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

PREDICCIÓN DEL COMPORTAMIENTO DE LA MALARIA EN COLOMBIA USANDO MODELOS DE MACHINE LEARNING

Sofy Johanna Certuche González
Jaime Martínez Santa
Zaira Idaly Pizo Gurrute

Director: Dr. Delia Ortega Lenis

Enero, 2026

Resumen

La malaria humana o paludismo es una enfermedad infecciosa transmitida por vectores, en este caso son los mosquitos hembras del género Anopheles que proliferan en zonas inferiores a 1600 metros en donde está localizada el 80 % de la Colombia rural, siendo susceptibles de contraer la infección alrededor de 25 millones de personas. La "Estrategia Técnica Mundial contra la Malaria 2016-2030" pactada por la OMS tiene como objetivo erradicar la enfermedad en 85 países que se consideran endémicos dentro de los cuales se encuentra Colombia por sus condiciones climáticas considerándose un problema de salud pública con un reporte del 10 % de los casos de malaria que se registran en la región de las Américas. Debido a la carga de la enfermedad se cuenta aproximadamente con 106 grupos de investigación sólo en la región y su comportamiento epidemiológico se ha tratado de explicar a través de modelos matemáticos (estadísticos y determinísticos), epidemiológicos (SI, SIR, SIS) y con aproximaciones desde la ciencia de datos (Deep Learning, Machine Learning).

Teniendo en cuenta la prevalencia de la enfermedad y su definición como problema de salud pública con estrategias exitosas basadas en el diagnóstico y tratamiento precoz, este proyecto tiene como objetivo desarrollar un modelo predictivo con técnicas de Machine Learning para efectuar una aproximación al comportamiento epidemiológico de la malaria en un departamento en Colombia durante el periodo 2015-2023 debido a que la comprensión de la enfermedad a partir de diferentes modelos va a permitir realizar predicciones temporales, prácticas y aplicables, optimizando tiempo y recursos.

Palabras Clave: Malaria, Machine Learning, Modelos, Epidemiología

Índice general

Índice de figuras

Índice de tablas

INTRODUCCIÓN

La malaria es una enfermedad cuya prevención, vigilancia y control representan un reto de salud pública, teniendo mayor presencia en la región amazónica y algunas zonas rurales y selváticas de países como Brasil, Colombia, Venezuela, Perú, Guayana, Surinam y Bolivia. En Colombia, según el Instituto Nacional de Salud, a la semana epidemiológica 44 de 2024, se tenía un acumulado de 110.343 casos de malaria, predominando la infección transmitida por las dos especies de parásitos más peligrosas, Plasmodium vivax con 61,9 % (68.333) y Plasmodium falciparum con 36,2 % (39.987).

A pesar de es una enfermedad prevenible y curable se continúa propagando, por lo cual el presente proyecto abordó el problema desde la ciencia de datos utilizando técnicas de Machine Learning para crear modelos que permitieron a través de su rendimiento evaluar el comportamiento epidemiológico de la malaria en Colombia durante el periodo 2015-2023 en la región pacífica especifícamente en el departamento del Chocó donde se presenta la mayor prevalencia.

El propósito del presente proyecto fue establecer varios modelos de Machine Learning que permitieron predecir el comportamiento de la malaria en Colombia, a partir de fuentes de datos oficiales (SIVIGILA), con el fin de contribuir con información pertinente a quienes diseñan las políticas públicas para la toma de decisiones de forma oportuna en las acciones de control y prevención de la enfermedad.

El presente documento contiene 5 capítulos: los tres primeros concernientes a la contextualización del proyecto: la definición del problema, objetivos de investigación, marco teórico y antecedentes; un cuarto capítulo corresponde a la preparación y construcción de conjuntos de validación y prueba del corpus de datos de malaria en un departamento en Colombia durante el periodo 2015-2023, se establecen además las variables asociadas al desarrollo de la malaria.

En el capítulo 5 se efectuó el entrenamiento de diferentes modelos de Machine Learning para explicar el comportamiento epidemiológico de la malaria en el departamento de Chocó. En el apartado 6 se desarrolló la evaluación del desempeño de los modelos predictivos de la malaria en Colombia durante el periodo 2015-2023 y en el capítulo final se generaron unas conclusiones y se habló de trabajos futuros.

Capítulo 1

DEFINICIÓN DEL PROBLEMA

1.1. Planteamiento del Problema

La malaria es una enfermedad infecciosa transmitida por vectores causada por el parásito Plasmodium y transmitida por el mosquito anopheles, estos parásitos son de 4 tipos: el Plasmodium vivax (72 %), el Plasmodium falciparum (27 %), el Plasmodium malariae y el Plasmodium ovale; sus fases pueden aparecer hasta 10 a 15 días después de la mordida el mosquito. En el 2022, se estimaron 249 millones de casos de malaria en 85 países endémicos con un aumento de 5 millones de casos comparados con el 2021 [1] [2] [3]. La "Estrategia Técnica Mundial contra la Malaria 2016-2030" tiene como objetivo erradicar la enfermedad en 85 países que la OMS considera como endémico dentro de los cuales se encuentra Colombia por sus condiciones climáticas considerándose un problema de salud pública al reportar el 10 % de los casos de malaria que se registran en la región de las Américas [4] [5] [6].

Globalmente la tasa de mortalidad (por cada 100000) varió de 29 en el año 2000 a 15 en el 2020 se incrementó de nuevo a 15.2 con un descenso a 14.3 en el 2022, cerca del 96 % de las muertes por malaria ocurrieron en 29 países [5]. El mayor número de casos en Colombia se concentra en la región Pacífica en el departamento del Chocó con un total de 218831 casos según el Instituto Nacional de Salud (INS) en el periodo 2015-2023 representando el 36.6 % de la carga de la enfermedad en el país. [7])

Actualmente se cuentan aproximadamente con 106 grupos de investigación sólo en nuestra región y las investigaciones efectuadas han tratado de explicar el comportamiento de la enfermedad a través de modelos matemáticos (estadísticos y determinísticos) [8] [9] [10], epidemiológicos (SI, SIR, SIS) [6], [11] y desde la ciencia de datos (Deep Learning, Machine Learning) [12],[13];estas aproximaciones se han efectuado en diferentes regiones del país pero no están orientados hacia la región Pacífica donde se encuentra el mayor número de casos con unos factores de riesgo demográficos, ambientales y sociales que la hacen diferente a otras sitios del país.

Teniendo en cuenta la prevalencia de la enfermedad y su definición como problema de salud pública cuyas estrategias exitosas están basadas en el diagnóstico y tratamiento precoz [14] [7], este proyecto pretende dar un enfoque con herramientas desde la Ciencia de datos debido a que la comprensión de la enfermedad a partir de diferentes modelos "permite realizar predicciones temporales, simples y prácticas, directamente comprobables y aplicables, economizando tiempo y recursos" [3].

1.2. Formulación del Problema

¿Cuál es el modelo de mejor desempeño para predecir el comportamiento epidemiológico de la malaria en un departamento en Colombia durante el periodo 2015-2023 aplicando técnicas de Machine Learning?

1.2.1. Sistematización

La formulación del problema lleva a las siguientes preguntas de sistematización:

¿Cuáles son las fuentes de datos disponibles para modelar el comportamiento de la malaria en Colombia?

¿Qué calidad tienen los datos históricos de malaria disponibles entre 2015 y 2023 y cómo se pueden tratar problemas como la falta de datos o datos incompletos?

¿Cuáles son las variables que muestran un efecto significativo sobre el comportamiento de casos de malaria en un departamento de Colombia?

¿Cuáles son las técnicas de aprendizaje automático más idóneas de acuerdo con el tipo de variable, para modelar el riesgo de malaria?

¿Qué métricas de evaluación permiten la comparación de modelos para su selección?

Capítulo 2

OBJETIVOS DEL PROYECTO

2.1. Objetivo General

Desarrollar un modelo predictivo con técnicas de Machine Learning para aproximarse al comportamiento epidemiológico de la malaria en un departamento de Colombia durante el periodo 2015-2023.

2.2. Objetivos Específicos

- Preparar y construir conjuntos de validación y prueba del corpus de datos de malaria en Colombia durante el periodo 2015-2023.

- Definir las variables asociadas al comportamiento de la malaria en un departamento de Colombia.

- Entrenar diferentes modelos de Machine Learning que permitan explicar el comportamiento epidemiológico de la malaria en un departamento de Colombia.

- Evaluar el desempeño de los modelos predictivos de la malaria en un departamento de Colombia durante el periodo 2015-2023 para desplegar los resultados del mejor de los modelos.

Capítulo 3

MARCO TEORICO Y ANTECEDENTES

3.1. Marco Teórico

En esta sección se contextualizan las enfermedades transmitidas por vectores, el contexto epidemiológico de la malaria y los diferentes modelos predictivos/explicativos de la enfermedad: epidemiológicos, matemáticos y desde la ciencia de datos

3.1.1. Enfermedades transmitidas por vectores

Las enfermedades transmitidas por vectores son aquellas ocasionadas por agentes infecciosos como virus, parásitos y bacterias, que son transmitidos específicamente por artrópodos como mosquitos, garrapatas, chinches, pulgas y piojos entre otros y constituyen aproximadamente el 17 % de las patologías infecciosas en el ser humano [4]. Un vector es "todo organismo vivo con la capacidad de transportar y transmitir de forma activa y constante cualquier microorganismo desde un hospedero vertebrado e infectado hacia otro susceptible" [15].

La distribución de este tipo de enfermedades se asocia a factores demográficos, ambientales y sociales. Entre los factores demográficos la sobrepoblación y el hacinamiento propician un aumento de vectores y de hospederos susceptibles, en la parte ambiental el cambio climático permite la adaptación de vectores a nuevas altitudes, el calentamiento global que produce la invasión del ser humano en ambientes silvestres y las condiciones climáticas de determinadas zonas al poseer características de temperatura, humedad y precipitación la hacen más propicia para el desarrollo de la enfermedad. Dentro de los factores sociales se destacan la carencia de medidas de higiene en los niveles personal, de vivienda y de comunidad, la marginación, pobreza bajo nivel educativo, migración, exposición ocupacional y urbanización descontrolada[16].

3.1.2. Contexto epidemiológico de la malaria

Los síntomas de la malaria no complicada (infección sintomática con tolerancia a la vía oral y ausencia de síntomas de malaria severa) son inicialmente inespecíficos e incluye malestar general, fatiga, fiebre, escalofríos, criodiaforesis, cefalea. La malaria severa se debe de sospechar en caso de presentar alteración del estado de conciencia y crisis convulsivas. El tiempo de incubación habitual

es de 13 a 28 días [3].

En cuanto al tratamiento las directrices de la OMS centran el manejo de la malaria en 3 estrategias: erradicación, de control y eliminación [1]. En la actualidad, las principales estrategias de control de la malaria se basan en "el diagnóstico y tratamiento precoces y efectivos mediante test de diagnóstico rápido (TDR) y terapia combinada con artemisinas, en la utilización de mosquiteras impregnadas con insecticidas, en el tratamiento preventivo intermitente en embarazadas y niños, en la lucha contra el mosquito transmisor, y en el desarrollo (y actual pilotaje) de vacunas" [14].

Los mosquitos hembras del género Anopheles proliferan en zonas inferiores a 1600 metros en donde está localizada el 80 % de la Colombia rural, siendo susceptibles de contraer la infección por malaria alrededor de 25 millones de personas. En Colombia se ha tenido un total de 2,004,049 casos en el periodo 1970-2019 con 2 picos en 1980, 4 picos epidémicos en la década de los 90,3 picos en la década de 1970s, 2000s, 2010s; el pico más intenso ocurrió en 1998. Para el 2022 los departamentos con mayor carga de la enfermedad fueron Chocó (34 %), Nariño (17 %) y Córdoba (12 %); la región del pacífico concentra el 55 % de los casos afectando población vulnerable como las comunidades indígenas y afrodescendientes. Los grupos de edad más afectados fueron entre los 15 y 29 años y los 10 y 24 años respectivamente afectando la población en edad productiva [17].En Colombia la malaria sigue siendo un problema prioritario de salud pública, según el Instituto Nacional de Salud, en 2024 se registraron 135.290 casos de malaria complicada con 28 muertes [7].

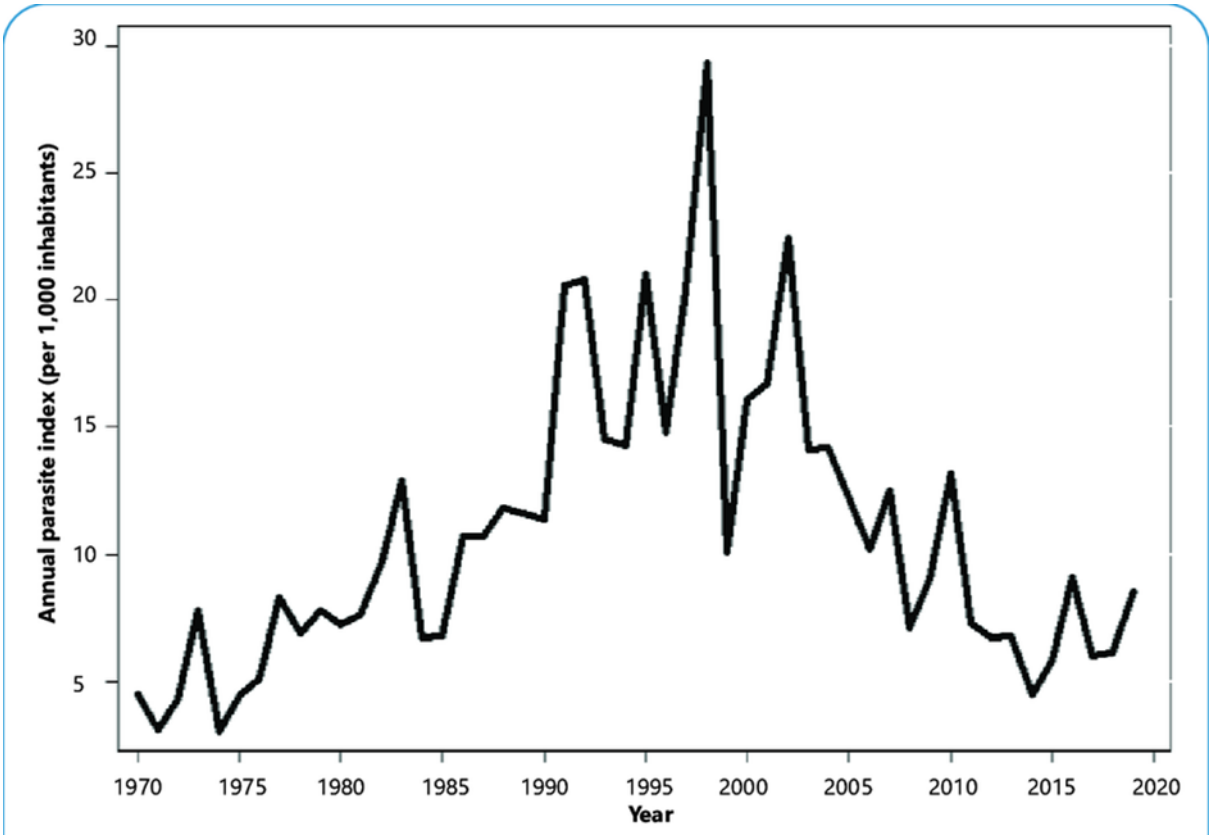

Figura 3.1: Malaria en Colombia 1970-2019
Fuente: https://www.researchgate.net/figure/Malaria-epidemics-in-Colombia-1970-2019_fig1_36026625

3.1.3. Modelos predictivos/explicativos en malaria

El uso de modelos predictivos/explicativos de la malaria se basa en la premisa de que el comportamiento de la dinámica de casos de de cada mes debe ser similar a lo registrado en el mismo mes de años anteriores, sin embargo esto no sucede así ya que pueden ocurrir cambios en el comportamiento de los brotes, los sistemas de vigilancia, las decisiones de salud pública y la dispersión de los casos notificados Los estudios realizados para entender el comportamiento de las enfermedades producidas por vectores dependen de la información con la que se cuente y del enfoque que se quiera dar, lo cual produce tiene limitaciones en Colombia debido a la cultura de los sistemas de información. Existen modelos matemáticos que analizan la dinámica epidemiológica de la malaria, pero existen algunos inconvenientes en su formulación porque están orientados a la realidad de una determinada población; los modelos epidemiológicos toman elementos matemáticos combinados con técnicas específicas como la regresión lineal múltiple, análisis de supervivencia, análisis de datos categóricos, estadística espacial y los métodos bayesianos para ensayos clínicos.

Estos modelos requieren conocer los niveles habituales de incidencia en cada población y momento del año además de la dinámica habitual de casos de nuevo diagnóstico en caso de la aparición de brotes. Los modelos en Ciencia de datos permite el análisis de grandes volúmenes de información y permiten aproximaciones al comportamiento epidemiológico de una enfermedad tan variable como la malaria con limitaciones en Colombia debido a su heterogeneidad sociodemográfica y climática lo que hace que su comportamiento sea diferente por regiones.

3.1.3.1 Modelos epidemiológicos En el contexto epidemiológico de las enfermedades infecciosas transmitidas por vectores la transmisión del agente infeccioso tiene un proceso cuyas categorías dependen del parásito y el tipo de infección [18], estas categorías están representadas por la notación S-E-I-R [19]. El primer grupo es la fracción de la población Susceptible (S) a la infección, luego la fracción de población Expuesta (E) quienes están infectados por el patógeno pero que no son capaces de transmitir la infección a otros durante el periodo latente. La próxima categoría son los infectados (I) a través de la interacción con los susceptibles y finalmente estos individuos quienes pueden recuperarse (R). Estas variaciones en la estructura de la enfermedad pueden dar diferentes modelos, ejemplo de esto son individuos infectados (I) que no se recuperan y mueren o individuos recuperados (R) quienes desarrollan inmunidad temporal o definitiva. Esto origina 8 diferentes modelos epidemiológicos: SI, SIS, SEI, SEIS, SIR, SIRS, SEIR y SEIRS [20]. En el estudio de la malaria se han planteado varios modelos teóricos [21]: la teoría de los sucesos, los modelos de plantación de procesos, el modelo de la historia natural de la enfermedad, modelo explicativo y el modelo de la alternancia [22]. La teoría de los sucesos se enfoca en la prevención de la enfermedad y el control de los mosquitos vectores, los modelos de plantación de procesos estudian la dinámica de la malaria con técnicas de diseño, el modelo de la historia natural de la enfermedad plantea la interacción entre agente, huésped y ambiente con sus posibles desenlaces: recuperación, la incapacidad o la muerte [23]. El modelo explicativo con un énfasis más clínico se centra en el origen de la enfermedad, los procesos patofisiológicos, la distribución y el manejo que se le da al paciente [24]. En contraste a esto, el modelo de alternancia se centra en la cadena de transmisión de la enfermedad [25].

3.1.3.2 Modelos matemáticos Existen varios modelos matemáticos para la malaria que a pesar de no ser tan complejos explican en forma exitosa los factores que influencian la transmisión de la enfermedad, los más conocidos son: (a) Modelo Ross (b)Modelo Macdonald y (c) Modelo Anderson-May Model. Estos modelos consideran factores como el periodo de latencia de la infección, la susceptibilidad diferenciada por grupo etario, la inmunidad adquirida y la heterogeneidad espacial genética del huésped y del parásito. Los modelos emplean un sistema de ecuaciones diferenciales ordinarias para expresar los cambios temporales en la prevalencia de la infección malárica en los humanos, pero se vuelven determinísticos debido a que las ecuaciones no incorporan variabilidad estocástica (aleatoria) en los parámetros componentes. Sin embargo, estos modelos se quedan cortos cuando se incorporan nuevas variables y se consideran más interacciones [18] por lo cual algunos autores sugieren que se debe incorporar a los modelos las posibles consecuencias de heterogeneidad en el contacto huésped -vector [26].

En Colombia algunos modelos matemáticos integran "escenarios como el cambio climático y su incidencia en la transmisión, así como la toma de decisiones en el sector de salud, mediante la representación de las múltiples interacciones entomológicas, epidemiológicas y climáticas de la transmisión" [27]. En general estos modelos matemáticos estocásticos o determinísticos tienden a integrar varios factores estructurales, espaciales y temporales que puedan explicar la transmisión de la enfermedad [18].

3.1.3.3. Modelos desde la ciencia de datos

3.1.3.3.1. Machine Learning Los modelos de Machine Learning (ML) han demostrado ser herramientas fundamentales en la lucha contra la malaria, ya que pueden manejar y analizar grandes volúmenes de datos, identificar patrones y realizar predicciones precisas sobre la propagación de la enfermedad. Estos modelos permiten a los profesionales de la salud pública tomar decisiones informadas sobre cómo distribuir los recursos y aplicar estrategias de prevención, predicción.

