainer* de SHAP (*SHapley Additive exPlanations*), que es una herramienta útil para entender cómo cada variable contribuye a las predicciones del modelo [31]. En este caso, se empleó la media absoluta de cada variable para todas las instancias del conjunto de prueba, como se observa en la Figura 22.

Figura 22. Diagrama de barras - Importancia de las variables predictoras del modelo ganador.



Fuente: Elaboración propia.

Lo anterior es muy relevante porque la variable predictora más importante para el modelo es el "índice de búsquedas web", el cual, como se ha indicado antes, es un recurso novedoso, especialmente a nivel local, para este tipo de iniciativas en ciencia de datos. Además, el resultado sobresale cuando se tiene en cuenta que el valor de "searches" es muy superior al de las otras variables, las cuales son bien conocidas por su uso y relevancia en este tipo de modelos predictivos.

Uso del modelo ganador con los datos de validación

A partir de lo anterior, se utilizó el modelo creado a partir del algoritmo ANN y las variables rezagadas para estimar la cantidad de casos de dengue del conjunto de datos de validación. En la Figura 23 se presentan los resultados en términos del valor estimado y el intervalo de confianza.

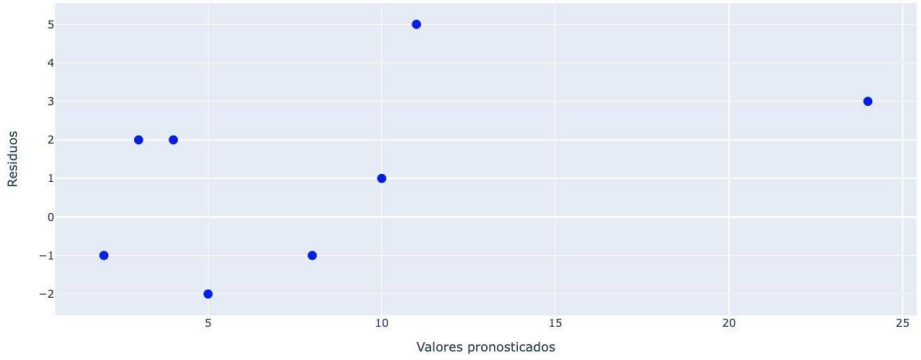
Figura 23. Valores estimados e intervalos de confianza versus valores reales del conjunto de validación.

Fuente: Elaboración propia.

Una vez aplicado el modelo ganador al conjunto de validación, se calculó el coeficiente de determinación, y su valor es 0,7843. Además, los valores reales estuvieron dentro de los rangos del intervalo de confianza, en cinco de los ocho casos. Es importante indicar que se trata de un resultado satisfactorio, especialmente cuando se consideran las limitaciones del conjunto de datos.

También se hizo una validación complementaria de la calidad del modelo ganador revisando sus errores. La Figura 24 presenta los residuos versus los valores ajustados para el conjunto de datos de validación. Como se puede evidenciar, los residuos están distribuidos aproximadamente en forma aleatoria alrededor del cero, no se evidencia algún tipo de patrón o tendencia inquietante, y no hay valores atípicos. Esta aproximación confirma que el modelo basado en ANN se ajusta bien a los datos, y que no hay indicios de algún problema en el modelo.

Figura 24. Diagrama de Dispersión - Residuos versus valores ajustados



Fuente: Elaboración propia.

6. CONCLUSIONES Y TRABAJOS FUTUROS

6.1. CONCLUSIONES

A través del análisis descriptivo de la dinámica del dengue grave en la ciudad de Cali, fue posible identificar picos en la ocurrencia de casos para los años 2016 y 2020, además de identificar que generalmente el primer cuatrimestre del año concentra la mayor cantidad de observaciones; por otro lado, se deja en evidencia que la proporción de casos grave frente a los casos de dengue leve ha presentado una clara tendencia al alza tras cada año; sumado a esto, se evidenció que el 51% de los casos de dengue grave estuvieron concentrados en 3 instituciones de Salud: La Clínica Infantil Club Noel, La Fundación Valle del Lili y el Hospital Universitario del Valle, indicando así que estas entidades deberán ser focalizadas por la Secretaría de Salud en sus programas de atención a la enfermedad del dengue.