En varios estudios se ha buscado determinar si algunos parámetros podrían usarse para predecir la malaria mediante el uso de modelos de regresión logística, Redes bayesianas, análisis discriminante lineal de Fisher y K-Vecinos más cercanos (K-NN), determinando en qué medida se podría estimar la probabilidad de que un paciente tenga o no la enfermedad de la malaria. [28]

Con los Modelos de clustering se agrupan áreas con características similares en cuanto a riesgo de malaria, las Redes neuronales recurrentes (RNN) se enfocan en datos secuenciales para predecir brotes futuros y las Redes Generativas Antagónicas (GANs) permiten generar datos sintéticos cuando los datos reales son limitados.

Ejemplo de estos modelos son los utilizados por la Fundación Gates (Bill & Melinda Gates Foundation) ha apoyado numerosos proyectos que utilizan machine learning para predecir y controlar la malaria en diversas regiones. Utilizan datos geoespaciales, climáticos, epidemiológicos y de salud pública para crear modelos predictivos de brotes [29]

3.1.3.3.2. Deep Learning Los modelos de Deep Learning (aprendizaje profundo) son extremada mente eficaces para abordar problemas complejos como la predicción de la malaria, ya que tienen la capacidad de aprender patrones ocultos y no lineales a partir de grandes volúmenes de datos. Es un subcampo más avanzado de Machine Learning, que utiliza redes neuronales profundas para trabajar con datos no estructurados (como imágenes y texto) y requiere una mayor cantidad de datos y potencia computacional. El Deep Learning es utilizado para entrenar modelos a partir de grandes conjuntos de datos no estructurados y estructurados los cuales son capaces de identificar patrones complejos que podrían ser difíciles de detectar mediante métodos tradicionales.

Estos modelos de Deep Learning pueden integrar y analizar datos de diferentes fuentes, como imágenes, series temporales, datos geoespaciales y registros epidemiológicos, lo que permite una

predicción más precisa de la propagación de la enfermedad. Los modelos más utilizados en la predicción de la malaria son:

- Redes Neuronales Artificiales (ANN), que se utilizan para clasificar y predecir brotes de malaria basados en datos climáticos y epidemiológicos.
- Redes Neuronales Convolucionales (CNN), aplicadas al análisis de imágenes satelitales y microscópicas para detectar áreas propensas a la transmisión y la presencia de vectores de malaria.
- Redes Neuronales Recurrentes (RNN) y LSTM, ideales para predecir brotes de malaria mediante el análisis de datos históricos y series temporales, considerando factores como el clima y la estacionalidad.
- Autoencoders, que reducen la dimensionalidad de los datos y ayudan a detectar anomalías en los patrones de propagación de la malaria.
- Generative Adversarial Networks (GANs), utilizadas para generar datos sintéticos cuando los datos reales son escasos, y simular escenarios de propagación de la enfermedad.
- Transformers y Redes de Atención, que permiten procesar secuencias de datos y enfocarse en factores clave para mejorar las predicciones y la interpretabilidad de los modelos.

Ejemplo de estos modelos son los utilizados por IBM Research que se centran en la utilización de Deep Learning y técnicas de Big Data para mejorar la predicción de brotes de malaria. IBM aplica su plataforma de inteligencia artificial y aprendizaje automático, como IBM Watson, para analizar y procesar grandes volúmenes de datos, incluyendo datos históricos de malaria, condiciones climáticas, patrones geoespaciales y otros factores ambientales que influyen en la propagación de la enfermedad.[30]

3.2. Antecedentes

La revisión efectuada en el proyecto fue realizada con las palabras claves: Malaria, Machine Learning, Modelos, Epidemiología. Se utilizaron como motores de búsqueda Jstor, Mendeley Science direct, Scopus. Web of Science, Access Engineering, Pudmed, los artículos y el material escogido como referencia se tomó de los últimos 10 años en idioma español e inglés.

Después de ejecutar el plan de trabajo y revisar los estudios previos identificados en el marco teórico se encontró que existen aproximaciones a los modelos desde la matemática, la epidemiología y la ciencia de datos, estos modelos describen escenarios con diversos factores como la demografía, el huésped, la distribución geográfica, las interacciones sociales, el clima y el desarrollo de la enfermedad en un marco espacial y temporal [31].Otros factores que juegan un papel importante son el grupo

etario, la especie del parásito, el tratamiento, las dinámicas poblacionales y el ambiente los cuales afectan la dinámica de la malaria a diferentes escalas [18].

Debido a la complejidad de la enfermedad y su predominio en regiones geográficas específicas es importante entrenar diferentes modelos predictivos que permitan una aproximación con herramientas proporcionadas por la ciencia de datos. Así es como en la última década, se han utilizado diversos métodos de clasificación, como árboles de decisión, redes neuronales, máquinas de vectores de soporte y clasificadores de árboles aleatorios, para el diagnóstico, predicción de la malaria [32]-[33]; algunos de estos estudios se han centrado en desarrollar tecnologías específicas para la discriminación celular o la detección de parásitos. La mayor parte del conocimiento sobre la malaria se encuentra en la combinación de segmentación, características y clasificación [31]

Los Modelos de series de tiempo (ARIMA, LSTM) se han aplicado en la investigación de predicción de la malaria. Se utilizan para entrenar y luego pronosticar puntos de tiempo futuros basándose en datos históricos [34]. Algunos de estos modelos son útiles cuando se usan con variables climáticas como predictivos de casos de malaria aunque poseen problemas de sobreestimación [35]. Los resultados sugieren una asociación entre la incidencia de la malaria y la variabilidad climática la cual varía entre regiones geográficas [36].

Ejemplos de estos estudios los encontramos en Tuta [37] que incluye en un modelo de redes neuronales variables de laboratorio clínico (hemoglobina, recuento de leucocitos, recuento de plaquetas, bilirrubina total) síntomas (presencia de disnea, vómitos, fiebre persistente), historial previo de malaria y, uso previo de medicamentos para la malaria. Este módelo encontró unos valores de sensibilidad entre 13 % al 47 %, valores predictivos positivos que oscilaron entre el 37 % y el 88 % y una especificidad variando del 79 % al 90 %. Otro estudio utilizando redes neuronales encontró un exactitud de 65.22 %, una especificidad de 57.89 % y una sensibilidad del 100 % [38]. La exactitud de otros estudios encuentran sensibilidades entre 77 y 85 % al comparar con otros modelos como Support Vector Machine y modelos de regresión múltiple [39][40].

Otro estudio efectuado en la India mostró que el Random Forest fue el mejor modelo al compararlo con otros con una accuracy de 90.92, este estudio tuvo en cuenta tambien diagnóstico, laboratorios y síntomas de la enfermedad [41]; en Senegal se obtuvieron resultados similares con un accuracy de 92 % para el random forest [42]. El estudio de para predecir riesgo de malaria en viajeros comparó varios modelos de Machine Learning encontrando los mejores resultados en un modelo XGBoost (XGB) obteniendo un AUC in la validación cruzada de 0.90 (95 %CI: 0.84–0.96) y un AUC de 0.80 con el Test set. Se obtuvo una sensibilidad de 90 % y una especificidad de 81 % [43]. El estudio realizado en Nigeria encontró como el modelo con mayor rendimiento el Adaboost con un accuracy of 98.2 %, precisión de 96.6 %, y una tasa de error de 1.8 %. con una fuerte relacion entre edad y sexo [44].

También existen propuestas de modelos de Machine Learning que le dan prioridad a las variables climáticas e incluyen precipitación, superficie de radiación, temperatura presión atmosférica y

humedad relativa. Estos modelos privilegian el modelo XGBoost [36]. El estudio de Gutierrez [45] muestra la asociación entre incidencia de malaria y temperatura, reportando que en 100 municipios de Colombia "la transmisión de la malaria se intensifica a medida que las temperaturas aumentan de 15 °C a aproximadamente 23,5 °C, después de lo cual el efecto disminuye a temperaturas superiores a 23,5 °C".

El estudio realizado en Ghana tomó los parámetros hematológicos, la edad y el género de los pacientes para predecir la malaria mediante el uso de modelos de regresión logística, naive Bayes, análisis discriminante lineal de Fisher y K vecinos más cercanos. Encontró que "el modelo de mejor rendimiento fue la regresión logística, con un área bajo la curva de 81.5 %, una especificidad y sensibilidad del 74.6 % y 79.89 %, respectivamente, con un valor predictivo positivo del 39.8 % y un valor predictivo negativo del 94.6 %" [46]

Los anteriores modelos muestran la utilidad del Machine Learning al efectuar modelos predictivos que dan diferentes métricas las cual tienen contextos dispares por lo cual es necesario desarrollar modelos en Colombia para regiones como la costa Pacífica pues presenta unas condiciones diferentes a la de los modelos expuestos anteriormente desarrollados en otros continentes.

Capítulo 4

PREPARACIÓN Y CONSTRUCCIÓN DE CONJUNTOS DE VALIDACIÓN Y PRUEBA DEL CORPUS DE DATOS DE MALARIA

4.1. Desarrollo

En este capítulo se especifican los métodos utilizados como el cargue de datos, la identificación de datos faltantes por variable y la imputación y estandarización de los datos, además se desarrollan los pasos requeridos para la obtención de los resultados como son la construcción de la base de datos y el análisis exploratorio de la base de datos de malaria en Colombia.

4.1.1. Construcción y entendimiento de la base de datos

Para el presente proyecto se utilizó como fuente principal el sistema oficial de vigilancia epidemiológica (SIVIGILA) regulado por el Instituto Nacional de Salud de Colombia (INS) cuyo acceso está disponible al público y permite recolectar los datos sobre la malaria en Colombia. SIVIGILA es la principal fuente de información epidemiológica en Colombia y posee características que la hacen idónea para la investigación al poseer cobertura conceptual (la información corresponde a los objetivos planteados en este proyecto), cobertura geográfica (Colombia), cobertura temporal (2007-2023) y amplitud porque en esta fuente de tipo secundario se obtiene la información de las variables geográficas, demográficas y de tratamiento en la malaria. Para la obtención de las variables climáticas de este periodo se obtuvo la información del dataset del IDEAM con registros mes a mes para el periodo de estudio.

El procesamiento de datos se efectuó con la información obtenida por la librería implementada en R sivirep versión 1.0.0 que proporciona funciones para la manipulación de datos y la generación de reportes automatizados basados en las bases de datos individualizadas de casos de SIVIGILA, esta librería fue desarrollada por la Pontificia Universidad Javeriana y el Instituto Nacional de Salud y permite el acceso libre a las bases de datos para el periodo de estudio 2015-2023

Al realizar la descarga de datos a través de SIVIREP se encontró una base de datos de 674534 registros y 72 variables para el estudio de la malaria en Colombia durante el periodo de estudio.

Propiedad	**Valor**
Número de filas	674.534
Número de columnas	72

Tabla 4.1: Resumen de las dimensiones del conjunto de datos

Estas variables incluyen información de la ficha epidemiológica que el INS destina para el estudio de enfermedades transmisibles o eventos de notificación obligatoria. Los datos de este dataset dividen la información en 2 grandes grupos: Datos del paciente y Ubicación geográfica del caso.

Se nombran las variables de la data:

No.	**Variable Data**	**Nombre Variable**	**Descripción**	**Opción de respuesta**
1	cod_eve	Código del evento en SIVIGILA	Código numérico que identifica el evento de interés en salud pública notificado y lo asocia al protocolo específico de vigilancia (malaria).	Numérico – INS
2	fec_not	Fecha de notificación	Fecha en la que el caso es notificado oficialmente al sistema de vigilancia epidemiológica.	DD/MM/AAAA
3	semana	Semana epidemiológica	Semana epidemiológica en la que se realiza la notificación del caso.	1–52 / 53
4	ano	Año epidemiológico	Año epidemiológico en el que se notifica el caso.	AAAA
5	cod_pre	Código UPGD notificadora	Código del prestador de servicios de salud (UPGD) que notifica el caso.	Numérico (REPS)

No.	Def Variable Data	Nombre Variable	Descripción	Opción de respuesta
6	cod_sub	Código subred/unidad	Código que identifica la sede o subíndice del prestador.	Numérico
7	edad	Edad del paciente	Edad cumplida del paciente al momento del evento.	Numérico
8	uni_med	Unidad de edad	Unidad de medida de la edad reportada.	1=Años, 2=Meses, 3=Días
9	nacionalidad	Código nacionalidad	Código que identifica la nacionalidad legal del paciente.	Catálogo INS/DANE
10	nombre_nacionalidad	Nombre nacionalidad	Nombre del país de nacionalidad del paciente.	Texto
11	sexo	Sexo	Sexo biológico reportado del paciente.	1=Masculino, 2=Femenino, 3=Indeterminado
12	cod_pais_o	Código país ocurrencia	Código del país donde ocurrió la exposición al evento.	DANE
13	cod_dpto_o	Código departamento ocurrencia	Código del departamento donde ocurrió el evento.	DANE
14	cod_mun_o	Código municipio ocurrencia	Código del municipio donde ocurrió el evento.	DANE
15	area	Área de ocurrencia	Área geográfica donde ocurrió el caso.	1=Cabecera municipal, 2=Rural disperso, 3=Centro poblado
16	ocupacion	Ocupación	Actividad laboral principal del paciente.	CIUO
17	tip_ss	Tipo aseguramiento	Tipo de régimen de afiliación al sistema de salud.	1=Contributivo, 2=Subsidiado, 3=Especial, 4=Excepción, 5=No asegurado

No.	Def Variable Data	Nombre Variable	Descripción	Opción de respuesta
18	cod_ase	Código EPS	Código de la EAPB a la que pertenece el paciente.	Numérico
19	per_etn	Pertenece a grupo étnico	Indica pertenencia étnica del paciente.	Sí / No
20	gru_pob	Grupo poblacional	Grupo poblacional especial al que pertenece el paciente.	1=Indígena, 2=ROM, 3=Raizal, 4=Palenquero, 5=Negro/Mulato, 6=Otro
21	nom_grupo	Nombre grupo poblacional	Nombre del grupo poblacional especial.	Texto
22	estrato	Estrato socioeconómico	Clasificación socioeconómica del lugar de residencia habitual.	1 a 6
23	gp_discapa	Discapacidad	Indica si presenta discapacidad.	Sí / No
24	gp_desplaz	Desplazado	Indica situación de desplazamiento forzado.	Sí / No
25	gp_migrant	Migrante	Identifica si el paciente es migrante.	Sí / No
26	gp_carcela	Privado de la libertad	Indica si está privado de la libertad.	Sí / No
27	gp_gestan	Gestante	Señala estado de gestación.	Sí / No
28	sem_ges	Semanas gestación	Número de semanas de gestación.	Numérico
29	gp_indigen	Indígena	Indica pertenencia a pueblo indígena.	Sí / No
30	gp_pobicfb	A cargo ICBF	Bajo protección del ICBF.	Sí / No
31	gp_mad_com	Madre comunitaria	Ejerce rol de madre comunitaria.	Sí / No

No.	Def Variable Data	Nombre Variable	Descripción	Opción de respuesta
32	gp_desmovi	Desmovilizado	Persona desmovilizada de grupo armado.	Sí / No
33	gp_psiquia	Trastorno psiquiátrico	Antecedente de trastorno psiquiátrico.	Sí / No
34	gp_vic_vio	Víctima de violencia	Víctima de cualquier tipo de violencia.	Sí / No
35	gp_otros	Otro grupo especial	Pertenece a otro grupo especial.	Sí / No
36	fuente	Fuente notificación	Fuente a través de la cual se notifica el caso.	1=UPGD, 2=BAI, 3=Laboratorio
37	cod_pais_r	Código país residencia	Código del país de residencia habitual.	DANE
38	cod_dpto_r	Código depto residencia	Código del departamento de residencia habitual.	DANE
39	cod_mun_r	Código municipio residencia	Código del municipio de residencia habitual.	DANE
40	cod_dpto_n	Código depto notificación	Código DANE del departamento desde el cual se notifica.	DANE
41	cod_mun_n	Código municipio notificación	Código DANE del municipio desde el cual se notifica.	DANE
42	fec_con	Fecha consulta	Fecha de primera consulta por síntomas.	DD/MM/AAAA
43	ini_sin	Inicio síntomas	Fecha de inicio de síntomas.	DD/MM/AAAA
44	tip_cas	Tipo de caso	Clasificación inicial del caso.	1=Sospechoso, 2=Confirmado lab, 3=Confirmado nexo
45	pac_hos	Hospitalización	Indica si requirió hospitalización.	Sí / No

No.	Def Variable Data	Nombre Variable	Descripción	Opción de respuesta
46	fec_hos	Fecha hospitalización	Fecha de hospitalización.	DD/MM/AAAA
47	con_fin	Condición final	Condición final del paciente.	1=Vivo, 2=Muerto
48	fec_def	Fecha defunción	Fecha de defunción cuando aplica.	DD/MM/AAAA
49	ajuste	Tipo ajuste	Indica si el caso fue ajustado.	0=Sin ajuste, 1=Lab, 2=Unidad análisis
50	fecha_nto	Fecha nacimiento	Fecha de nacimiento del paciente.	DD/MM/AAAA
51	cer_def	Certificado defunción	Número de certificado de defunción cuando aplica.	Sí / No
52	cbmte	Causa básica muerte	Causa básica de muerte según CIE-10.	CIE-10
53	fec_arc_xl	Fecha archivo Excel	Fecha de validación o contraste con SISPRO.	DD/MM/AAAA
54	fec_aju	Fecha ajuste	Fecha en la que se realizó el ajuste del caso.	DD/MM/AAAA
55	fm_fuerza	Fuerza pública	Indica si pertenece a fuerza pública.	Ejército/Policía/Otro
56	fm_unidad	Unidad fuerza pública	Nombre de la unidad militar o policial.	Texto
57	fm_grado	Grado fuerza pública	Grado o rango dentro de la fuerza pública.	Texto
58	confirmados	Caso confirmado	Cumple criterios de confirmación para malaria.	Sí / No
59	consecutive_origen	Consecutivo sistema origen	Número consecutivo en sistema previo a SIVIGILA.	Numérico
60	va_sispro	Validado en SISPRO	Validado o cruzado con SISPRO.	Sí / No
61	estado_final_de_ caso	Código estado final	Clasificación final definitiva del caso.	INS

No.	Def Variable Data	Nombre Variable	Descripción	Opción de respuesta
62	nom_est_f_caso	Nombre estado final	Descripción del estado final del caso.	Texto
63	nom_upgd	Nombre UPGD	Nombre de la UPGD notificadora.	Texto
64	pais_ocurrencia	País ocurrencia	Nombre del país donde ocurrió el evento.	Texto
65	nombre_evento	Evento	Nombre del evento notificado.	Malaria
66	departamento_ocurrencia	Departamento ocurrencia	Nombre del departamento donde ocurrió el evento.	Texto
67	municipio_ocurrencia	Municipio ocurrencia	Nombre del municipio donde ocurrió el evento.	Texto
68	pais_residencia	País residencia	Nombre del país de residencia habitual.	Texto
69	departamento_residencia	Departamento residencia	Nombre del departamento de residencia habitual.	Texto
70	municipio_residencia	Municipio residencia	Nombre del municipio de residencia habitual.	Texto
71	departamento_notificacion	Departamento notificación	Departamento desde el cual se realiza la notificación.	Texto
72	municipio_notificacion	Municipio notificación	Municipio desde el cual se realiza la notificación.	Texto

La tabla muestra la incidencia de la malaria en Colombia para el periodo 2015-2023 el cual muestra un comportamiento irregular con picos en el 2016 y el 2023.

Año	Frecuencia
2015	56.666
2016	84.778
2017	55.136
2018	63.143
2019	80.418
2020	81.368
2021	73.985
2022	73.561
2023	105.479

Tabla 4.3: Frecuencia de registros por año en malaria

Al mirar la base de datos se encuentran 5 posibles eventos, para el presente estudio se tomó la decisión de escoger los eventos 470 y 490 correspondientes a la Malaria Falciparum y a la Malaria Vivax que representan el 97 % de casos de malaria en Colombia.

Código	Nombre del Evento	Frecuencia
460	Malaria asociada (formas mixtas)	9.932
470	Malaria falciparum	317.062
490	Malaria vivax	336.951
495	Malaria complicada	10.447
540	Mortalidad por malaria	142

Tabla 4.4: Frecuencia de eventos relacionados con malaria reportados

Para la depuración de la base de datos también se tuvo en cuenta el país de ocurrencia del evento para tener en cuenta sólo los casos ocurridos en Colombia obteniendo el dataset malaria Colombia con 665.088 registros

Tabla 4.5: Frecuencia de registros por país de procedencia

#	País	Frecuencia
1	AFGANISTÁN	2
2	ANGOLA	3
3	ANTÁRTIDA	1
4	ARABIA SAUDÍ	2
5	ARUBA	1
6	BANGLADESH	2
7	BARBADOS	1
8	BENIN	1
9	BOLIVIA	1
10	BRASIL	214
11	CABO VERDE	1

Continúa en la siguiente página

#	País	Frecuencia
12	CAMERÚN	5
13	CHILE	1
14	CHIPRE	1
15	COLOMBIA	665088
16	COMORAS	3
17	COSTA DE MARFIL	5
18	COSTA RICA	2
19	CROACIA	1
20	ECUADOR	41
21	ERITREA	9
22	ESLOVENIA	1
23	ESPAÑA	1
24	ETIOPÍA	4
25	GAMBIA	6
26	GHANA	3
27	GUAYANA FRANCESA	4
28	GUINEA	2
29	GUINEA ECUATORIAL	6
30	GUYANA	2
31	INDIA	3
32	ISLAS MALVINAS (FALKLAND)	1
33	ISLAS MARIANAS DEL NORTE	3
34	KENYA	3
35	MACAO	1
36	MACEDONIA, ANTIGUA REPÚBLICA DE YUGOSLAVIA	1
37	MALAWI	1
38	MÉXICO	1
39	MOZAMBIQUE	2
40	NICARAGUA	3
41	NIGERIA	5
42	PAÍSES BAJOS	1
43	PANAMÁ	14
44	PERÚ	324
45	QATAR	1
46	REINO UNIDO DE GRAN BRETAÑA E IRLANDA DEL	2
47	REPÚBLICA CENTROAFRICANA	15
48	REPÚBLICA DEMOCRÁTICA DEL CONGO	4
49	REPÚBLICA DOMINICANA	2
50	REPÚBLICA POPULAR DEL CONGO	1
51	REPÚBLICA UNIDA DE TANZANIA	2
52	SAN PEDRO Y MIGUELON	1
53	SIERRA LEONA	1
54	SINGAPUR	5
55	SUDÁFRICA	3
56	TERRITORIO OCUPADO DE PALESTINA	1
57	TOGO	1
58	UGANDA	3
59	VANUATU	2
60	VENEZUELA	8713

Se revisan las 72 variables para el estudio y se escogen 13 variables que se consideraron importantes para la posterior implementación del modelo. Este nuevo dataset está conformado por 557236 registros y 13 variables.

Tabla 4.6: Primeras filas del conjunto `malaria_data`

cod_eve	nombre_evento	país_ocurrencia	departamento	municipio	sem	año	edad	sexo	etnia	SS	área	grupo_quinquenio	n
460	MALARIA ASOCIADA (FORMAS MIXTAS)	COLOMBIA	AMAZONAS	AMAZONAS. MUNICIPIO DESCONOCIDO	02	2016	21	F	4	N	3	20-24	1
460	MALARIA ASOCIADA (FORMAS MIXTAS)	COLOMBIA	AMAZONAS	AMAZONAS. MUNICIPIO DESCONOCIDO	10	2016	21	M	4	S	3	20-24	1
460	MALARIA ASOCIADA (FORMAS MIXTAS)	COLOMBIA	AMAZONAS	AMAZONAS. MUNICIPIO DESCONOCIDO	10	2016	29	F	4	N	3	25-29	1
460	MALARIA ASOCIADA (FORMAS MIXTAS)	COLOMBIA	AMAZONAS	AMAZONAS. MUNICIPIO DESCONOCIDO	10	2016	39	M	1	S	3	35-39	1
460	MALARIA ASOCIADA (FORMAS MIXTAS)	COLOMBIA	AMAZONAS	AMAZONAS. MUNICIPIO DESCONOCIDO	20	2017	43	F	6	S	3	40-44	1
460	MALARIA ASOCIADA (FORMAS MIXTAS)	COLOMBIA	AMAZONAS	AMAZONAS. MUNICIPIO DESCONOCIDO	30	2018	4	F	6	N	2	0-4	1

Se realiza la identificación de datos faltantes en las 13 variables de interés (nombre del evento,sexo, edad,grupo etnico, afiliacion al sistema de salud, area de residencia, municipio y departamento de ocurrencia, semana, año, numero de registros) visualizadas en la figura 4.1 encontrando que no es necesario efectuar imputación o estandarización de datos al no existir faltantes.

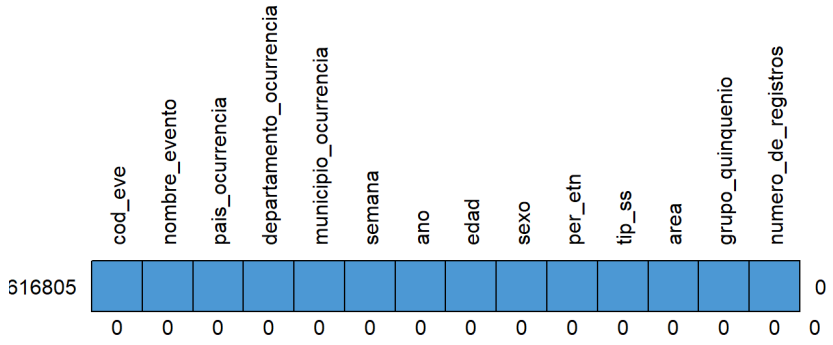

Figura 4.1: Figura Datos Faltantes

Para la obtención del dataset de las variables climáticas la información se obtuvo de la base de datos del IDEAM de donde se tomaron las variables que la mayoría de estudio realizan con las enfermedades producidas por vectores: precipitación, humedad y temperatura (Humedad relativa mínima mensual, Humedad relativa media mensual, Humedad relativa máxima mensual, Precipitación,Temperatura mínima media mensual,Temperatura media mensual, Temperatura máxima media mensual).

La base de datos final estuvo constituida por 13 variables sociodemográficas (obtenidas de SIVIGILA) y 5 variables climáticas (Obtenidas del IDEAM)

4.1.2. Analisis exploratorio de los datos

Se realizo un análisis univariado (distribución de frecuencias y representaciones gráficas) y bivariado (representaciones gráficas y análisis de correlación) de las variables.

Analisis Univariado: Se efectuo en las variables sexo, grupo etnico, afiliacion al sistema de salud, area de resisdencia y departamento de ocurrencia.

Sexo

Sexo	Frecuencia
F	260100
M	356705

Tabla 4.7: Frecuencia de registros por sexo

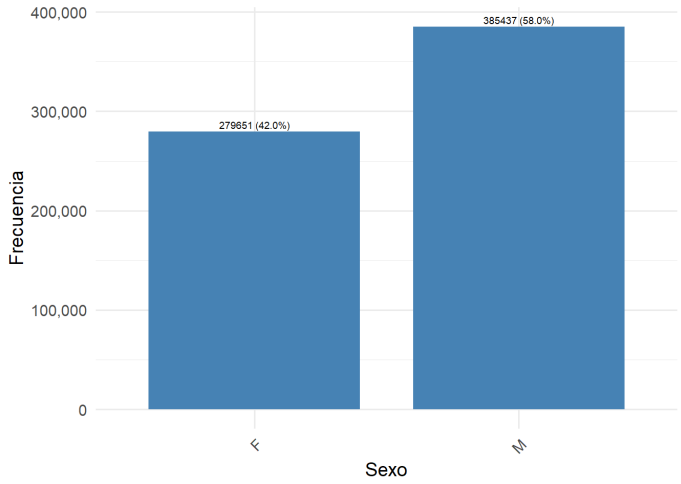

Figura 4.2: Frecuencia por sexo.

La mayor proporción de casos en hombres (58 %) sugiere un patrón diferencial de exposición al vector, posiblemente asociado a actividades laborales de mayor riesgo en zonas endémicas. Este resultado justifica la inclusión del sexo como variable explicativa en los modelos posteriores, aun cuando su efecto pueda ser moderado frente a factores ambientales.

Análisis Bivariado: Se efectuaron los siguientes cruces:

- Grupo Quinquenio vs Género
- Departamento ocurrencia vs Género
- Departamento ocurrencia vs Grupo etario

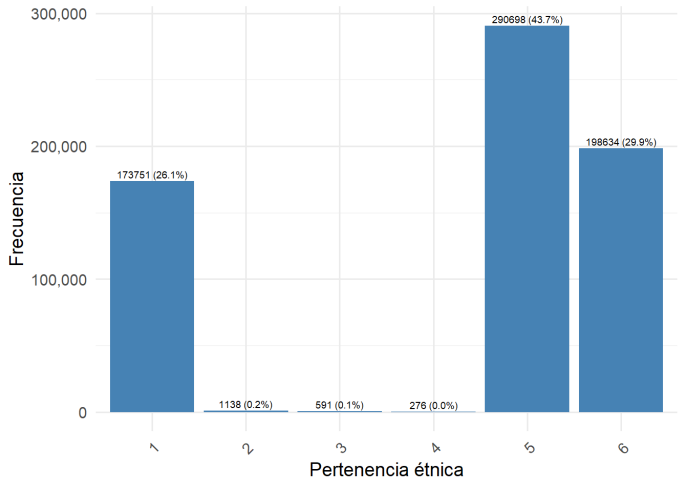

Figura 4.3: Frecuencia por grupo étnico

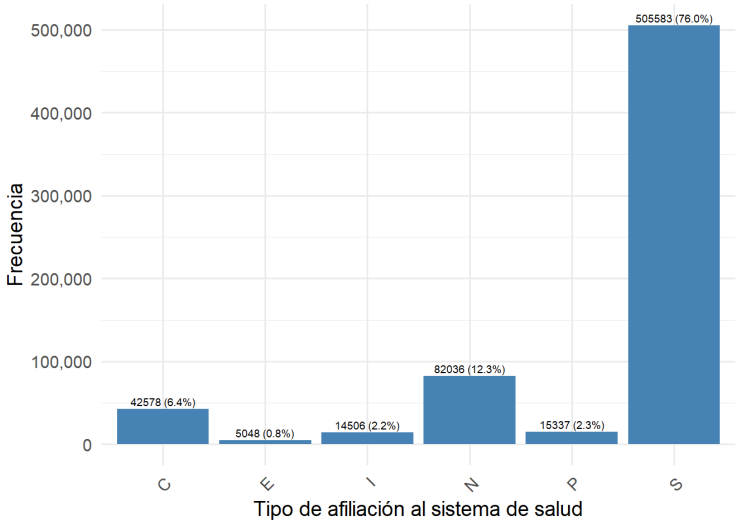

Figura 4.4: Frecuencia de afiliación al sistema de salud.

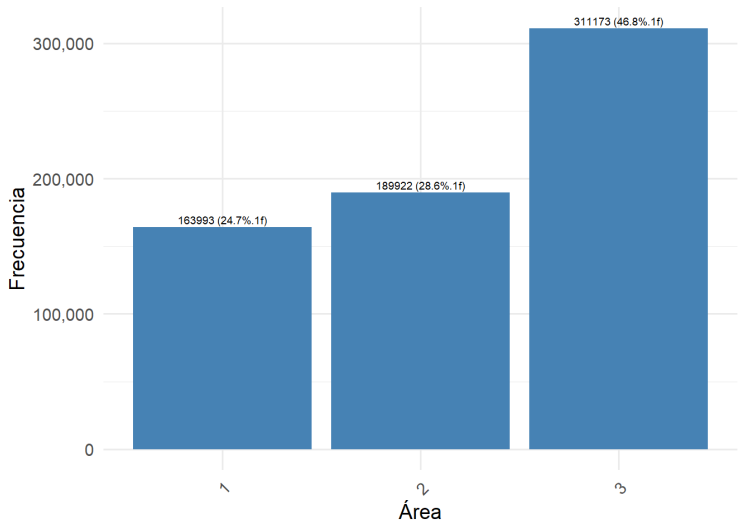

Figura 4.5: Distribución de los casos de malaria por área de residencia.

Las Figuras 4.3, 4.4 y 4.5 evidencian un patrón consistente de concentración de casos en poblaciones históricamente vulnerables. La mayor carga de malaria en población afrocolombiana e indígena, afiliados al régimen subsidiado y residentes en áreas rurales sugiere que la transmisión de la enfermedad está estrechamente asociada a condiciones territoriales y poblacionales específicas. Este comportamiento resalta la importancia de considerar características demográficas y de localización geográfica en el análisis multivariado, complementando el estudio de los factores climáticos.

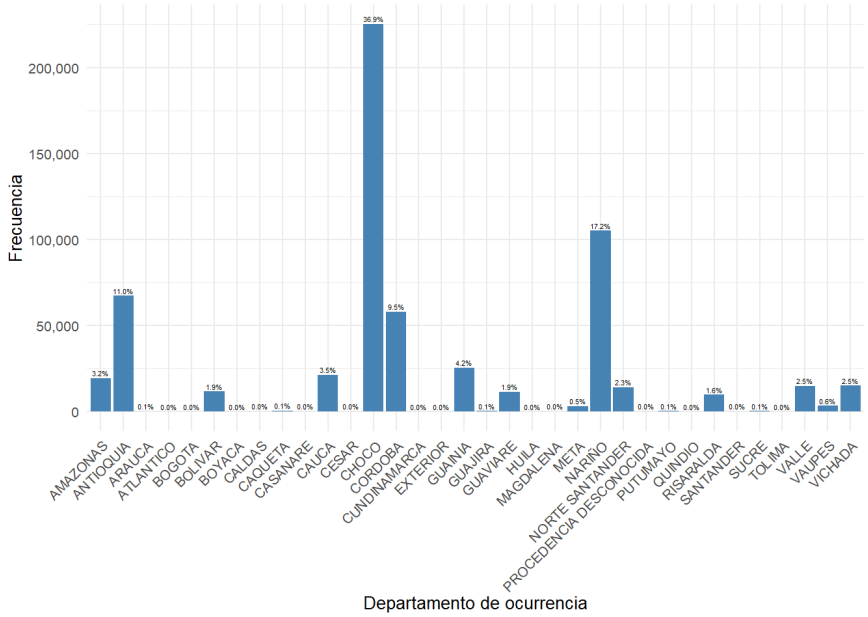

Figura 4.6: Departamentos con mayor ocurrencia de casos de malaria en Colombia.

Se destacan Chocó, Antioquia, Cauca, Córdoba, Guainía y Nariño, en su mayoría localizados en la región del Pacífico, lo que refleja la influencia de condiciones ambientales y socioeconómicas propias de esta zona..

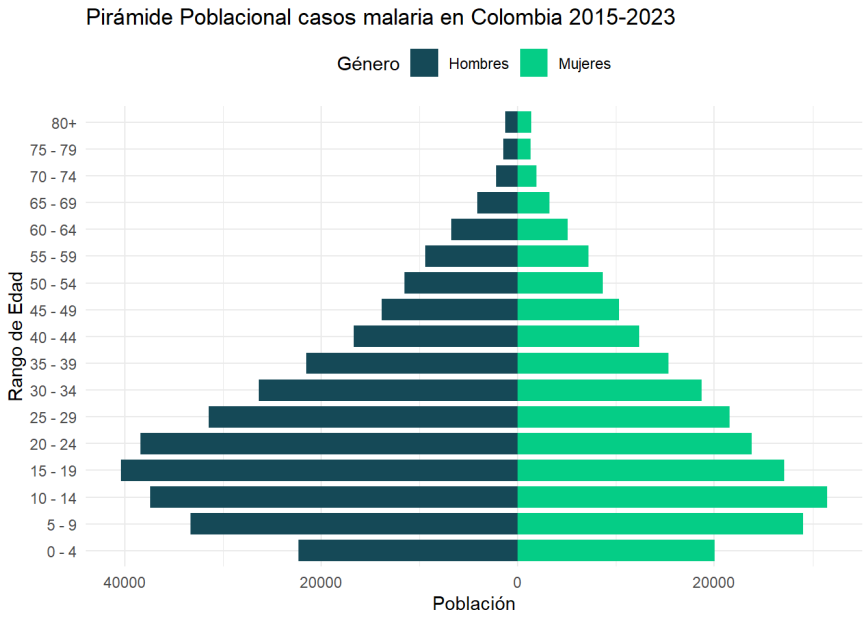

Figura 4.7: Pirámide poblacional de los casos de malaria.

La pirámide poblacional nos muestra una distribución asimétrica por género con predominio masculino independiente del grupo etario,se encuentra mayor distribución de los casos de malaria en el rango de edad entre los 10 y los 24 años donde se encuentra concentrada la mayor parte de casos de malaria en Colombia durante el periodo de estudio.

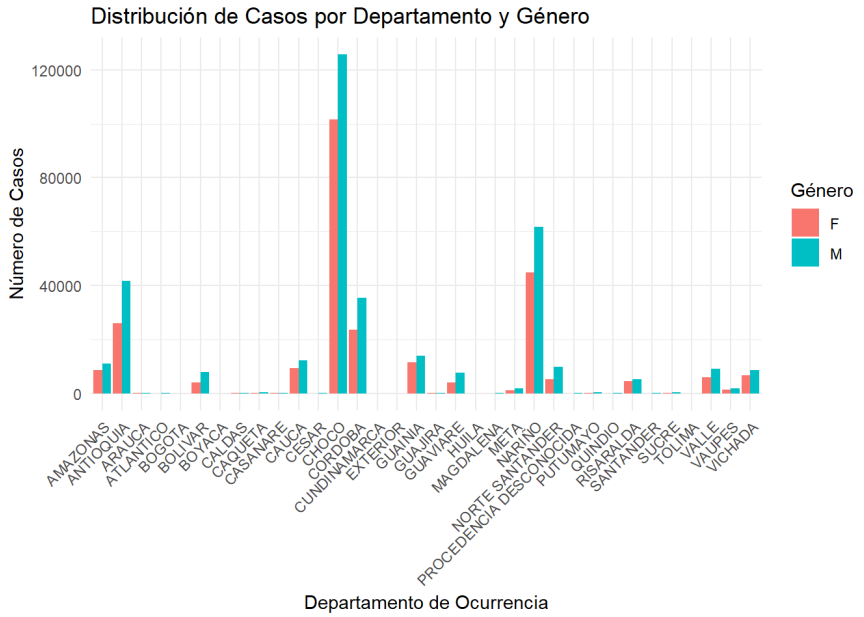

Figura 4.8: Casos de malaria por departamento y sexo.

Los departamentos con el mayor número de casos reportados de malaria son Antioquia, Chocó y Nariño. En todos ellos, se observa que el porcentaje de casos en hombres es superior al de mujeres, lo que podría reflejar diferencias en la exposición, actividad laboral o factores sociales asociados al género.

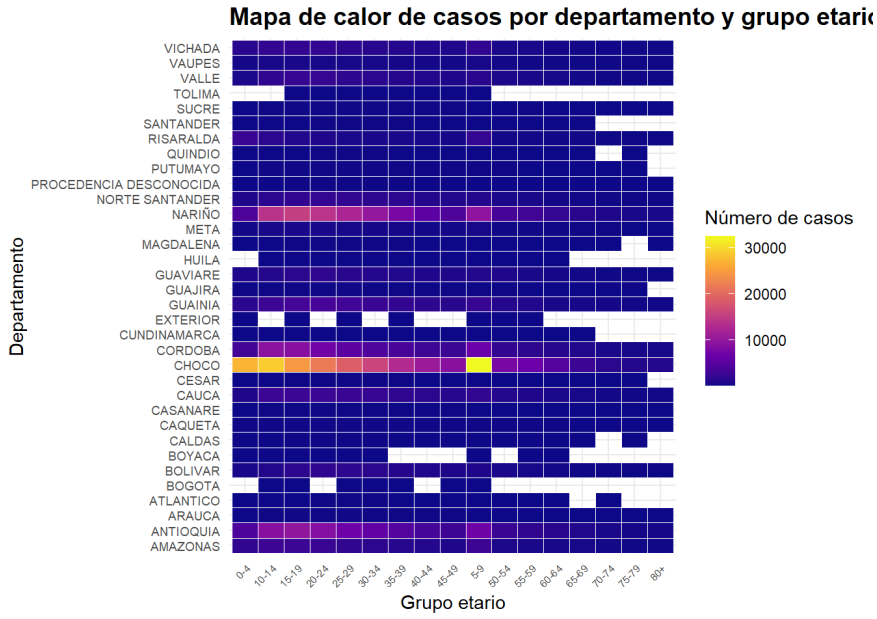

Figura 4.9: La mayor incidencia se concentra en adultos jóvenes y de mediana edad (0-49 años).

Los grupos 0-4 y 5-9 años presentan mayor numero de casos, lo cual es preocupante porque indica que niños pequeños están siendo expuestos al vector.

4.2. Variables asociadas al comportamiento de la malaria en Colombia

4.2.1. Datos del paciente

Género: El compromiso por especie de malaria se comporta igual en ambos sexos y los hombres son más afectados que las mujeres (relación hombre: mujer 1,5:1,0) [47]. El género masculino es el más afectado como se mostró en el análisis descriptivo siendo esta una variable asociada al comportamiento de la malaria en Colombia [48] - [49]

Edad: Algunos estudios muestran que la malaria predomina en edades adultas en el grupo etario entre 31 y 40 años de edad. Este comportamiento es similar en las regiones de mayor presencia de la malaria y en los departamentos de mayor prevalencia excepto en el Chocó donde tiene un predominio a través de todo el ciclo vital afectando varios grupos etarios [49]-[50].

Pertenencia étnica: Los habitantes de raza negra tienen un peso importante en ambas regiones y en la zona de Urabá hay presencia de unos pocos indígenas kunas (tule), emberas (katíos) y zenúes. La

población afrocolombiana y la población indígena son predominantes en las regiones de predominio de la malaria por lo cual esta es una variable asociada al comportamiento de la malaria en Colombia [51].

Tipo de seguridad social: El tipo de afiliación al *Sistema General de Seguridad Social en Salud* en Colombia "es un buen indicador del nivel socioeconómico, y es un factor predictor de mayor morbilidad y mortalidad prematura asociada con los factores determinantes sociales de la salud". Esta variable está asociada al comportamiento de la malaria en Colombia pues "las desigualdades en salud se generan por diferencias en las condiciones sociales y económicas, lo cual influye en el riesgo de enfermar y la forma de enfrentar la enfermedad" [51].