En este proyecto se demostró que el uso de redes neuronales artificiales (ANN) es una herramienta eficaz para la predicción de la cantidad de casos de dengue grave en la Ciudad de Cali. Los resultados de la evaluación de los modelos indicaron que ANN tuvieron un mejor desempeño que los otros algoritmos empleados, a saber, *Random Forest*, SVM y *LightGBM*.

También se demostró que el ejercicio de la ingeniería de variables es clave en la creación de modelos predictivos para el dengue grave en Cali. La inclusión de rezagos temporales en las variables predictoras permitió un entrenamiento más eficaz y permitió que todos los algoritmos tuvieran mejores resultados. En lo que respecta al algoritmo ANN, se observó que la inclusión de registros de períodos anteriores de las variables predictoras permitió que el modelo captara mejor las tendencias y patrones históricos de la enfermedad, lo cual, a su vez, le dio la capacidad de hacer las mejores predicciones obtenidas. Estos resultados podrían tener implicaciones relevantes para las iniciativas de creación de modelos predictivos de enfermedades epidémicas, y sugiere la inclusión de este tipo de aproximaciones a la ingeniería de variables en la búsqueda de mejorar la efectividad de los modelos.

Adicionalmente se evidenció que el uso de técnicas de regularización en los modelos basados en ANN les permitió tener un desempeño destacado en una situación de cantidad de datos limitada, en contraste con otros algoritmos que vieron comprometida su capacidad para generalizar y experimentaron problemas de sobreajuste. Esto permitió validar que las características del conjunto de datos requieren un enfoque experimental y la evaluación cuidadosa de diferentes opciones para lograr un desempeño satisfactorio. En última instancia, se debe elegir el modelo y la técnica que mejor se adapte a las necesidades y objetivos del proyecto.

Finalmente, este proyecto demostró que el uso de datos informales, es decir, la que no está basada en fuentes oficiales, como la lluvia, el índice del niño-3.4, la temperatura, y especialmente las búsquedas web pueden ser útiles para la predicción de la cantidad de casos de dengue grave en Cali. Los resultados mostraron que la inclusión de esas variables en el proceso de construcción del modelo basado en ANN le permitió predecir la cantidad de casos de la enfermedad. El hallazgo de que las variables de búsquedas web resultó ser la más relevante para el modelo ganador sugiere que se trata de una variable valiosa para mejorar las capacidades del sistema de vigilancia y control del dengue grave y la predicción de brotes en la Ciudad de Cali.

6.2. TRABAJOS FUTUROS

Las investigaciones futuras podrían encontrar útil incorporar nuevas variables de interés para proponer modelos más robustos, reconociendo, entre otras cosas, diferentes enfoques utilizados en los procesos de vigilancia epidemiológica del dengue. Valdría la pena explorar variables como los datos procedentes de pruebas especializadas de laboratorio, de modo que se pueda determinar los serotipos del virus que estén generando brotes en determinadas áreas de la ciudad, además de poder refinar el reporte de la cantidad de infecciones reales, descartando enfermedades que producen cuadros sintomáticos similares al dengue [1].