4.2.2. Ubicación geográfica del caso

Departamento de ocurrencia: Los departamentos de la región Pacífica en Colombia son los de mayor prevalencia de la enfermedad en Colombia siendo el Chocó el de mayor número de casos (37 %) además de presentar unas condiciones sociales y económicas vulnerables lo que favorece la aparición y el desarrollo de enfermedades transmitidas por vectores [52].

Este departamento además cuenta con la mayor completitud en cuanto a las variables climáticas y tiene una distribución en todos los grupos etarios razón por la cual es el departamento que puede ser más representativo para modelar el comportamiento de la malaria en Colombia

Departamento de ocurrencia: Los departamentos de la región Pacífica en Colombia son los de mayor prevalencia de la enfermedad en Colombia siendo el Chocó el de mayor número de casos (37 %) además de presentar unas condiciones sociales y económicas vulnerables lo que favorece la aparición y el desarrollo de enfermedades transmitidas por vectores [52].

4.2.3. Variables climáticas

Temperatura: Colombia tiene un clima predominantemente tropical pero diverso debido a su variada topografía y a su situación ecuatorial. Las temperaturas son relativamente constantes durante todo el año. Las temperaturas medias oscilan entre 24 y 30 grados centígrados. Entre 15 y 20^{o}C, es notablemente más fresco en la zona alta de los Andes.

Precipitación: Por encima de los 3.000 metros, existe incluso un clima glaciar de alta montaña. En Colombia hay dos estaciones templadas secas y lluviosas al año, de abril a mayo y de octubre a noviembre. Las temperaturas y precipitaciones, sin embargo, dependen en gran medida de la altitud y la proximidad del Caribe y el Pacífico. Entre diciembre y marzo, sin embargo, es soleado y seco en casi todas partes.La región amazónica y la costa del Pacífico reciben la mayor cantidad de precipitaciones, lo que da lugar a frondosas selvas tropicales y a una gran diversidad de especies. La región andina del suroeste recibe las precipitaciones principalmente de los vientos alisios del

Pacífico, mientras que las laderas orientales de los Andes están más bajo la influencia de los vientos del Atlántico.La costa del Caribe y las llanuras del valle del Magdalena tienen un clima tropical de sabana. En general, esta región se considera la más seca, ya que los vientos alisios del Caribe apenas introducen aire marino húmedo en el país. La precipitación promedio en Colombia es de 2576,67 mm, en el periodo de estudio se presentó el valor mínimo en los últimos 100 años: 2085.08 en el 2015.

Humedad: Los valores de humedad relativa dependen necesariamente de la temperatura del momento. En las zonas tropicales continentales, en donde las variaciones de la temperatura durante el día son generalmente grandes, la humedad relativa cambia considerablemente en el curso del día. En Colombia la humedad en promedio supera el 65 % en promedio siendo este valor mas alto en la region pacifica donde su promedio es mayor al 80 %.

Capítulo 5

ENTRENAMIENTO DE DIFERENTES MODELOS DE MACHINE LEARNING.

5.1. Momento de definición y delimitación el problema

El presente estudio se centra en el análisis del comportamiento epidemiológico de la malaria en el departamento del Chocó, una de las regiones con mayor carga de la enfermedad en Colombia. Durante el periodo de estudio (2015–2023), el Chocó aportó el 36,63 % de los casos de malaria reportados en el país, concentrados en 31 municipios, lo que equivale a 218 831 casos del total de 597 255 casos registrados en Colombia.

Esta distribución evidencia una alta concentración territorial de la enfermedad, ya que más de un tercio de los casos nacionales se presentan en un solo departamento, mientras que el 63,37 % restante se distribuye entre los demás departamentos del país.

Variables para poder cumplir con el objetivo propuesto.

Variable respuesta:
Número de casos de malaria por mes y año en el departamento del Chocó.

Variables predictoras:
Desde el punto de vista del modelamiento estadístico, se redujo la dimensionalidad del modelo (evitando exceso de categorías) para facilitar la interpretación epidemiológica y permitir identificar patrones diferenciales de riesgo asociados a la edad [53]. Cada variable predictora se dejó con 2 o 3 categorías para cumplir este objetivo. Las variables contempladas para la ejecución de los modelos (Ver tabla 6.1) fueron:

Año y semanas: Las semanas epidemiológicas corresponden a 52 en 1 año, para efectos de unificación en la base de datos con las variables climáticas se efectuó la conversión a las semanas calendario de cada mes entre el 2015 y el 2023. La información en el dataset quedó mes a mes.

Edad: Se tomó la clasificación efectuada por el DANE en 3 grupos de edad. Los menores de 15 años son considerados de alta vulnerabilidad inmunológica, especialmente en zonas endémicas donde la exposición constante al vector genera una inmunidad parcial que tarda años en desarrollarse; los

niños presentan tasas de infección y complicaciones más altas [54] [55] [56].
Los adultos de 15 a 64 años corresponden a la población económicamente activa, con mayor exposición ocupacional al vector (agricultura, minería, pesca, trabajo en bosques). Además, las conductas de movilidad laboral incrementan la probabilidad de infección [57]. Los mayores de 64 años, aunque suelen presentar menor exposición, su inmunosenescencia (disminución natural del sistema inmune) los hace más propensos a complicaciones graves en caso de infección [58]. Esta clasificación coincide con la utilizada por el DANE (Departamento Administrativo Nacional de Estadística de Colombia) y por organismos internacionales (OMS, ONU, Banco Mundial) para describir la estructura poblacional en niños, adultos y adultos mayores.

Sexo: Se consideraron las dos categorías clásicas correspondientes al sexo biológico.

Pertenencia étnica: Se establecieron tres clasificaciones: Afrodescendiente, Indígena y Otras.

Seguridad social: Se realizaron tres agrupaciones. Las dos principales corresponden a los regímenes Subsidiado y Contributivo. En la categoría Otros se incluyeron las personas sin afiliación a un régimen tradicional o sin ningún tipo de afiliación.

Área: El DANE reconoce tres clasificaciones territoriales. Para este análisis, las categorías Centro poblado y Rural disperso se agruparon como Rural. Por razones prácticas y de comparación con otros estudios, se utilizó la clasificación general Rural y Urbana.

Variables climáticas: Los modelos de Machine Learning en malaria toman las variables climáticas: clima, humedad y precipitación, el uso de más variables climáticas puede generar multicolinealidad, afectando la estabilidad y la interpretabilidad de los modelos predictivos [59]. La temperatura determina la supervivencia del vector, la tasa de picadura y el desarrollo del parásito dentro del mosquito [60], la humedad influye en la longevidad del mosquito y su capacidad de vuelo y la precipitación crea criaderos temporales, afectando la densidad del vector [61].

Tabla 5.1: Variables predictoras y opciones de respuesta

VARIABLES	**OPCIONES DE RESPUESTA**
Año	2015 – 2023
Edad	< 15 15–59 > 60
Sexo	F M
Pertenencia étnica	Indígena Afrocolombiano Otros (Rom, Raizal, Palenquero, Otros)
Tipo de seguridad social	Contributivo Subsidiado Otros (especial, sin afiliación, excepción, indeterminado)
Área	Urbana Rural (Centro poblado – rural disperso)
Variables climáticas	Humedad relativa mínima mensual Humedad relativa media mensual Humedad relativa máxima mensual Precipitación Temperatura mínima media mensual Temperatura media mensual Temperatura máxima media mensual

5.2. Análisis Descriptivo

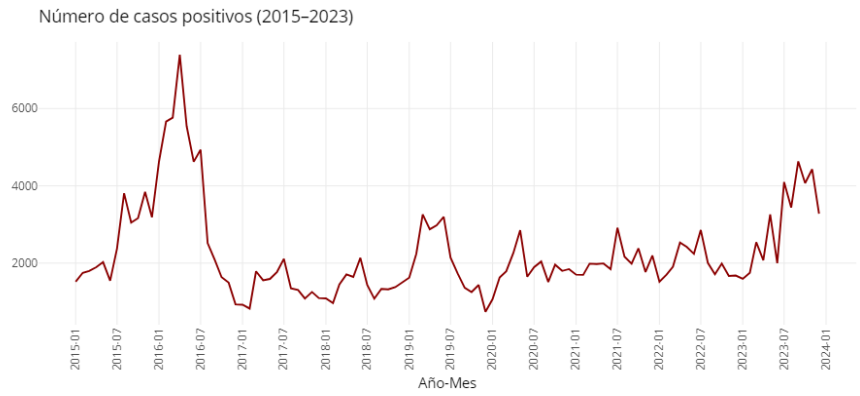

Figura 5.1: Número de casos positivos de malaria (2015–2023)

En la gráfica se muestra la evolución mensual del número de casos positivos de malaria en el departamento del Chocó entre 2015 y 2023. Se observa una variación estacional marcada, con picos y valles recurrentes a lo largo del tiempo, destacándose el mayor pico en abril de 2016.

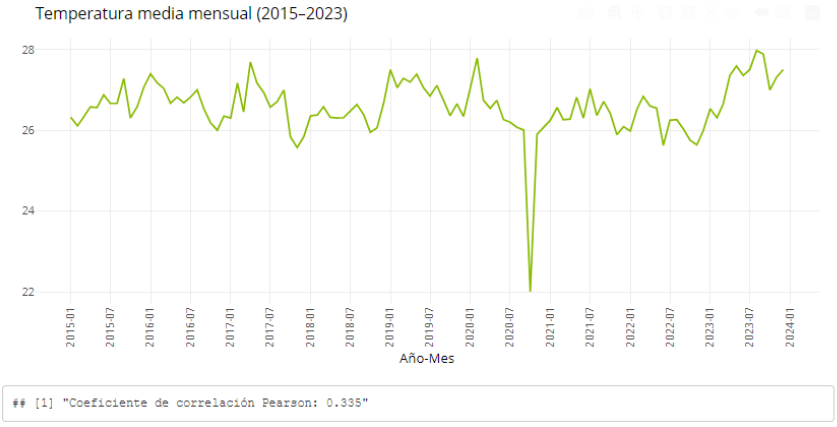

Figura 5.2: Temperatura media mensual (2015–2023)

La gráfica presenta la temperatura media mensual, la cual oscila entre 22°C y 28°C, evidenciando una relativa estabilidad con fluctuaciones estacionales leves; la mayor caída se presentó en octubre de 2020. Cabe destacar que una temperatura cálida favorece la supervivencia y desarrollo del vector, y que la transmisión de la malaria se intensifica a medida que la temperatura aumenta desde 15 °C hasta aproximadamente 23,5 °C [45]. El coeficiente de correlación de Pearson entre el número de casos positivos de malaria y la temperatura media mensual en el Chocó durante el período 2015–2023 es $r = 0{,}335$, lo que indica una relación positiva moderada; es decir, la temperatura influye en la incidencia de la malaria, aunque no constituye el único factor determinante.

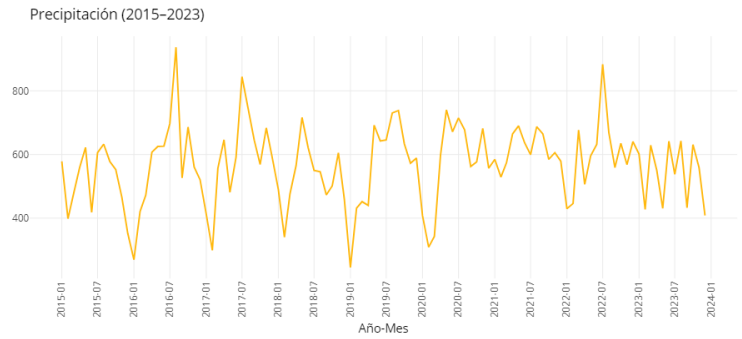

Figura 5.3: Precipitación 2015-2023

En la gráfica se muestra la precipitación mensual en Chocó. Se observan variaciones importantes por mes, esto es, meses con lluvias altas que podrían coincidir con aumentos de casos de malaria, ya que las lluvias intensas pueden crear criaderos temporales para mosquitos o afectar el acceso a servicios de salud lo que podría desencadenar un reporte tardío. La precipitación puede ser usada como predictor en alertas tempranas.

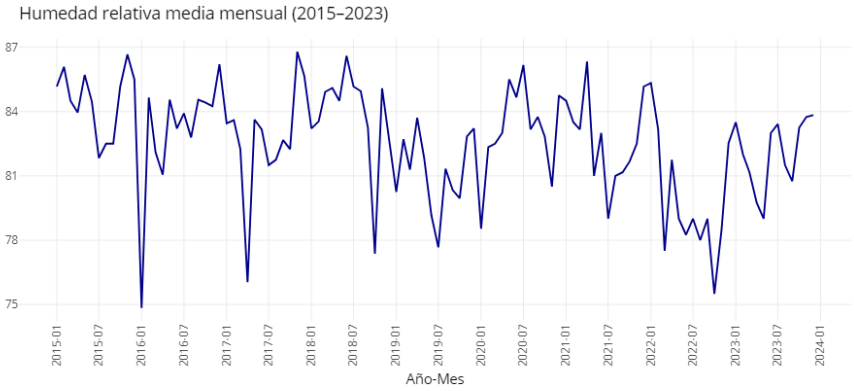

Figura 5.4: Humedad Relativa 2015-2023

La figura muestra la humedad relativa media por mes. En general, se observa una humedad consistentemente alta (75 % - 87 %). Una humedad alta constante favorece la longevidad del vector.

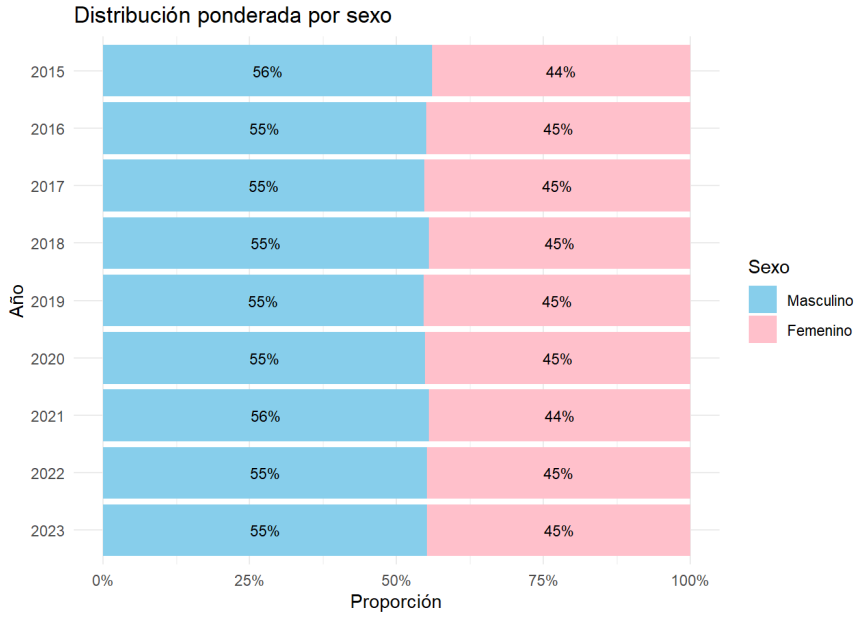

Figura 5.5: Distribución Ponderada por sexo

La figura se muestra la proporción ponderada de casos por sexo. Alrededor del 56 % corresponde al sexo masculino y el 44 % al femenino.

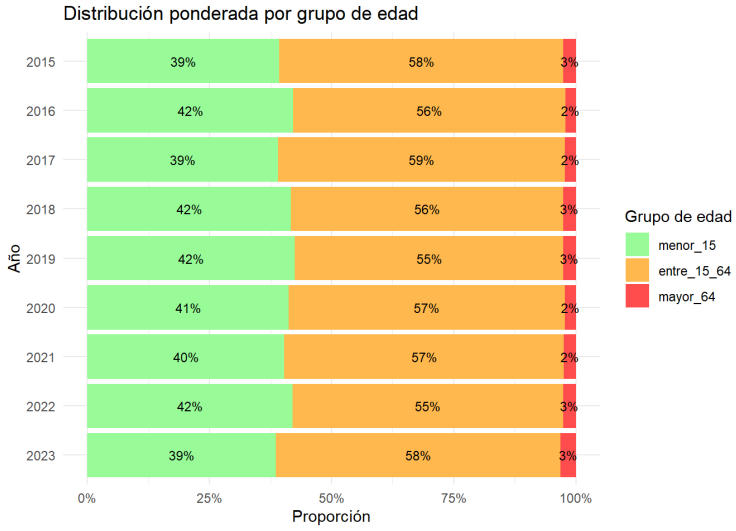

Figura 5.6: Distribución Ponderada por grupo de edad

La figura muestra las proporciones ponderadas por año para los tres grupos de edades evaluados, esto es, menores de 15 años, entre 15 y 64 años, y mayores de 64 años. En todos los años, el grupo entre 15 y 64 años concentra la mayor proporción de casos positivos de malaria en Chocó (55 %–59 %) , seguido por menores de 15 años (39 %–42 %) , por último se tiene una minoría de casos en mayores de 64 años (2 %–3 %) .

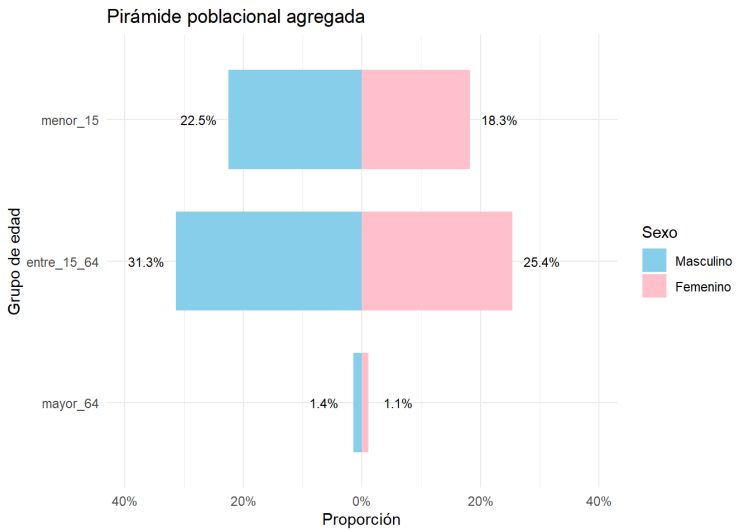

Figura 5.7: Pirámide Poblacional Agregada

La figura muestra las proporciones conjuntas por sexo y grupo de edad agregadas durante los años 2015 a 2023. Los hombres tienen mayor representación respecto a las mujeres en los tres grupos de edades.

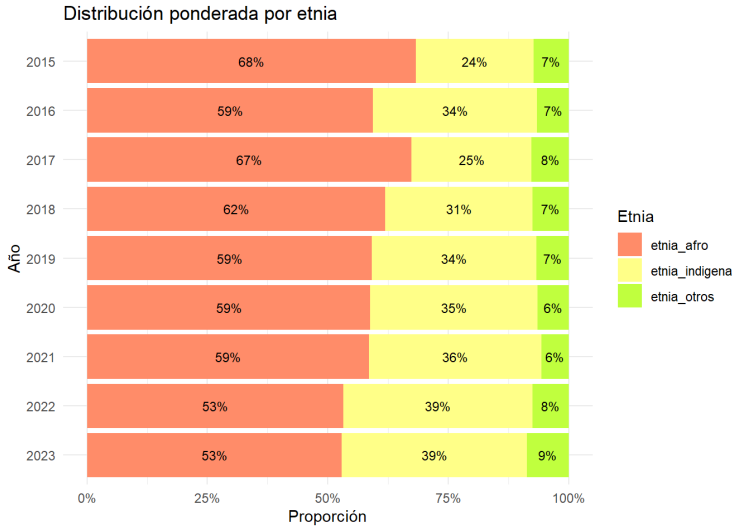

Figura 5.8: Distribución Ponderada por Etnia

En la figura se muestran las proporciones por grupos étnicos (afrodescendiente, indígena, y otros) entre los años 2015 y 2023. Se evidencia dominio de la población afrodescendiente (53 %–68 %) , seguido por indígena (24 %–39 %) , con una minoría en otras etnias (6 %–9 %). En los años 2022 y 2023 hay una tendencia al aumento de proporción de casos en indígenas.

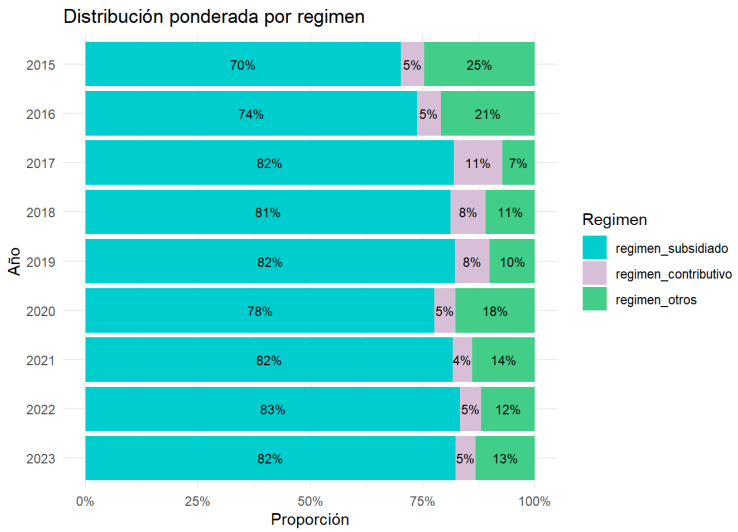

Figura 5.9: Distribución Ponderada por Régimen

En la figura se muestran las proporciones por año por régimen de salud, esto es, régimen subsidiado, régimen contributivo, y otros. La gran mayoría de casos de malaria reportados corresponden a afiliados al régimen subsidiado, generalmente población de menores recursos (70 %–84 %).

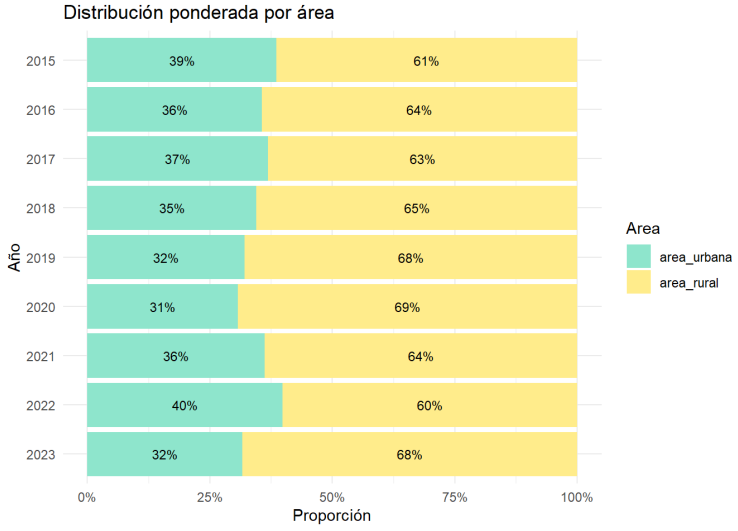

Figura 5.10: Distribución Ponderada por Área

En la figura se muestran las proporciones por año por área urbana y área rural. Se presenta mayor proporción de casos en el área rural (61 %–69 %) . Cabe tener en cuenta que la transmisión predominantemente rural es consistente con la naturaleza de los vectores y con el acceso limitado a control de los mismos.

5.3. Análisis de Serie de Tiempo

5.3.1. Análisis exploratorio de la serie temporal

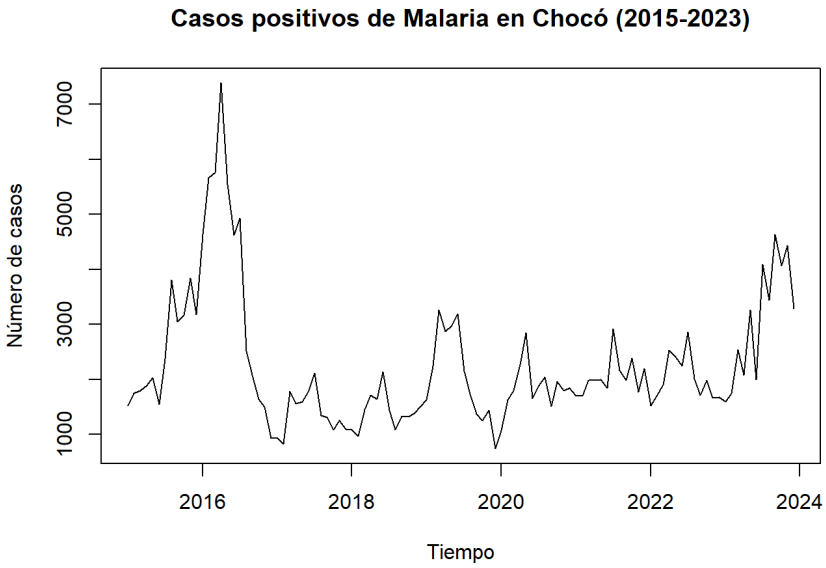

Figura 5.11: Casos positivos de Malaria en Chocó (2015-2023)

En cuanto a la evolución temporal de los casos positivos de malaria en el departamento del Chocó, 2015-2023 se observa un pico epidémico pronunciado en 2016, seguido de una disminución significativa y un incremento progresivo a partir de 2020, lo que evidencia una dinámica temporal no trivial y potencialmente influenciada por factores ambientales y epidemiológicos.

5.3.2. Evaluación de estacionariedad

5.3.2.1. Prueba Dickey–Fuller aumentada (ADF)

Con el fin de evaluar la idoneidad de la serie temporal para su modelación, se analizó su estacionariedad y estructura de dependencia temporal mediante la prueba Dickey-Fuller aumentada (ADF) y el estudio de las funciones de autocorrelación (ACF) y autocorrelación parcial (PACF).

Para evaluar la estacionariedad de la serie temporal se aplicó la prueba Dickey-Fuller aumentada (ADF), cuyos resultados se muestran a continuación:

Estadístico ADF: 3.6389
Orden de rezagos: 4

Valor p: 0.03299

La hipótesis nula de la prueba ADF establece que la serie presenta una raíz unitaria (no es estacionaria), mientras que la hipótesis alternativa indica que la serie es estacionaria.Dado que el valor p (0.03299) es menor que el nivel de significancia del 5 % (= 0,05), se rechaza la hipótesis nula. En consecuencia, se concluye que la serie temporal es estacionaria y no requiere diferenciación adicional para su modelación

Dado que el valor p (0.03299) es menor que el nivel de significancia del 5 % ($\alpha = 0{,}05$), se rechaza la hipótesis nula. En consecuencia, se concluye que la serie temporal es estacionaria y no requiere diferenciación adicional para su modelación

5.3.3. Análisis de autocorrelación (ACF) y autocorrelación parcial (PACF)

Posteriormente, se analizaron las funciones de autocorrelación (ACF) y autocorrelación parcial (PACF) de la serie original con el objetivo de identificar la estructura de dependencia temporal.

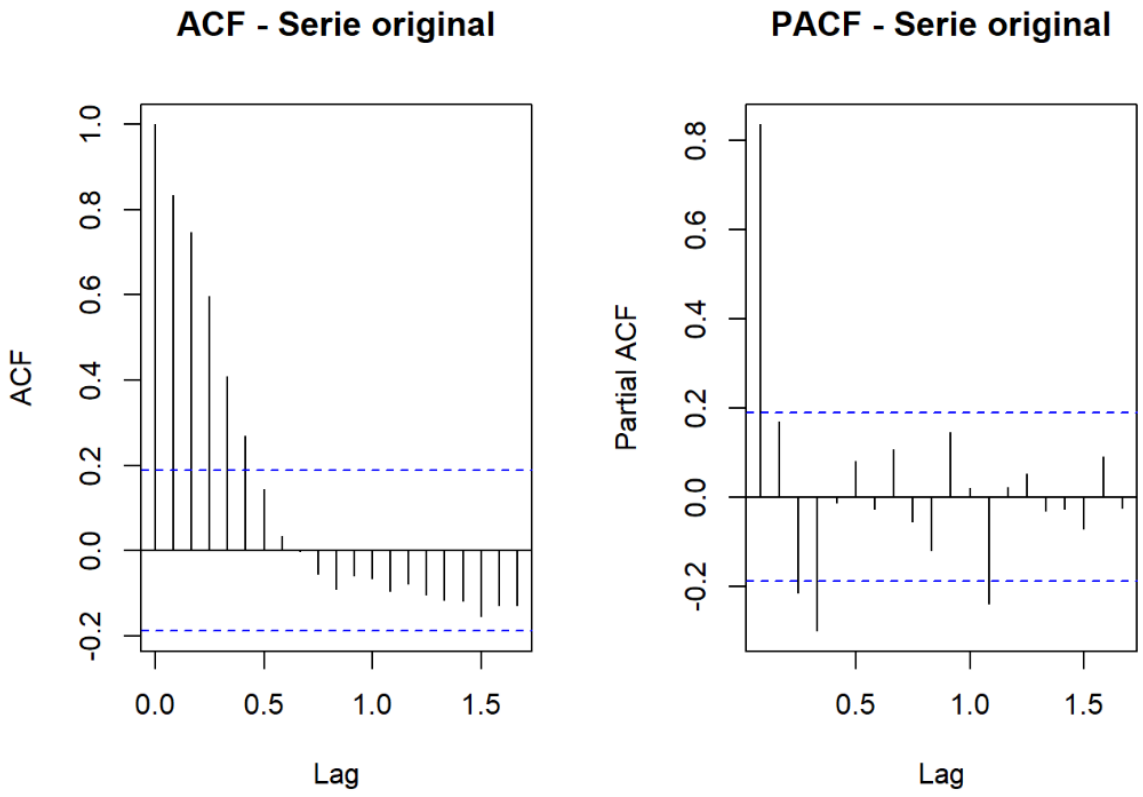

Figura 5.12: ACF y PACF de la serie temporal de casos positivos de malaria en Chocó (2015–2023)

La gráfica de la ACF muestra autocorrelaciones positivas y estadísticamente significativas en los primeros rezagos, las cuales decrecen de forma gradual hasta ubicarse dentro de los límites de confianza. Este comportamiento es característico de procesos autorregresivos y resulta coherente con una serie estacionaria.

Por su parte, la PACF presenta un rezago inicial claramente significativo, mientras que los rezagos posteriores permanecen, en su mayoría, dentro de los límites de confianza. Este patrón sugiere que la dependencia temporal de la serie se explica principalmente por el primer retardo.

En conjunto, los resultados de la prueba ADF y el análisis de las funciones ACF y PACF confirman que la serie presenta una estructura temporal estable y adecuada para la estimación de modelos de series de tiempo. En particular, la evidencia empírica sugiere como modelo candidato inicial un modelo autorregresivo de primer orden, correspondiente a un ARIMA(1,0,0), el cual será contrastado con modelos alternativos mediante criterios de información y análisis de residuos.

5.3.4. Estimación del modelo ARIMA

Una vez verificada la estacionariedad de la serie temporal mediante la prueba Dickey-Fuller aumentada y el análisis de las funciones ACF y PACF, se procedió a la estimación de modelos ARIMA utilizando el procedimiento automático auto.arima, el cual selecciona el modelo óptimo con base en criterios de información, principalmente AIC y AICc.

El modelo seleccionado fue un ARIMA(1,0,3) con media distinta de cero, lo cual es consistente con el resultado previo de estacionariedad (d=0)

5.3.5. Resultados del modelo estimado

5.3.5.1. Coeficientes estimados

El modelo ARIMA(1,0,3) presenta los siguientes coeficientes estimados:

- **Componente autorregresivo (AR(1)):**

  $$\hat{\phi}_1 = 0{,}7453 \quad (SE = 0{,}0926)$$

  Este coeficiente es positivo y de magnitud considerable, lo que indica una alta persistencia temporal; es decir, el valor actual de la serie depende en gran medida del valor observado en el período anterior.

- **Componentes de promedio móvil (MA):**

  $$\hat{\theta}_1 = -0{,}1027, \quad \hat{\theta}_2 = 0{,}4170, \quad \hat{\theta}_3 = 0{,}2313$$

  Estos términos capturan choques aleatorios de períodos anteriores que afectan el comportamiento actual de la serie, complementando la estructura autorregresiva identificada previamente.

- **Media de la serie:**

  $$\hat{\mu} = 2235{,}927$$

  La inclusión de una media distinta de cero es consistente con una serie estacionaria que fluctúa alrededor de un nivel promedio constante.

5.3.5.2. Medidas de ajuste

El modelo presenta una varianza del error estimada de:

$$\sigma^2 = 353\,312$$

y un valor de log-verosimilitud de:

$$\ell = -841{,}51$$

lo que indica un ajuste adecuado a los datos observados.

5.3.6. Criterios de selección del modelo

AIC: 1695.02
AICc: 1695.85
BIC: 1711.11

Estos valores respaldan la selección del modelo ARIMA(1,0,3), al lograr un equilibrio adecuado entre calidad de ajuste y parsimonia.

5.3.7. Discusión y conclusiones del modelamiento

Estos valores justifican la selección del modelo ARIMA(1,0,3) frente a otras especificaciones evaluadas, al lograr un equilibrio entre calidad de ajuste y parsimonia.

En conjunto, los resultados indican que la serie temporal presenta una dinámica explicada tanto por un componente autorregresivo dominante como por efectos de choques aleatorios de corto plazo, capturados por los términos de promedio móvil. La ausencia de diferenciación confirma que la serie es estacionaria en nivel, en concordancia con los resultados de la prueba ADF.

El modelo ARIMA(1,0,3) se considera, por tanto, una especificación adecuada para describir el comportamiento temporal de la serie y constituye una base sólida para la generación de pronósticos, los cuales deberán ser validados mediante el análisis de residuos y pruebas de diagnóstico adicionales.

5.3.8. Diagnóstico del modelo

5.3.8.1. Análisis de residuos

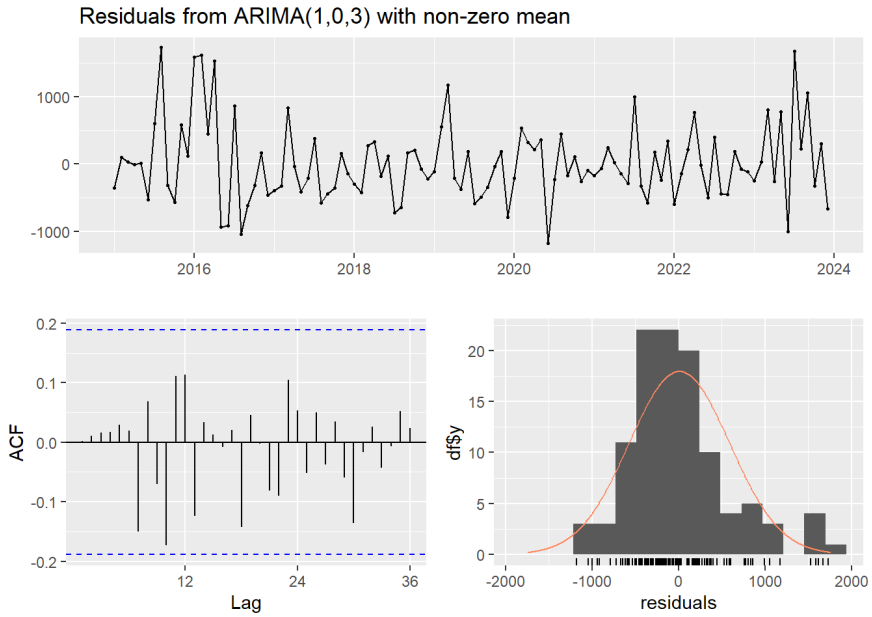

Figura 5.13: Análisis de residuos del modelo ARIMA(1,0,3) aplicado a la serie temporal de malaria en Chocó (2015–2023)

Con el objetivo de evaluar la adecuación del modelo ARIMA$(1, 0, 3)$, se realizó un análisis de los residuos mediante inspección gráfica, análisis de la función de autocorrelación y la aplicación de la prueba de Ljung-Box.

El gráfico de los residuos evidencia un comportamiento aleatorio alrededor de cero, sin patrones sistemáticos, lo que sugiere que los errores del modelo se comportan como ruido blanco. Asimismo, la función de autocorrelación (ACF) de los residuos no presenta picos estadísticamente significativos, indicando ausencia de dependencia temporal remanente.

5.3.8.2. Prueba de Ljung–Box

Adicionalmente, se aplicó la prueba de Ljung-Box a los residuos del modelo, obteniéndose un estadístico $Q^* = 17{,}949$ con 18 grados de libertad y un valor $p = 0{,}459$. Dado que el valor p es mayor que el nivel de significancia del 5 %, no se rechaza la hipótesis nula de ausencia de autocorrelación.

En consecuencia, se concluye que el modelo ARIMA$(1, 0, 3)$ captura adecuadamente la estructura temporal de la serie y presenta residuos con comportamiento de ruido blanco, por lo que resulta apropiado para el análisis y la generación de pronósticos.

5.3.9. Pronóstico de la Serie Temporal

5.3.9.1. Pronóstico a 12 meses

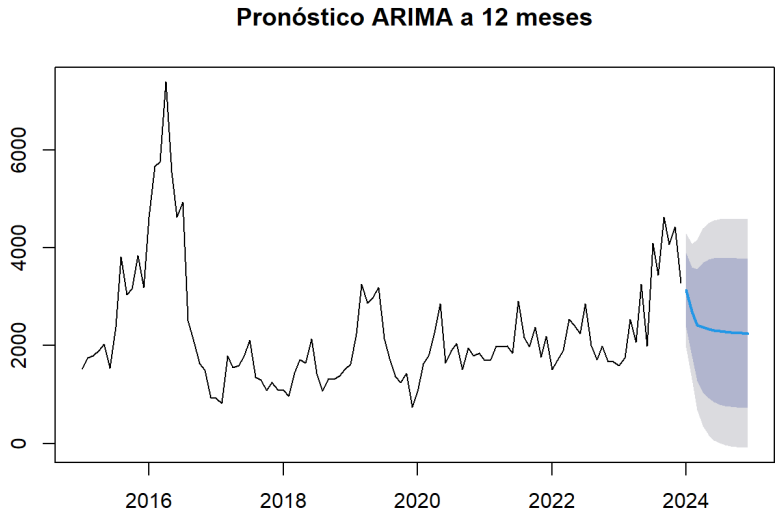

Figura 5.14: Pronóstico a 12 meses de los casos positivos de malaria en el departamento del Chocó mediante el modelo ARIMA(1,0,3)

La gráfica presenta la evolución histórica de la serie temporal junto con el pronóstico generado mediante un modelo ARIMA para un horizonte de 12 meses. La línea negra corresponde a los valores observados, mientras que la línea azul representa la predicción puntual del modelo para el período futuro. Las bandas sombreadas indican los intervalos de predicción, que reflejan el nivel de incertidumbre asociado a las estimaciones.

5.3.9.2. Tabla de pronósticos

Se observa que el pronóstico central sugiere una tendencia a la estabilización con leve descenso en los valores proyectados respecto a los últimos registros observados. No obstante, los intervalos de predicción se amplían progresivamente a medida que aumenta el horizonte temporal, lo cual evidencia una incertidumbre creciente en las estimaciones futuras, característica inherente a los modelos de series de tiempo.

Este comportamiento indica que, si bien el modelo proporciona una referencia útil sobre la posible evolución de la serie, las proyecciones deben interpretarse con cautela, especialmente en el mediano plazo. La información generada resulta pertinente para apoyar la planeación de escenarios, el fortalecimiento de acciones de vigilancia y monitoreo, y la toma de decisiones estratégicas,

considerando tanto la predicción puntual como los rangos de variabilidad esperados.

La Tabla 5.2 presenta los valores pronosticados para el período enero–diciembre de 2024, junto con sus respectivos intervalos de confianza al 80 % y 95 %, obtenidos a partir del modelo ARIMA$(1, 0, 3)$.

Tabla 5.2: Pronóstico mensual con intervalos de confianza (80 % y 95 %)

Periodo	Pronóstico	Lo 80	Hi 80	Lo 95	Hi 95
Ene 2024	3133.38	2371.62	3895.13	1968.37	4298.38
Feb 2024	2695.49	1790.00	3600.98	1310.66	4080.32
Mar 2024	2423.94	1290.03	3557.86	689.77	4158.11
Abr 2024	2376.05	1051.36	3700.74	350.12	4401.99
May 2024	2340.36	920.74	3759.99	169.23	4511.49
Jun 2024	2313.76	844.05	3783.48	66.03	4561.50
Jul 2024	2293.94	797.13	3790.75	4.76	4583.11
Ago 2024	2279.16	767.51	3790.82	-32.71	4591.04
Sep 2024	2268.15	748.32	3787.98	-56.24	4592.54
Oct 2024	2259.94	735.58	3784.30	-71.37	4591.25
Nov 2024	2253.83	726.96	3780.69	-81.32	4588.97
Dic 2024	2249.27	721.01	3777.53	-88.00	4586.54

5.3.10. Conclusión del análisis de series de tiempo

En conjunto, el análisis de series de tiempo permitió caracterizar la dinámica temporal de la malaria en el departamento del Chocó, evidenciando una serie estacionaria en nivel, con una fuerte dependencia autorregresiva y efectos de choques de corto plazo.

El modelo ARIMA(1,0,3) presentó un ajuste adecuado y cumplió satisfactoriamente los supuestos estadísticos, por lo que se considera una referencia robusta para la generación de pronósticos y para la comparación con modelos multivariados y de aprendizaje automático desarrollados en la siguiente sección.

5.4. Criterios para la selección de variables

5.4.1. Definición de variables climáticas

Incluir todas las variables climáticas medidas (temperatura mínima, media y máxima; humedad mínima, media y máxima; y precipitación) puede generar multicolinealidad, afectando la estabilidad

y la interpretabilidad de los modelos predictivos [62].