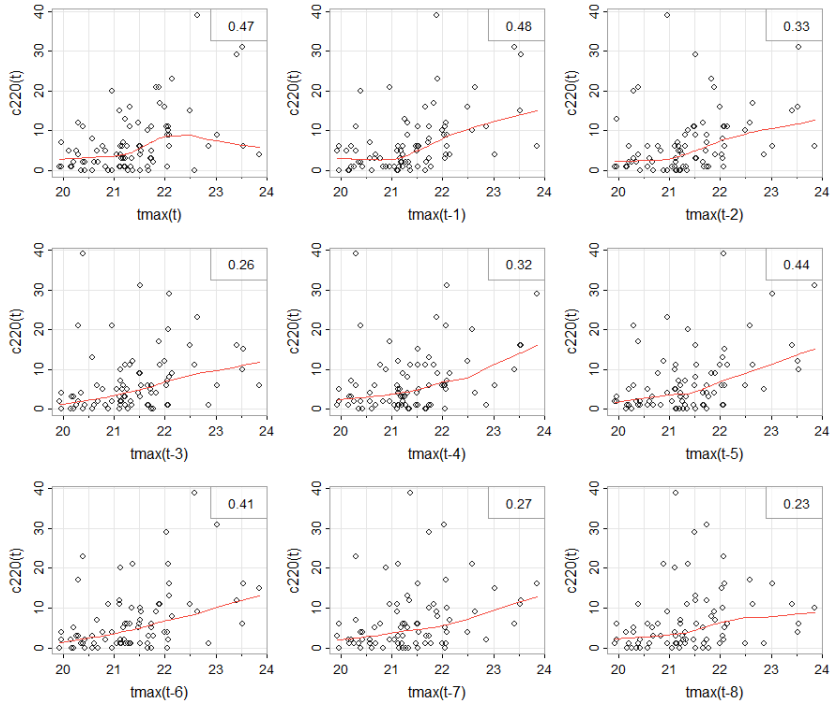
También sería útil explorar los registros de las actividades clásicas enfocadas en el vector, como los muestreos entomológicos de *Aedes aegypti*, por tratarse de un tipo de inductor directo con potencial para mejorar la efectividad de los modelos en la prevención de brotes.

En esa misma línea, pero desde un recurso relativamente novedoso, valdría la pena involucrar información relacionada con el *World Mosquito Program*. Se trata de un programa internacional que utiliza mosquitos portadores de una bacteria para prevenir la propagación del dengue [32]. La efectividad de dicho programa podría cambiar, en mayor o menor medida, las dinámicas históricas del dengue y, por tanto, convertirlo en una variable susceptible de inclusión en los procesos de modelación.

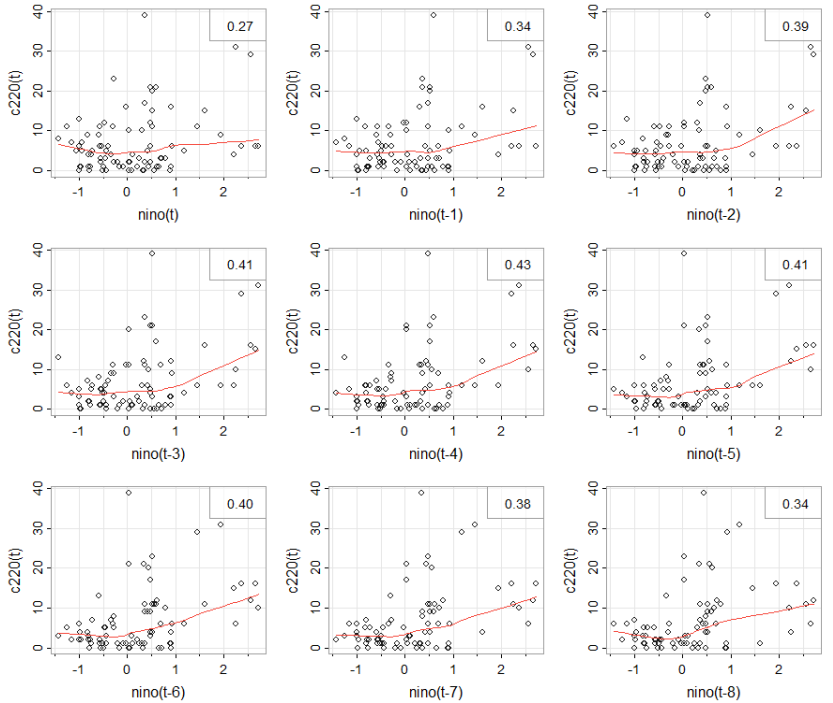
7. ANEXOS

Anexo 1: *Lags* para cada variable.