5.4.2. Matriz de Correlación

El análisis de la matriz de correlaciones constituye una herramienta fundamental en la etapa de exploración y selección de variables para modelos de series de tiempo, particularmente en contextos epidemiológicos donde los factores ambientales y sociodemográficos pueden presentar relaciones estructurales entre sí. La matriz de correlaciones permite evaluar de manera simultánea la intensidad y el sentido de la asociación lineal entre las variables explicativas y la variable respuesta, así como identificar posibles problemas de multicolinealidad entre predictores. En el presente estudio, este análisis resulta especialmente relevante debido a la alta dependencia esperada entre variables climáticas relacionadas, como las distintas medidas de temperatura, humedad y precipitación. La utilización de la matriz de correlaciones facilita la selección de un conjunto parsimonioso de variables, priorizando aquellas con mayor asociación empírica con la incidencia de malaria y evitando redundancias informativas que puedan afectar la estabilidad, interpretabilidad y capacidad de generalización de los modelos predictivos.

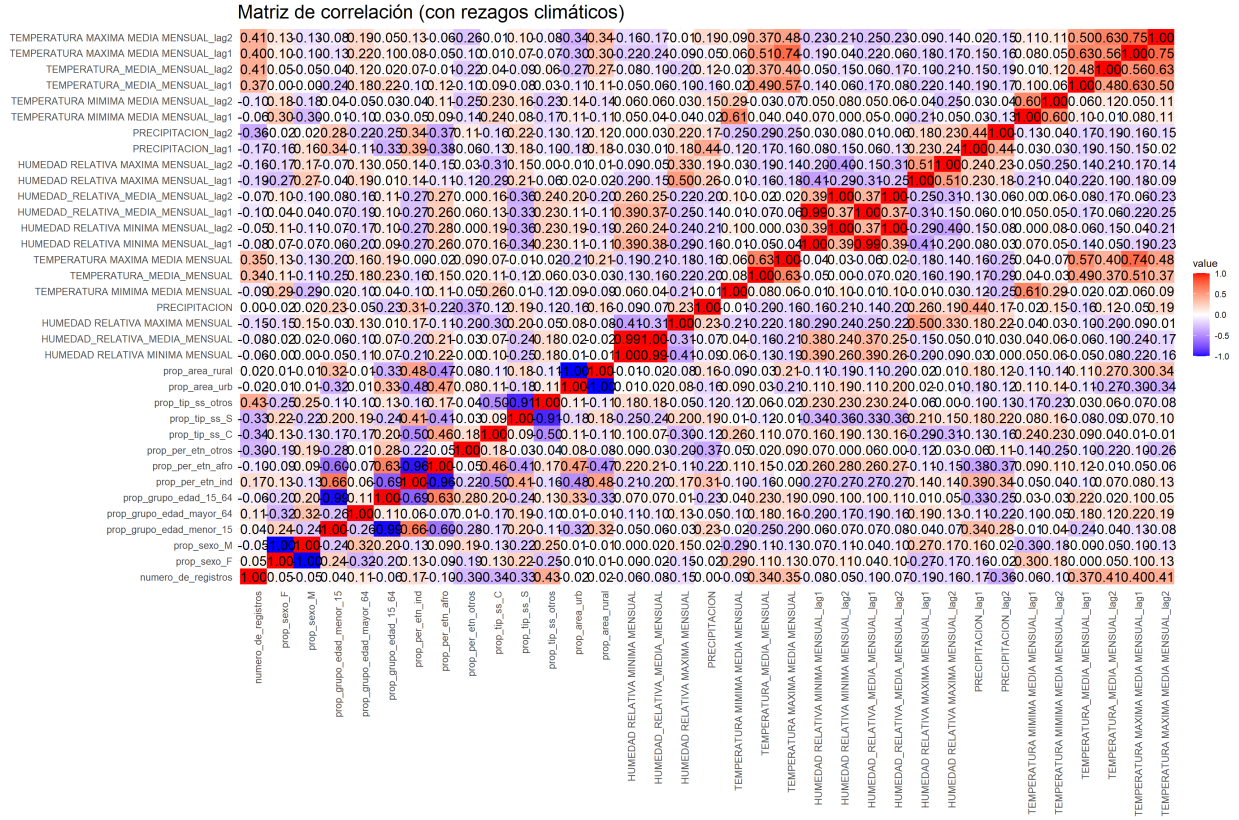

Figura 5.15: Matriz de correlación de Pearson entre el número de casos de malaria y las variables demográficas, territoriales y climáticas con rezagos temporales.

5.5. Selección de variables para el modelo predictivo

De acuerdo con los resultados obtenidos a partir de la matriz de correlación, se seleccionan las variables que harán parte de los modelos de análisis.

5.5.1. Variables climáticas

Las variables climáticas fueron analizadas tanto en su forma contemporánea como mediante rezagos temporales, con el propósito de evaluar su influencia sobre la dinámica de los casos de malaria en Colombia. En el contexto de la malaria, los efectos de las condiciones climáticas no suelen manifestarse de manera inmediata, sino a través de procesos biológicos y ambientales que introducen retardos temporales en la aparición de los casos. Incluir todas las variables climáticas medidas (temperatura mínima, media y máxima; humedad mínima, media y máxima; y precipitación) puede generar multicolinealidad, afectando la estabilidad y la interpretabilidad de los modelos predictivos.[62]

La temperatura presenta una asociación positiva moderada con el número de registros de malaria, especialmente cuando se consideran rezagos temporales, siendo la temperatura media mensual rezagada la que muestra una relación más consistente con la variable respuesta. Este comportamiento sugiere un efecto retardado de la temperatura sobre la incidencia de la enfermedad, lo cual es epidemiológicamente plausible, dado que influye en el desarrollo del mosquito vector y del parásito, procesos que requieren un periodo de maduración antes de reflejarse en un aumento de casos humanos. La alta correlación observada entre las distintas medidas de temperatura (mínima, media y máxima) evidencia una dependencia estructural entre estos predictores, lo que justifica la selección de una única variable representativa para evitar problemas de multicolinealidad y preservar la parsimonia del modelo.

La precipitación muestra correlaciones negativas de magnitud baja a moderada con el número de registros de malaria, particularmente en los rezagos de uno y dos meses, lo que indica que episodios de lluvia intensa podrían estar asociados a una disminución posterior de los casos. Este patrón puede interpretarse como el resultado de la alteración o eliminación de criaderos del mosquito causada por excesos de precipitación, reflejando la naturaleza no lineal del efecto de la lluvia sobre la transmisión de la malaria. La mayor relevancia de los rezagos temporales resalta la importancia de incorporar retardos en el análisis, ya que los efectos de la precipitación sobre la dinámica vectorial no se manifiestan de forma inmediata.

La humedad relativa presenta correlaciones débiles con el número de registros de malaria, tanto en su forma contemporánea como rezagada, lo que sugiere que su variabilidad mensual no exhibe una relación lineal fuerte con la incidencia de la enfermedad en el periodo analizado. No obstante, desde un punto de vista epidemiológico, la humedad desempeña un papel complementario en la supervivencia y longevidad del mosquito vector, por lo que su inclusión puede ser relevante como

factor modulador dentro de modelos multivariados. La alta correlación interna entre las distintas medidas de humedad relativa respalda la selección de una única variable representativa, con el fin de evitar redundancias informativas y mejorar la estabilidad del modelo.

Así pues las variables seleccionadas de acuerdo con el coeficiente de correlación son:

- **Temperatura:** `TEMPERATURA_MEDIA_MENSUAL_lag2` (2 rezagos)
- **Precipitación:** `PRECIPITACION_lag2` (2 rezagos)
- **Humedad relativa:** `HUMEDAD_RELATIVA_MAXIMA_MENSUAL_lag1` (1 rezago)

5.5.2. Variables socioeconómicas

Las variables socioeconómicas, expresadas como proporciones, presentan colinealidad perfecta o cuasi perfecta, dado que corresponden a particiones de una misma población cuya suma es igual a uno.

Por esta razón, no se incluyen simultáneamente en los modelos, y se selecciona únicamente una variable representativa por grupo con el fin de evitar problemas de identificación y estimación.

Las variables socioeconómicas seleccionadas son:

- **Grupo de edad:** `prop_grupo_edad_15_64`
- **Área de residencia:** `prop_area_rural`
- **Seguridad social:** `prop_tip_ss_S`
- **Grupo étnico:** `prop_per_etn_afro`

Las variables de sexo fueron igualmente evaluadas. A pesar de presentar una baja correlación con la variable respuesta y de la colinealidad perfecta existente entre las proporciones de hombres y mujeres, se decide incluir una única variable representativa para capturar posibles diferencias asociadas al sexo:

- **Sexo:** `prop_sexo_M`

Se opta en conservar el conjunto completo de variables climáticas rezagadas y sociodemográficas seleccionadas, permitiendo que los modelos de aprendizaje automático determinen de manera automática la relevancia relativa de cada predictor, minimizando así decisiones subjetivas en la etapa de selección.

5.6. Modelos de Aprendizaje Automático

5.6.1. Modelos descartados

La decisión de excluir determinados enfoques de aprendizaje automático en el presente estudio se sustenta en una evaluación crítica de su pertinencia frente a la naturaleza epidemiológica de la malaria en Colombia, la estructura temporal de los datos disponibles y los objetivos analíticos planteados. En este proceso, se priorizan criterios de estabilidad estadística, coherencia metodológica e interpretabilidad, de acuerdo con las recomendaciones establecidas en la literatura especializada en ciencia de datos aplicada [62].

En particular, se descarta el uso de arquitecturas de redes neuronales recurrentes de alta complejidad, tales como modelos apilados con múltiples capas o arquitecturas híbridas profundas. Si bien estos enfoques pueden capturar relaciones altamente complejas, requieren volúmenes de datos considera blemente mayores para alcanzar una adecuada capacidad de generalización. En el contexto de series epidemiológicas mensuales de longitud moderada, su utilización incrementa el riesgo de sobreajuste y reduce la estabilidad de las estimaciones. Por esta razón, se opta por una arquitectura recurrente parsimoniosa, adecuada al tamaño muestral disponible y coherente con los objetivos aplicados e interpretativos del estudio.

En relación con el tipo de variable respuesta analizada, se descartan modelos de clasificación probabilística, como Naive Bayes y métodos afines, dado que el fenómeno de interés se expresa mediante conteos de casos y no a través de categorías discretas. La aplicación de enfoques clasificatorios habría requerido una transformación artificial de la variable dependiente, con la consecuente pérdida de información cuantitativa y una interpretación epidemiológica poco adecuada para los fines del estudio [62].

Por otra parte, se decide no incorporar métodos de aprendizaje no supervisado, tales como técnicas de agrupamiento o reducción de dimensionalidad, dado que el propósito del estudio no se centra en la identificación de estructuras latentes sin referencia a una variable respuesta, sino en la modelación explícita de la relación entre factores climáticos y la incidencia de malaria. En este sentido, dichos enfoques no resultan pertinentes para responder a las preguntas de investigación formuladas.

Asimismo, los métodos basados en proximidad local, como el k-Nearest Neighbors en su formulación estándar, fueron excluidos debido a su sensibilidad a la escala de las variables y a la ausencia de mecanismos explícitos para incorporar la dependencia temporal de los datos. Esta limitación es particularmente relevante en el análisis de enfermedades transmitidas por vectores, donde los efectos de las condiciones climáticas suelen manifestarse con retardos temporales que no pueden ser capturados adecuadamente por este tipo de modelos (WHO, 2017).

En síntesis, la exclusión de estos enfoques responde a una estrategia metodológica orientada a privilegiar modelos que ofrezcan un balance adecuado entre capacidad predictiva, estabilidad estadística e interpretabilidad, asegurando resultados coherentes con la naturaleza epidemiológica del fenómeno analizado y con los objetivos del estudio.

5.6.2. Modelos considerados

La selección de los modelos de aprendizaje automático utilizados en este estudio se basa en la necesidad de comprender adecuadamente la relación entre las variables climáticas y sociodemográficas consideradas y el comportamiento de los registros de malaria en Colombia, garantizando un equilibrio entre capacidad predictiva, robustez estadística e interpretabilidad. Para ello, se priorizan enfoques capaces de modelar relaciones no lineales y posibles interacciones complejas entre los predictores, características habituales en fenómenos epidemiológicos influenciados por factores ambientales y sociales [62].

En primer lugar, se considera una Red Neuronal Recurrente de tipo Long Short-Term Memory (LSTM) teniendo en cuenta la naturaleza temporal de los registros de malaria y en la necesidad de modelar explícitamente las dependencias dinámicas presentes en la serie epidemiológica. Las arquitecturas LSTM permiten capturar patrones temporales de corto y mediano plazo y representar de forma más adecuada la evolución de la incidencia de la malaria. La implementación del modelo se realiza bajo una arquitectura parsimoniosa y regularizada, acorde con el tamaño muestral disponible y los objetivos aplicados del estudio, garantizando un equilibrio entre capacidad predictiva, estabilidad estadística e interpretabilidad en un contexto de análisis epidemiológico.

Como modelo base interpretable, se incorpora un Árbol de Decisión, el cual permite segmentar el espacio de predictores en regiones homogéneas y generar reglas de decisión fácilmente comprensibles. Este enfoque resulta útil tanto como herramienta exploratoria como punto de comparación metodológico, al ofrecer una representación clara de la forma en que las variables explicativas influyen sobre los registros de malaria. Sin embargo, su tendencia al sobreajuste y su limitada capacidad predictiva en escenarios complejos justifican su uso principalmente como modelo de referencia.

Adicionalmente, se emplea el algoritmo de Random Forest, un método de ensamble basado en la agregación de múltiples árboles de decisión construidos sobre muestras bootstrap y subconjuntos aleatorios de variables. Este enfoque permite reducir la varianza del estimador y mejorar la estabilidad de las predicciones, lo que resulta particularmente relevante en el análisis de datos ambientales y epidemiológicos caracterizados por alta variabilidad. Asimismo, Random Forest proporciona medidas de importancia de variables, lo que aporta información complementaria para la interpretación de los factores asociados a la malaria, aun cuando sacrifica interpretabilidad directa frente a un árbol individual.

Finalmente, se incluye un modelo de Gradient Boosting, representativo de los métodos aditivos secuenciales de alto desempeño predictivo. Este enfoque ajusta de manera iterativa modelos débiles, corrigiendo progresivamente los errores cometidos en etapas previas, lo que favorece la captura de estructuras complejas en los datos. Su incorporación permite evaluar el impacto del boosting sobre la precisión del modelo, reconociendo al mismo tiempo su sensibilidad al ajuste de hiperparámetros y su menor transparencia interpretativa.

En conjunto, la selección de estos cuatro modelos permite abordar el problema de estudio desde perspectivas metodológicas complementarias: desde enfoques altamente flexibles hasta modelos interpretables y métodos de ensamble robustos. Esta estrategia facilita una comparación integral entre desempeño predictivo y consideraciones prácticas como interpretabilidad y estabilidad, alineándose con las buenas prácticas en ciencia de datos aplicada a la salud pública y con los lineamientos para el análisis epidemiológico de enfermedades transmisibles.[62]

5.6.2.1. Estrategia de modelación y validación

Para soportar la determinación objetiva del mejor modelo, se definió una estrategia de modelación y validación basada en un esquema de partición temporal del conjunto de datos.La implementación de dicha estrategia se desarrolló mediante la siguiente rutina:

1. Dividir el conjunto original de datos en dos partes, respetando el orden cronológico: un subconjunto correspondiente al 80 % de los registros, destinado al entrenamiento de los modelos, y el 20 % restante, correspondiente a los periodos más recientes, reservado para la evaluación fuera de muestra.
2. Entrenar cada modelo utilizando exclusivamente el subconjunto de entrenamiento.
3. Evaluar el desempeño de los modelos mediante el conjunto de prueba (datos fuera de muestra), calculando las métricas correspondientes.
4. Comparar los resultados obtenidos y seleccionar el modelo con mejor desempeño según los criterios definidos.
5. Confirmar la capacidad de generalización del modelo seleccionado mediante el análisis de los resultados obtenidos en el conjunto reservado para validación final.

5.6.3. Modelos de aprendizaje automático

Previo a la especificación de cada uno de los modelos considerados, se define un esquema común de entrenamiento y evaluación basado en una partición temporal de la serie epidemiológica. En particular, los primeros periodos observados se utilizan para el entrenamiento de los modelos, mientras que los periodos más recientes se reservan para la evaluación fuera de muestra. Este criterio se aplica de manera consistente a todos los modelos de aprendizaje automático evaluados, con el fin de

evitar fuga de información y garantizar una comparación homogénea del desempeño predictivo en un contexto temporal.

La configuración de los hiperparámetros de los modelos de aprendizaje automático se realiza mediante un procedimiento sistemático de búsqueda, evaluando diferentes combinaciones de parámetros sobre el conjunto de entrenamiento. La selección final se basa en la minimización de métricas de error predictivo, bajo un esquema de validación coherente con la estructura temporal de la serie. Este enfoque permite reducir decisiones arbitrarias, mejorar la estabilidad de los modelos y garantizar una comparación metodológicamente consistente entre los distintos algoritmos considerados.

5.6.4. Red Neuronal Recurrente Long Short-Term Memory (LSTM)

La inclusión de una red neuronal recurrente de tipo Long Short-Term Memory (LSTM) en este estudio se fundamenta en la necesidad de modelar explícitamente la dependencia temporal presente en los registros de malaria en Colombia. A diferencia de las redes neuronales feedforward, las arquitecturas recurrentes están diseñadas para procesar secuencias ordenadas, permitiendo capturar patrones dinámicos y efectos retardados que son inherentes a los fenómenos epidemiológicos.

La transmisión de la malaria es el resultado de procesos biológicos y ambientales complejos, en los cuales las condiciones climáticas y sociodemográficas no impactan de manera inmediata sobre la incidencia de la enfermedad, sino a través de mecanismos acumulativos y retardados. En este contexto, las redes LSTM resultan particularmente adecuadas, ya que incorporan mecanismos de memoria que permiten retener información relevante de periodos anteriores y modelar relaciones no lineales dependientes del tiempo, superando las limitaciones de los modelos neuronales no secuenciales.

No obstante, considerando el tamaño muestral disponible y el enfoque aplicado del estudio, se descarta el uso de arquitecturas recurrentes profundas o con múltiples capas apiladas, optando por una red LSTM de arquitectura simple y regularizada. De este modo, el modelo LSTM se emplea como un enfoque flexible y especializado en series de tiempo, manteniendo criterios de parsimonia, estabilidad estadística e interpretabilidad acordes con estudios epidemiológicos en salud pública.

5.6.4.1. Formulación estadística del modelo

Sea y_t el número de casos de malaria en el periodo t, y sea $\mathbf{X}_t = (x_{t1}, x_{t2}, \ldots, x_{tp})$ el vector de predictores climáticos rezagados y variables sociodemográficas observadas en el mismo periodo.

Una red neuronal recurrente del tipo *Long Short-Term Memory* (LSTM) modela la relación entre la secuencia de predictores $\{\mathbf{X}_{t-\ell}, \ldots, \mathbf{X}_t\}$ y la variable respuesta y_t mediante un sistema de estados internos con memoria, definido por las siguientes ecuaciones:

$$\mathbf{f}_t = \sigma\left(\mathbf{W}_f[\mathbf{h}_{t-1}, \mathbf{X}_t] + \mathbf{b}_f\right), \tag{5.1}$$
$$\mathbf{i}_t = \sigma\left(\mathbf{W}_i[\mathbf{h}_{t-1}, \mathbf{X}_t] + \mathbf{b}_i\right), \tag{5.2}$$
$$\tilde{\mathbf{c}}_t = \tanh\left(\mathbf{W}_c[\mathbf{h}_{t-1}, \mathbf{X}_t] + \mathbf{b}_c\right), \tag{5.3}$$
$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tilde{\mathbf{c}}_t, \tag{5.4}$$
$$\mathbf{o}_t = \sigma\left(\mathbf{W}_o[\mathbf{h}_{t-1}, \mathbf{X}_t] + \mathbf{b}_o\right), \tag{5.5}$$
$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t). \tag{5.6}$$

Donde $\mathbf{f}_t$, $\mathbf{i}_t$ y $\mathbf{o}_t$ representan las compuertas de olvido, entrada y salida, respectivamente; $\mathbf{c}_t$ es el estado de memoria interna de la celda; $\mathbf{h}_t$ es el estado oculto de la red en el periodo t; $\mathbf{W}.$ y $\mathbf{b}.$ son matrices de pesos y términos de sesgo; $\sigma(\cdot)$ es la función sigmoide logística y $\odot$ denota el producto elemento a elemento.

La predicción del número de casos se obtiene a partir del estado oculto final:

$$\hat{y}_t = \beta_0 + \boldsymbol{\beta}^\top \mathbf{h}_t. \tag{5.7}$$

El modelo se estima minimizando una función de pérdida de tipo error cuadrático medio:

$$\mathcal{L} = \frac{1}{n}\sum_{t=1}^{n}\left(y_t - \hat{y}_t\right)^2. \tag{5.8}$$

mediante algoritmos de optimización basados en gradiente, como Adam (Adaptive Moment Estimation), el cual combina adaptativamente información del primer y segundo momento del gradiente para mejorar la estabilidad y velocidad de convergencia.