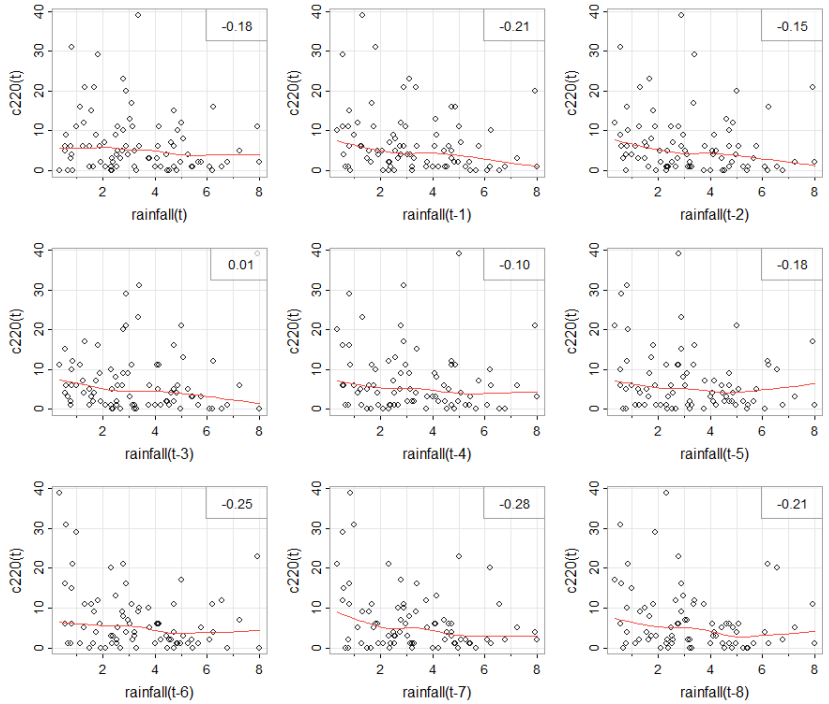
***Lags*: Temperatura Máxima VS c220**



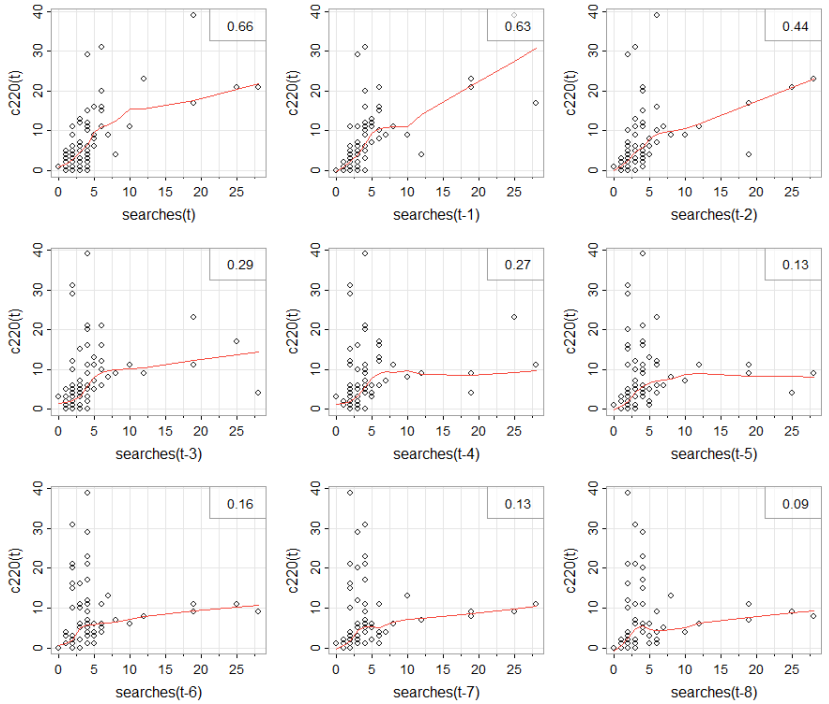
***Lags*: Índice del Niño 3.4 VS c220**



Lags: Precipitaciones VS c220



Lags: Búsquedas Web VS c220



8. REFERENCIAS BIBLIOGRÁFICAS

[1] G. T. R. H, «Dengue Fever : The Need of a Sustained and Integrated Epidemiological Surveillance Approach», *Rev. Salud Bosque*, vol. 12, n.º 1, Art. n.º 1, dic. 2022, doi: 10.18270/rsb.v12i1.4128.

[2] «Boletín Epidemiológico». https://www.ins.gov.co/buscador-eventos/Paginas/Vista-Boletin-Epidemilogico.aspx (accedido 21 de junio de 2022).

[3] «Dengue y dengue grave». https://www.who.int/es/news-room/fact-sheets/detail/dengue-and-severe-dengue (accedido 21 de junio de 2022).

[4] H. Gutierrez-Barbosa, S. Medina-Moreno, J. C. Zapata, y J. V. Chua, «Dengue Infections in Colombia: Epidemiological Trends of a Hyperendemic Country», *Trop. Med. Infect. Dis.*, vol. 5, n.º 4, Art. n.º 4, dic. 2020, doi: 10.3390/tropicalmed5040156.

[5] Gordon S. Linoff y Michael J. A. Berry, *Data Mining Techniques : For Marketing, Sales, and Customer Relationship Management*, vol. 3rd ed. Indianapolis, Ind: Wiley, 2011. Accedido: 22 de junio de 2022. [En línea]. Disponible en: https://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=520245&lang=es&site=ehost-live

[6] G. James, D. Witten, T. Hastie, y R. Tibshirani, *An Introduction to Statistical Learning with Applications in R*, 8.ª ed. Springer, 2017.

[7] R. Bedre, «Linear regression in Python (using sklearn and statsmodels)», *Data science blog*, 23 de marzo de 2020. https://www.reneshbedre.com/blog/linear-regression.html (accedido 20 de mayo de 2023).

[8] Tibco, «¿Qué es un bosque aleatorio?», *TIBCO Software*. https://www.tibco.com/es/reference-center/what-is-a-random-forest (accedido 20 de mayo de 2023).

[9] «Support Vector Machine (SVM)». https://la.mathworks.com/discovery/support-vector-machine.html (accedido 20 de mayo de 2023).

[10] F. Pedregosa *et al.*, «Scikit-learn: Machine Learning in Python», *J. Mach. Learn. Res.*, vol. 12, n.º 85, pp. 2825-2830, 2011.

[11] C. Irrgang, J. Saynisch, y M. Thomas, «Estimating global ocean heat content from tidal magnetic satellite observations», *Sci. Rep.*, vol. 9, n.º 1, Art. n.º 1, may 2019, doi: 10.1038/s41598-019-44397-8.

[12] J. Q. Espinosa, «Dengue en Colombia: epidemiología de la reemergencia a la hiperendemia», *Rev. Salud Bosque*, vol. 5, n.º 1, Art. n.º 1, sep. 2015, doi: 10.18270/rsb.v5i1.186.

[13] Y. N. Bellini Saibene, M. Volpacchio, S. Banchero, y R. Mezher, «Desarrollo y uso de herramientas libres para la explotación de datos de los radares meteorológicos del INTA», sep. 2014.

[14] J. Espinoza Zúñiga, «Aplicación de metodología CRISP-DM para segmentación geográfica de una base de datos pública», *Ing. Investig. Tecnol. - UNAM*, vol. XXI, n.º 1, pp. 1-17, 2020, doi: https://orcid.org/0000-0001-6828-2145.

[15] E. Sylvestre *et al.*, «Data-driven methods for dengue prediction and surveillance using real-world and Big Data: A systematic review», *PLoS Negl. Trop. Dis.*, vol. 16, n.º 1, p. e0010056, ene. 2022, doi: 10.1371/journal.pntd.0010056.

[16] L. C. Madoff, D. N. Fisman, y T. Kass-Hout, «A New Approach to Monitoring Dengue Activity», *PLoS Negl. Trop. Dis.*, vol. 5, n.º 5, p. e1215, may 2011, doi: 10.1371/journal.pntd.0001215.