5.6.4.2. Preprocesamiento de los datos

Previo al entrenamiento del modelo, los predictores son centrados y escalados para garantizar estabilidad numérica y favorecer la convergencia del algoritmo de optimización. Posteriormente, los datos son reorganizados en formato secuencial, construyendo ventanas temporales de longitud fija que preservan el orden cronológico de la serie.

La partición de los datos se realiza bajo un esquema temporal 80/20, donde el conjunto de entrenamiento corresponde a los primeros periodos de la serie y el conjunto de prueba incluye los periodos más recientes. Este procedimiento permite evaluar la capacidad predictiva del modelo en un escenario realista de pronóstico, evitando fugas de información temporal.

5.6.4.3. Arquitectura y configuración del modelo

Dado el tamaño muestral y la longitud de la serie temporal, se adopta una arquitectura parsimoniosa compuesta por una única capa LSTM con un número reducido de unidades ocultas, seguida de una capa densa con salida lineal. Esta configuración permite capturar dependencias temporales relevantes sin incurrir en una complejidad excesiva que pueda comprometer la generalización del modelo.

Adicionalmente, se incorpora regularización mediante dropout y control de la magnitud de los pesos, con el fin de mitigar el sobreajuste y mejorar la estabilidad del estimador.

5.6.4.4. Arquitectura y ajuste de hiperparámetros

El modelo de Red Neuronal Recurrente Long Short-Term Memory (LSTM) se especifica con el propósito de capturar las dependencias temporales presentes en la serie de tiempo correspondiente al número de registros de malaria. Esta arquitectura resulta adecuada para el análisis de series temporales, dado que incorpora mecanismos de memoria interna que permiten modelar efectos acumulativos y retardados. La configuración empleada consiste en una capa LSTM con un número fijo de unidades ocultas, seguida de una capa densa de salida con activación lineal, apropiada para un problema de regresión. La elección de una arquitectura parsimoniosa responde a la longitud moderada de la serie analizada y a la necesidad de controlar el riesgo de sobreajuste, priorizando la estabilidad del entrenamiento y la capacidad de generalización del modelo.

Desde el punto de vista funcional, el modelo LSTM implementa transformaciones no lineales mediante funciones de activación internas, principalmente funciones sigmoides y la función tangente hiperbólica (tanh), lo que lo clasifica como un método no paramétrico y no lineal. El entrenamiento se realiza mediante la minimización del error cuadrático medio, utilizando un optimizador de tipo gradiente adaptativo, y el ajuste de los principales hiperparámetros —como el número de unidades, la tasa de dropout, el tamaño del lote y el número de épocas— se define de manera empírica e informada. Dado el carácter aplicado del estudio y el tamaño limitado del conjunto de datos, no se implementa un procedimiento exhaustivo de búsqueda de hiperparámetros, decisión coherente con las recomendaciones metodológicas para evitar el sobreajuste sin mejoras sustanciales en el desempeño predictivo fuera de muestra[62].

La configuración de los hiperparámetros empleados en el entrenamiento del modelo LSTM se definió de la siguiente manera:

Listing 5.1: Hiperparámetros del entrenamiento del modelo LSTM

```
history <- lstm_model$fit(
  x = X_train,
  y = y_train,
  epochs = as.integer(100),
  batch_size = as.integer(8),
  validation_split = 0.2,
  verbose = as.integer(0)
)
```

- **x = X_train**: Datos de entrada utilizados para entrenar el modelo. En este estudio corresponden a las secuencias temporales estructuradas para alimentar la red LSTM.
- **y = y_train**: Valores objetivo asociados a cada secuencia de entrenamiento, correspondientes a la variable respuesta que se desea predecir.
- **epochs = 100**: Número de iteraciones completas sobre el conjunto de entrenamiento. El modelo procesa la totalidad de los datos en 100 ocasiones con el fin de optimizar los parámetros internos.
- **batch_size = 8**: Tamaño del lote utilizado durante el entrenamiento. Los pesos del modelo se actualizan cada 8 observaciones, lo que favorece estabilidad numérica y eficiencia computacional.
- **validation_split = 0.2**: Proporción de los datos de entrenamiento reservada para validación interna. El 20 % de las observaciones se utilizan para evaluar el desempeño del modelo en cada época, sin participar en la actualización de los pesos.
- **verbose = 0**: Parámetro que controla la salida en consola durante el entrenamiento. Un valor de 0 indica que no se mostrará información del progreso del proceso iterativo.

5.6.4.5. Estrategia de entrenamiento y regularización

El entrenamiento del modelo LSTM se realiza de manera iterativa mediante retropropagación a través del tiempo (Backpropagation Through Time), minimizando el error cuadrático medio. Se emplea un optimizador adaptativo y un número limitado de épocas, junto con mecanismos de regularización, para controlar la varianza del modelo y evitar un ajuste excesivo a los datos de entrenamiento.

El número de unidades LSTM, la tasa de dropout y el número de épocas han sido seleccionados mediante búsqueda en grilla sobre el conjunto de entrenamiento.

5.6.4.6. Evaluación del desempeño predictivo

El desempeño predictivo del modelo de Red Neuronal Recurrente tipo *Long Short-Term Memory* (LSTM) se evaluó sobre el conjunto de prueba, correspondiente al 20 % más reciente de la serie temporal, utilizando métricas calculadas en la escala original de la variable de interés. En particular, se estimaron la raíz del error cuadrático medio (RMSE), el error absoluto medio (MAE) y el coeficiente de determinación R^2, con el fin de evaluar tanto la magnitud del error como la capacidad explicativa del modelo.

Tabla 5.3: Desempeño predictivo del modelo LSTM

Modelo	RMSE	MAE	R^2
Red Neuronal Recurrente LSTM	1150.57	903.94	-0.47

El modelo de Red Neuronal Recurrente *Long Short-Term Memory* (LSTM) presenta valores elevados de error *RMSE* y *MAE*, junto con un coeficiente de determinación R^2 negativo, lo cual indica que su desempeño predictivo es inferior al de un modelo base que simplemente predice el valor promedio de la serie. A pesar de su idoneidad teórica para el análisis de series de tiempo, los resultados sugieren que el modelo no logra capturar adecuadamente la dinámica de los registros de malaria en el conjunto de datos analizado. Este comportamiento puede atribuirse a la longitud moderada de la serie, que limita la capacidad de generalización de arquitecturas recurrentes, así como a la sensibilidad del modelo a la especificación de hiperparámetros. En consecuencia, el LSTM no aporta mejoras predictivas sustanciales en este contexto y su utilidad se restringe a un análisis metodológico comparativo.

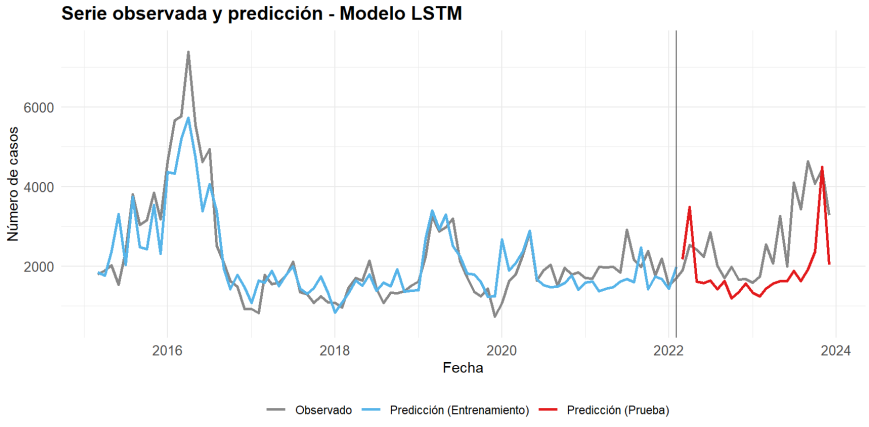

Figura 5.16: Serie observada y predicha del número de casos de malaria mediante el modelo LSTM

5.6.5. Árbol de Decisión

El árbol de decisión se incluye en el análisis debido a su alta interpretabilidad y a su capacidad para identificar umbrales críticos en variables climáticas y sociodemográficas asociadas a la transmisión de la malaria. Este tipo de modelo resulta especialmente útil en estudios epidemiológicos, ya que permite traducir relaciones complejas en reglas de decisión fácilmente comprensibles, facilitando la interpretación desde una perspectiva de salud pública.

5.6.5.1. Fundamentación estadística

El árbol de decisión particiona recursivamente el espacio de predictores en regiones disjuntas, de tal forma que la respuesta dentro de cada región sea lo más homogénea posible. Formalmente, el modelo puede expresarse como:

$$\hat{y}_t = \sum_{m=1}^{M} c_m \, \mathbb{I}(\mathbf{x}_t \in R_m),$$

donde:

- $\mathbb{I}(\cdot)$ es la función indicadora,
- $\mathbf{x}_t \in R_m$ indica que la observación t pertenece a la región terminal R_m,
- M es el número de regiones terminales,
- c_m representa el valor constante asignado a la región R_m, correspondiente al promedio de la variable respuesta dentro de dicha región.

El criterio de partición se basa en la minimización de la varianza intra-nodo, lo que permite identificar divisiones que expliquen de manera eficiente la variabilidad observada en los registros de malaria.

5.6.5.2. Preprocesamiento de los datos

A diferencia de otros modelos de aprendizaje automático, el árbol de decisión no requiere estandarización de las variables explicativas, ya que las particiones se realizan de forma independiente de la escala de los predictores. Se emplea el mismo esquema de partición temporal de los datos utilizado en los modelos anteriores, respetando el orden cronológico de la serie epidemiológica.

5.6.5.3. Configuración del modelo

La complejidad del árbol se controla mediante el parámetro de complejidad (*complexity parameter*, *cp*), el cual penaliza la inclusión de nuevas particiones que no aportan mejoras sustanciales en la reducción del error. Este control resulta fundamental para mitigar el sobreajuste y mejorar la capacidad de generalización del modelo.

la implementación computacional del modelo se realizo utilizando la función `rpart()`, especificando un enfoque de regresión mediante el método `.ªnova"` y fijando el parámetro de complejidad en $cp = 0{,}01$, como se muestra a continuación:

Listing 5.2: Configuración del modelo de árbol de decisión

```
tree_model <- rpart(
  numero_de_registros ~ .,
  data = df_modelo |> dplyr::select(-fecha),
  method = "anova",
  control = rpart.control(cp = 0.01)
)
```

5.6.5.4. Arquitectura y ajuste de hiperparámetros

El modelo de Árbol de Decisión se emplea como un enfoque interpretable para modelar la relación entre las variables explicativas y el número de registros de malaria. Este tipo de modelo segmenta el espacio de predictores mediante reglas jerárquicas, permitiendo identificar de forma explícita los umbrales y condiciones bajo los cuales las variables influyen sobre la variable respuesta. Su inclusión resulta pertinente como modelo base, dada su facilidad de interpretación y su utilidad para el análisis exploratorio de relaciones no lineales.

El ajuste del modelo se realiza mediante la minimización del error cuadrático medio, controlando su complejidad a través del parámetro de poda (complexity parameter), el cual limita el crecimiento del árbol y reduce el riesgo de sobreajuste. La selección de este parámetro se define de manera empírica e informada, considerando la estabilidad del modelo y su capacidad de generalización, en coherencia con las recomendaciones metodológicas para modelos basados en particiones recursivas [62].

Para el modelo de Árbol de Decisión, los hiperparámetros utilizados fueron los siguientes:

- **cp (complexity parameter) = 0.01**: Define el umbral mínimo de reducción de la desviación necesario para dividir un nodo. Este valor permite equilibrar la complejidad del árbol y la capacidad de generalización, evitando sobreajuste.
- **method = .ªnova"**: Se emplea para problemas de regresión, minimizando la suma de cuadrados dentro de los nodos.

Estos hiperparámetros aseguran que el árbol capture las relaciones relevantes entre las variables explicativas y la variable respuesta, manteniendo un nivel adecuado de interpretabilidad y generalización.

5.6.5.5. Estrategia de entrenamiento

El árbol se ajusta minimizando la varianza intra-nodo.

5.6.5.6. Evaluación del desempeño predictivo

El desempeño predictivo de los modelos se evaluó utilizando un esquema de validación temporal, en el cual el 80 % inicial de la serie se destinó al entrenamiento y el 20 % más reciente se reservó como conjunto de prueba. Este enfoque permite evaluar la capacidad de los modelos para generalizar y producir predicciones en periodos no observados, respetando la estructura temporal de los datos.

La evaluación se realizó mediante los siguientes parámetros de desempeño calculados sobre el conjunto de prueba y en la escala original de la variable de interés: la raíz del error cuadrático medio (RMSE), el error absoluto medio (MAE) y el coeficiente de determinación (R^2). El RMSE penaliza de forma más severa los errores grandes, el MAE mide el error promedio absoluto, y el coeficiente R^2 cuantifica la proporción de la variabilidad de la serie que es explicada por el modelo.

Los resultados obtenidos para el modelo de Árbol de Decisión se presentan en la siguiente Tabla:

Tabla 5.4: Desempeño predictivo del modelo de Árbol de Decisión

Modelo	RMSE	MAE	R^2
Árbol de Decisión	741.87	549.29	0.39

El modelo de Árbol de Decisión muestra un desempeño notablemente superior al de la red neuronal recurrente, con errores considerablemente menores (RMSE = 741.87 y MAE = 549.29) y un coeficiente de determinación positivo (R^2 = 0,39). Estos resultados indican que el modelo logra explicar una proporción moderada de la variabilidad observada en los registros de malaria, capturando relaciones relevantes entre las variables explicativas y la variable respuesta.

No obstante, el valor de R^2 sugiere que una fracción importante de la variabilidad permanece sin explicar, lo cual es consistente con las limitaciones conocidas de los árboles de decisión individuales para modelar estructuras complejas y dependencias temporales. A pesar de ello, su alto nivel de interpretabilidad y su desempeño intermedio lo convierten en una herramienta exploratoria valiosa dentro del análisis.

5.6.6. Modelo Random Forest

El algoritmo Random Forest se emplea con el objetivo de reducir la varianza asociada a un árbol de decisión individual y capturar relaciones no lineales complejas entre las variables explicativas y la variable respuesta. Este enfoque resulta particularmente adecuado en contextos epidemiológicos caracterizados por alta variabilidad temporal y patrones no estacionarios, como es el caso de la malaria en Colombia.

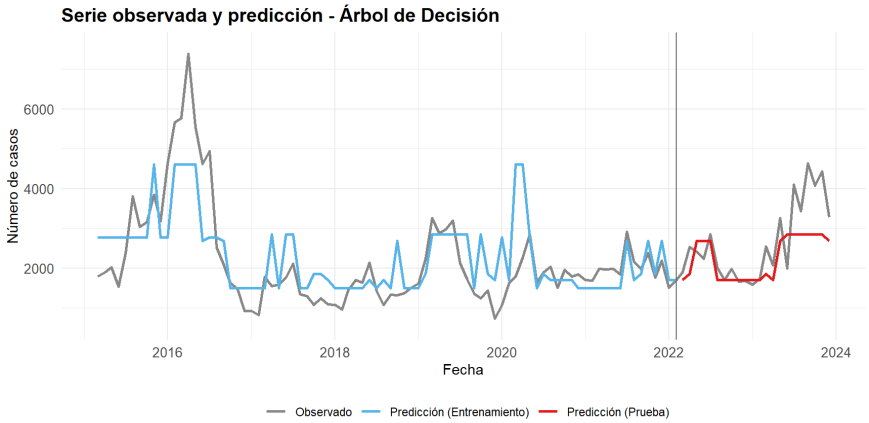

Figura 5.17: Serie observada y predicha del número de casos de malaria utilizando un Árbol de Decisión

5.6.6.1. Fundamentación estadística

Random Forest se basa en un esquema de ensamble mediante *bootstrap aggregation* (*bagging*), en el cual múltiples árboles de decisión son entrenados sobre muestras bootstrap del conjunto de datos. La predicción final se obtiene como el promedio de las predicciones individuales de cada árbol, lo que puede expresarse como:

$$\hat{y}_t = \frac{1}{B} \sum_{b=1}^{B} \hat{y}_t^{(b)}, \tag{5.9}$$

donde B representa el número total de árboles del ensamble y $\hat{y}_t^{(b)}$ corresponde a la predicción generada por el árbol b en el periodo t. Este procedimiento permite reducir significativamente la varianza del modelo sin incrementar de forma sustancial el sesgo.

5.6.6.2. Preprocesamiento de los datos

A diferencia de otros enfoques de modelación, Random Forest no requiere el escalamiento de las variables explicativas, dado que las particiones se realizan a partir de reglas basadas en umbrales. Asimismo, se mantuvo la estructura temporal de los datos, respetando la secuencia cronológica de las observaciones para evitar filtraciones de información futura.

5.6.6.3. Configuración del modelo

El modelo fue configurado utilizando un número elevado de árboles y un subconjunto aleatorio de predictores en cada división, con el propósito de aumentar la diversidad entre los árboles y mejorar la capacidad predictiva del ensamble. Esta aleatorización contribuye a reducir la correlación entre los árboles individuales y a fortalecer la robustez del modelo.

5.6.6.4. Arquitectura y ajuste de hiperparámetros

El modelo Random Forest se implementa como un método de ensamble basado en la agregación de múltiples árboles de decisión construidos a partir de muestras bootstrap y subconjuntos aleatorios de variables. Este enfoque permite capturar relaciones no lineales complejas y reducir la varianza del estimador, mejorando la estabilidad de las predicciones frente a un árbol individual. Su aplicación resulta adecuada en contextos donde se busca un equilibrio entre capacidad predictiva y robustez del modelo.

El ajuste del Random Forest se controla principalmente mediante el número de árboles y el número de variables consideradas en cada partición. Estos hiperparámetros se definen de forma empírica, atendiendo a criterios de estabilidad del error y consistencia de las predicciones. Dado el carácter aplicado del estudio, no se realiza una búsqueda exhaustiva de hiperparámetros, priorizando configuraciones que ofrezcan buen desempeño predictivo sin incrementar innecesariamente la complejidad del modelo[62].

Los hiperparámetros utilizados en el ajuste del modelo fueron:

- **ntree = 500**: Número de árboles a construir en el ensamble. Un valor mayor mejora la estabilidad de las predicciones y reduce la varianza del modelo.
- **importance = TRUE**: Activa el cálculo de la importancia de cada variable, permitiendo identificar los predictores más relevantes.

La configuración seleccionada permite que el modelo capture patrones complejos en los datos sin generar sobreajuste, facilitando además la interpretación mediante la importancia de las variables y la comparación entre valores observados y predichos.

5.6.6.5. Evaluación del desempeño predictivo

La evaluación del desempeño predictivo del modelo se realizó mediante un esquema de validación temporal, en el cual el 80 % inicial de la serie se destinó al entrenamiento y el 20 % más reciente se reservó como conjunto de prueba. El desempeño se cuantificó utilizando el error absoluto medio (MAE), la raíz del error cuadrático medio (RMSE) y el coeficiente de determinación (R^2). Los resultados obtenidos se muestran en la tabla siguiente:

Tabla 5.5: Desempeño predictivo del modelo Random Forest

Modelo	RMSE	MAE	R^2
Random Forest	903.44	666.20	0.09

El modelo de Random Forest presenta un desempeño predictivo moderado, con valores de error RMSE = 903.44 y MAE = 666.20, superiores a los obtenidos por el Árbol de Decisión, pero con un coeficiente de determinación positivo aunque bajo ($R^2 = 0{,}09$). Este resultado indica que, si bien el modelo logra capturar de forma parcial la variabilidad de la serie, su capacidad explicativa global es limitada en el conjunto de prueba.

Este comportamiento puede estar asociado a la estructura temporal de los datos, dado que el algoritmo de Random Forest no incorpora explícitamente dependencias temporales y trata las observaciones como independientes. En este contexto, el modelo no logra capitalizar plenamente las ventajas del enfoque de ensamble, lo que reduce su efectividad frente a otros métodos más adecuados para el análisis de series de tiempo.

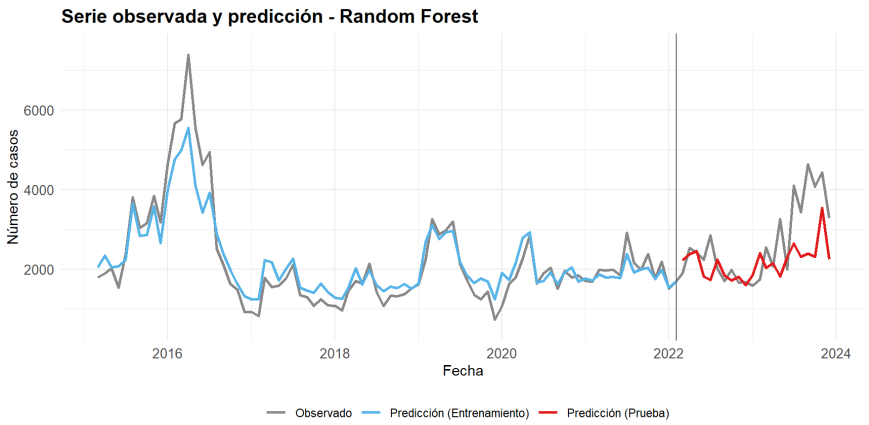

Figura 5.18: Serie observada y predicha del número de casos de malaria mediante el modelo Random Forest

5.6.7. Modelo Gradient Boosting

El modelo Gradient Boosting se incorpora al análisis por su alta capacidad predictiva y su habilidad para modelar interacciones no lineales complejas entre las variables climáticas y sociodemográficas,

aspectos fundamentales en la dinámica de transmisión de la malaria.