[17] A. A. M. González, F. G. O. Beltrán, y L. F. S. Guzmán, «Modelo bayesiano para el estudio de la enfermedad del dengue en el departamento de Atlántico, Colombia, años 2010 a 2013», *Perspect. Geográfica*, vol. 22, n.º 2, Art. n.º 2, dic. 2017, doi: 10.19053/01233769.7603.

[18] L. Sepúlveda-Salcedo, O. Vasilieva, H. Martínez, y J. Castro, «Ross Macdonald: Un modelo para la dinámica del dengue en Cali, Colombia», *Rev. Salud Pública*, vol. 17, pp. 749-761, feb. 2016, doi: 10.15446/rsap.v17n5.44685.

[19] «Introduction to Statistical Time Series, 2nd Edition | Wiley», *Wiley.com*. https://www.wiley.com/en-us/Introduction+to+Statistical+Time+Series%2C+2nd+Edition-p-9780471552390 (accedido 22 de febrero de 2023).

[20] «Determining Epidemic Threshold for Dengue Incidences in Singapore Based on Extreme Value Theory - A*STAR OAR». https://oar.a-star.edu.sg/communities-collections/articles/16979 (accedido 23 de marzo de 2023).

[21] «POWER | Data Access Viewer». https://power.larc.nasa.gov/data-access-viewer/ (accedido 17 de septiembre de 2022).

[22] A. S. Gagnon, A. B. G. Bush, y K. E. Smoyer-Tomic, «Dengue epidemics and the El Niño Southern Oscillation», *Clim. Res.*, vol. 19, n.º 1, pp. 35-43, nov. 2001, doi: 10.3354/cr019035.

[23] «Climate Prediction Center - Monitoring & Data: Ocean Niño Index Changes Description». https://origin.cpc.ncep.noaa.gov/products/analysis_monitoring/ensostuff/ONI_change.shtml (accedido 8 de febrero de 2023).

[24] E. Sánchez, «Variación interanual en la estructura vertical de las ondas de Kelvin ecuatoriales y su impacto en las ondas atrapadas a la costa frente a Perú y Chile», ago. 2015. doi: 10.13140/RG.2.1.3453.1687.

[25] «Google Trends», *Google Trends*. https://trends.google.es/trends/explore?date=all&geo=CO-VAC&q=dengue (accedido 13 de diciembre de 2022).

[26] M. Kuhn y K. Johnson, *Applied Predictive Modeling*. New York, NY: Springer, 2013. doi: 10.1007/978-1-4614-6849-3.

[27] N. R. Njeri, «Data Preparation For Machine Learning Modelling», *Int. J. Comput. Appl. Technol. Res.*, vol. 11, n.º 06, pp. 231-235, jun. 2022, doi: 10.7753/IJCATR1106.1008.

[28] M. M. Ahsan, M. A. P. Mahmud, P. K. Saha, K. D. Gupta, y Z. Siddique, «Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance», *Technologies*, vol. 9, n.º 3, Art. n.º 3, sep. 2021, doi: 10.3390/technologies9030052.

[29] «The Box–Cox Transformation: Review and Extensions | Semantic Scholar». https://www.semanticscholar.org/paper/The-Box%E2%80%93Cox-Transformation%3A-Review-and-Extensions-Atkinson-Riani/6f5265bf112ab1ec1168a7e9b99a939e24e260f2 (accedido 18 de marzo de 2023).

[30] J. Brownlee, *Machine Learning Mastery With Python: Understand Your Data, Create Accurate Models and Work Projects End-To-End*. 2021.

[31] «Welcome to the SHAP documentation — SHAP latest documentation». https://shap.readthedocs.io/en/latest/index.html (accedido 29 de marzo de 2023).

[32] «Home (ES - Inicio)», *World Mosquito Program*. https://www.worldmosquitoprogram.org/es/inicio (accedido 11 de febrero de 2023).