5.6.7.1. Fundamentación estadística

Gradient Boosting se basa en un modelo aditivo secuencial, donde los modelos débiles se ajustan de forma iterativa corrigiendo los errores del modelo previo. Formalmente, el modelo puede expresarse como:

$$F_m(x) = F_{m-1}(x) + \nu h_m(x) \tag{5.10}$$

donde $F_m(x)$ corresponde al modelo ensamblado en la iteración m, $F_{m-1}(x)$ representa el modelo acumulado hasta la iteración anterior, $h_m(x)$ es el modelo débil ajustado en dicha iteración (típicamente un árbol de decisión) y ν es la tasa de aprendizaje (*learning rate*), la cual controla la magnitud de la actualización y actúa como mecanismo de regularización.

5.6.7.2. Preprocesamiento de los datos

Al igual que otros métodos basados en árboles, Gradient Boosting no requiere el escalamiento de las variables explicativas, dado que las particiones se realizan a partir de reglas de umbral. Se mantuvo el esquema temporal original de los datos, respetando la secuencia cronológica de las observaciones.

5.6.7.3. Configuración del modelo

Se controla complejidad mediante learning rate, profundidad y número de árboles.

5.6.7.4. Arquitectura y ajuste de hiperparámetros

El modelo de Gradient Boosting se incorpora como un método aditivo secuencial orientado a mejorar progresivamente el ajuste del modelo mediante la combinación de múltiples modelos débiles. Este enfoque permite capturar relaciones complejas entre las variables explicativas y la variable respuesta, corrigiendo iterativamente los errores cometidos en etapas previas del entrenamiento, lo que resulta especialmente útil en problemas de predicción con estructuras no lineales.

El proceso de ajuste del modelo se rige por la minimización de una función de pérdida de tipo cuadrático, y su complejidad se controla mediante hiperparámetros como la tasa de aprendizaje (shrinkage), la profundidad de los árboles base y el número total de iteraciones. Estos parámetros se seleccionan de manera empírica e informada, considerando el balance entre precisión predictiva y estabilidad del modelo. En concordancia con buenas prácticas metodológicas, se evita una exploración

exhaustiva del espacio de hiperparámetros para reducir el riesgo de sobreajuste en un conjunto de datos de tamaño moderado. [62].

Los hiperparámetros escogidos para el modelo Gradient Boosting fuerón los siguientes:

```
gbm_model <- gbm(
  numero_de_registros ~ .,
  data = df_train %>% dplyr::select(-fecha),
  distribution = "gaussian",
  n.trees = 1000,
  interaction.depth = 3,
  shrinkage = 0.01,
  n.minobsinnode = 10,
  verbose = FALSE
)
```

- **n.trees = 1000**: Número de árboles en el ensamble, determinando la cantidad de iteraciones secuenciales del boosting.
- **interaction.depth = 3**: Profundidad máxima de cada árbol débil, controlando la complejidad y la capacidad de capturar interacciones entre variables.
- **shrinkage = 0.01**: Tasa de aprendizaje, que regula la contribución de cada árbol al modelo final y evita sobreajuste.
- **n.minobsinnode = 10**: Número mínimo de observaciones requeridas en un nodo terminal para permitir una división.

La configuración seleccionada permite que el modelo capture patrones complejos en los datos, aunque requiere un ajuste cuidadoso debido a la alta sensibilidad del boosting a los hiperparámetros. La comparación entre los valores predichos y los observados facilita evaluar su desempeño y entender la influencia de las variables en el ensamble.

5.6.7.5. Estrategia de entrenamiento

El entrenamiento se realizó de manera secuencial corrigiendo residuos del modelo previo. Los parámetros learning rate, profundidad y número de árboles fueron ajustados mediante un procedimiento de búsqueda sistemática.

5.6.7.6. Evaluación del desempeño predictivo

El desempeño predictivo del modelo de Gradient Boosting se evaluó utilizando un esquema de validación temporal, en el cual el 80 % inicial de la serie se destinó al entrenamiento y el 20 % más

reciente se reservó como conjunto de prueba. La evaluación se realizó mediante la raíz del error cuadrático medio (RMSE), el error absoluto medio (MAE) y el coeficiente de determinación R^2, calculados sobre la escala original de la variable de interés.

Tabla 5.6: Desempeño predictivo del modelo Gradient Boosting

Modelo	**RMSE**	**MAE**	R^2
Gradient Boosting	1144.22	812.66	-0.46

Como se muestra en la Tabla 5.6, el modelo de Gradient Boosting presenta los valores de error más altos entre los modelos evaluados, con un RMSE de 1144.22 y un MAE de 812.66, además de un coeficiente de determinación negativo ($R^2 = -0{,}46$), lo que evidencia un desempeño predictivo deficiente en el conjunto de prueba. Estos resultados indican que el modelo no logra generalizar adecuadamente y que su capacidad explicativa es inferior incluso a la de un modelo promedio. Esta situación puede estar asociada a la alta sensibilidad del método boosting al ajuste de hiperparámetros, así como a la presencia de ruido y a la variabilidad inherente de la serie epidemiológica analizada. En consecuencia, el modelo de Gradient Boosting no resulta adecuado para la predicción de los registros de malaria en el contexto de este estudio.

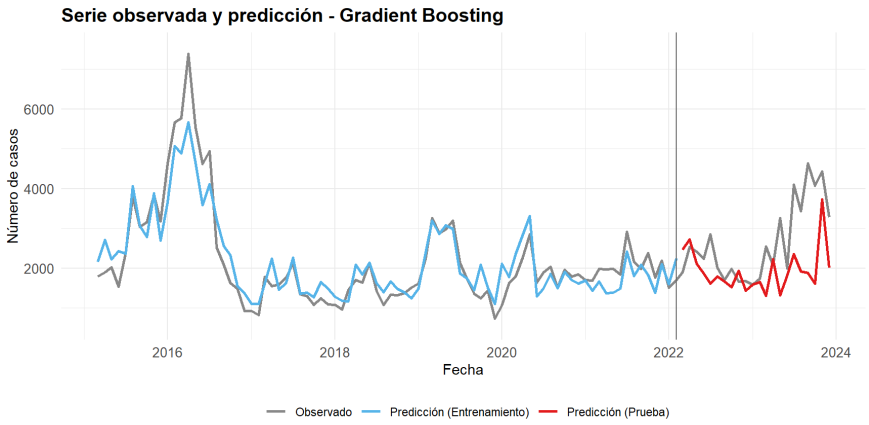

Figura 5.19: Serie observada y predicha del número de casos de malaria mediante el modelo Gradient Boosting

Capítulo 6

EVALUACION Y METRICAS DE LOS MODELOS PREDICTIVOS

Este capítulo tiene como objetivo evaluar y comparar el desempeño predictivo de los modelos estadísticos y de aprendizaje automático utilizados para analizar la dinámica de la malaria en Colombia durante el período 2015–2023. La evaluación rigurosa de los modelos constituye una etapa fundamental del proceso analítico, ya que permite cuantificar la precisión de las predicciones, identificar fortalezas y limitaciones metodológicas, y sustentar de manera objetiva la selección del enfoque más adecuado para el fenómeno estudiado.

La comparación se fundamenta en métricas de error ampliamente reconocidas en la literatura especializada y se complementa con un análisis cualitativo de los compromisos metodológicos asociados a cada modelo, así como con el estudio de la importancia relativa de las variables explicativas en los enfoques basados en árboles de decisión.

6.1. Criterios y métricas de evaluación del desempeño predictivo

La evaluación del desempeño predictivo de los modelos se fundamenta en el uso conjunto de la Raíz del Error Cuadrático Medio (RMSE), el Error Absoluto Medio (MAE) y el coeficiente de determinación R^2. Las métricas RMSE y MAE permiten cuantificar la magnitud promedio del error de predicción en las mismas unidades de la variable respuesta, diferenciándose en la penalización que otorgan a los errores de gran magnitud, siendo el RMSE más sensible a valores extremos.

Por su parte, el coeficiente de determinación R^2 evalúa la capacidad explicativa global del modelo, al medir la proporción de la variabilidad total de los registros de malaria que es explicada por las predicciones. En conjunto, estas tres métricas proporcionan una evaluación integral del desempeño, combinando criterios de precisión, robustez frente a valores atípicos y capacidad explicativa, aspectos fundamentales en el análisis de series epidemiológicas.

Las métricas se definen como:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{6.1}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{6.2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{6.3}$$

6.2. Comparación del desempeño predictivo de los modelos

La Tabla 6.1 presenta un resumen comparativo del desempeño predictivo de los modelos evaluados, junto con una síntesis de sus principales limitaciones metodológicas.

Tabla 6.1: Comparación del desempeño de modelos predictivos

Modelo	**RMSE**	**MAE**	R^2	**Limitaciones**
Red Neuronal Recurrente LSTM	1195.31	941.70	-0.59	Captura dependencias temporales, pero es sensible al tamaño muestral, con limitada generalización e interpretabilidad.
Árbol de Decisión	741.87	549.29	0.39	Modelo exploratorio, interpretable pero limitado en capacidad predictiva.
Random Forest	903.44	666.20	0.09	Modelo robusto, menos interpretable que el árbol simple.
Gradient Boosting	1144.22	812.66	-0.46	Modelo potente, sensible a hiperparámetros y menos interpretable.

6.3. Discusión de resultados

Los resultados obtenidos evidencian que el desempeño de los modelos evaluados es heterogéneo y está fuertemente condicionado por la naturaleza de la serie epidemiológica analizada. La red neuronal recurrente LSTM, a pesar de su idoneidad teórica para series de tiempo, presenta dificultades para generalizar adecuadamente en el contexto de una serie de longitud moderada, mostrando un comportamiento sensible a la variabilidad temporal y a la especificación del modelo.

El Árbol de decisión, si bien ofrece una estructura altamente interpretable y permite identificar relaciones relevantes entre las variables explicativas y la incidencia de malaria, presenta una capacidad predictiva limitada para capturar patrones complejos.

Por su parte, los modelos Random Forest y Gradient Boosting, exhiben un comportamiento más estable, aunque su desempeño no es uniforme y se ve afectado por la ausencia de una modelación explícita de la dependencia temporal. En conjunto, estos resultados sugieren que ningún enfoque resulta claramente dominante en todos los escenarios, lo que pone de manifiesto la complejidad del fenómeno estudiado y refuerza la necesidad de evaluar los modelos desde una perspectiva integral que considere simultáneamente precisión, estabilidad e interpretabilidad.

La complejidad de la enfermedad ayuda a explicar el rendimiento de estos modelos en el departamento del Chocó: la ubicación en una zona tropical húmeda que está directamente afectada por los fenómenos de la Niña y el Niño que tienen un carácter cíclico, los factores económicos como la presencia de minería ilegal y zonas de deforestación que favorecen la aparición de la malaria en la región, los factores del acceso a la salud de la población donde hay incoordinación entre la secretaria departamental (encargada de la promoción y la prevención) y las IPS (encargadas del diagnóstico y tratamiento). Estos factores climáticos, sociales y de acceso a la salud hacen que en este departamento se produzcan condiciones diferenciales en el desarrollo de la enfermedad como la convivencia de los 2 eventos más prevalentes de la malaria: Malaria vivax y Malaria falciparum con una reincidencia de la enfermedad en un 25 %, esto hace que el comportamiento epidemiológico de la enfermedad sea diferente al de otros departamentos en Colombia.

6.4. Importancia de las variables explicativas

6.4.1. Comparación visual de la importancia de variables

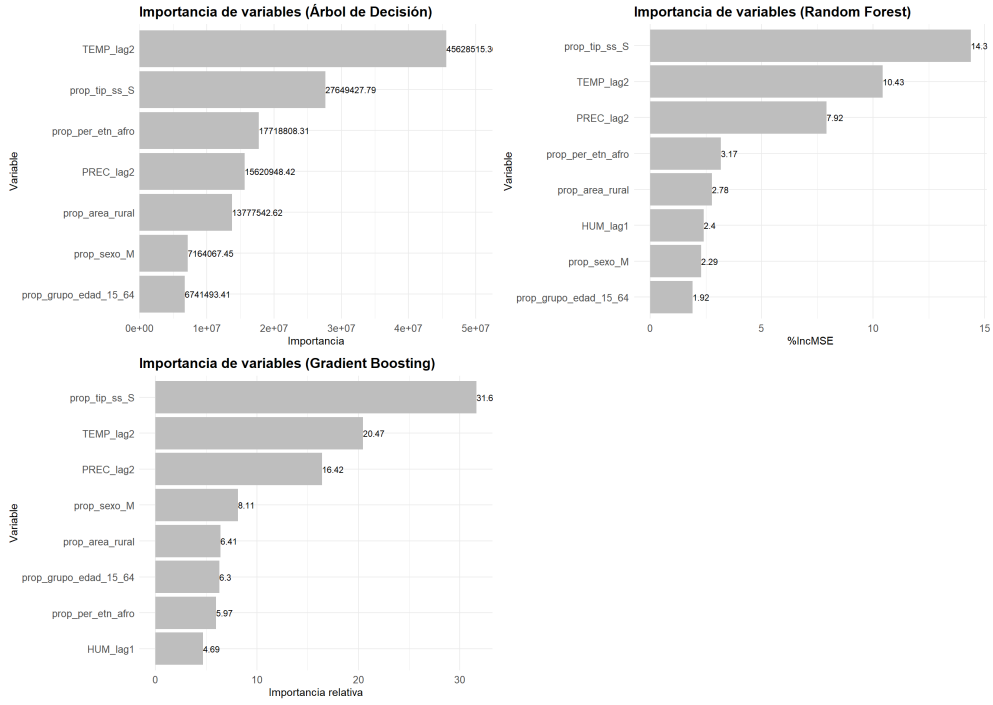

Figura 6.1: Importancia relativa de las variables explicativas en los modelos de aprendizaje automático

6.4.2. Análisis de la importancia de variables

Las gráficas de importancia obtenidas a partir del **Árbol de Decisión**, **Random Forest** y **Gradient Boosting** muestran un patrón consistente en la relevancia de las variables explicativas. En los tres modelos, la temperatura media mensual rezagada (`TEMP_lag2`) aparece como el predictor más importante, seguida por variables climáticas asociadas a la disponibilidad y persistencia de criaderos, como la precipitación rezagada (`PREC_lag2`) y la humedad relativa rezagada (`HUM_lag1`).

Entre los factores sociodemográficos, la proporción de población adscrita al régimen subsidiado (`prop_tip_ss_S`) y la proporción de población afrodescendiente (`prop_per_etn_afro`) presentan una contribución relevante y estable, lo que sugiere la influencia de condiciones estructurales en la dinámica de la malaria. El árbol individual concentra la importancia en pocas variables dominantes, mientras que los modelos de ensamble distribuyen la relevancia de manera más equilibrada, reflejando estimaciones más robustas.

No se presenta una gráfica de importancia para la **red neuronal artificial** debido a que este modelo no proporciona una medida interna y directamente interpretable de importancia de variables. En este estudio, la red neuronal se emplea principalmente como un enfoque comparativo en términos de desempeño predictivo global, manteniendo la interpretación de variables centrada en los modelos basados en árboles, que ofrecen mayor transparencia.

Capítulo 7

CONCLUSIONES Y TRABAJOS FUTUROS

7.1. Conclusiones

Las variables asociadas al comportamiento de la malaria en el Chocó durante el periodo 2015-2023 estuvieron vinculadas a factores demográficos, ambientales y sociales, los cuales reflejan la vulnerabilidad de la población y el efecto del clima sobre la dinámica temporal de la enfermedad.

Aunque el desempeño de los modelos evaluados (Red Neuronal Recurrente, Árbol de Decisión, Random Forest y Gradient Boosting) no permite generalizar resultados, estos evidencian un patrón consistente en la relevancia de las variables explicativas, lo que aporta información valiosa sobre los determinantes de la malaria en la región.

Los modelos de Machine Learning desarrollados sugieren que ninguno resulta claramente dominante en todos los escenarios, lo que refleja la complejidad del fenómeno estudiado y subraya la necesidad de evaluar los modelos desde una perspectiva integral, considerando simultáneamente precisión, estabilidad e interpretabilidad.

El trabajo aborda un problema de alta relevancia en salud pública y presenta un nivel de complejidad adecuado al integrar información epidemiológica con técnicas de Machine Learning. El tema implica desafíos importantes asociados al manejo de series temporales, la heterogeneidad de los datos y la interpretación de resultados en un contexto real.

7.2. Trabajos futuros

Se sugieren futuros estudios de Machine Learning que aborden el estudio de la malaria en el departamento del Choco utilizando tecnicas como el aprendizaje conjunto al combinar varios modelos para mejorar el rendimiento predictivo, el aprendizaje por transferencia aprovechando el conocimiento adquirido por un modelo preentrenado en una tarea o conjunto de datos iniciales y aplicarlo a una tarea o conjunto de datos nuevos, pero relacionados; el ajuste óptimo del modelo aumentando los datos para ampliar el conjunto de entrenamiento y la disminucion de sesgos al diversificar las fuentes de datos e incluir información representativa de otras fuentes, contextos y datos demográficos del departamento del Choco en el periodo de estudio. Se requiere que estas nuevas

propuestas tengan en cuenta otras fuentes de información globales como son las de OPS en malaria, las estaciones climáticas de otros países cercanos al área, cruces entre las fuentes de información en salud como la secretaria departamental y las IPS de la región, reportes sobre deforestación y minería ilegal en la región. El presente trabajo se realizó en el periodo de la pandemia de COVID por lo cual sería interesante contrastar estos datos para evidenciar posibles subregistros de la información. Futuros estudios podrían explorar con mayor profundidad elementos de innovación metodológica o comparaciones más avanzadas que amplíen el alcance científico del trabajo.

Bibliografía

[1] S. Blair. "Retos para la eliminación de la malaria en Colombia: un problema de saber o de poder". En: *Biomédica* 32 (2012), págs. 31-148.

[2] V. A. Olano et al. "Mapas preliminares de la distribución de especies de Anopheles vectores de malaria en Colombia". En: *Biomédica* 21.4 (2001), págs. 402-408.

[3] J. O. Rodríguez-Velásquez et al. "Dinámica de la epidemia de malaria en Colombia: Predicción probabilística temporal". En: *Revista de Salud Pública* 19 (2017), págs. 52-59.

[4] OPS/OMS. *Diez enfermedades transmitidas por vectores que ponen en riesgo a la población de las Américas*. 2017.

[5] World Health Organization. *World Malaria Report 2023*. World Health Organization, 2023.

[6] J. S. Serna-Trejos y S. G. Bermúdez-Moyano. "Contexto epidemiológico de malaria en Colombia, 2022". En: *Revista Cubana de Medicina* 61.4 (2022).

[7] Instituto Nacional de Salud. *Boletín epidemiológico semanal. Semana epidemiológica 45. Malaria*. ISSN 2357-6189, Colombia. 2024.

[8] J. P. R. Leiton et al. "Influencia de la fuerza de infección y la transmisión vertical en la malaria: Modelado Matemático". En: *Revista Facultad de Ciencias Básicas* 13.1 (2017), págs. 4-18.

[9] M. Casals, K. Guzmán y J. A. Caylà. "Modelos matemáticos utilizados en el estudio de las enfermedades transmisibles". En: *Revista Española de Salud Pública* 83 (2009), págs. 689-695.

[10] S. E. M. Varela, H. D. Morales y J. S. N. Cano. *Dinámica de los grupos de investigación en malaria en Colombia*. Inf. téc. Universidad Nacional de Colombia, 2015.

[11] E. A. Gómez Hernández et al. "Modelo multiparche y multigrupo para la transmisión de la malaria". En: *Ciencia en Desarrollo* 11.1 (2020), págs. 49-61.

[12] I. Jdey, G. Hcini y H. Ltifi. "Deep learning and machine learning for malaria detection: overview, challenges and future directions". En: *International Journal of Information Technology & Decision Making* 23.5 (2024), págs. 1745-1776.

[13] Y. W. Lee, J. W. Choi y E. H. Shin. "Machine learning model for predicting malaria using clinical information". En: *Computers in Biology and Medicine* 129 (2021), pág. 104151.

[14] Z. H. Ortiz. "Epidemiología de la malaria". En: *RIECS* 6.S1 (2021), págs. 14-17.

[15] M. Torres-Castro et al. "Las enfermedades transmitidas por vector: importancia y aspectos epidemiológicos". En: *Bioagrociencias* 13.1 (2020).

[16] R. López-Vélez y R. Molina-Moreno. "Cambio climático en España y riesgo de enfermedades infecciosas y parasitarias transmitidas por artrópodos y roedores". En: *Revista Española de Salud Pública* 79 (2005), págs. 177-190.

[17] J. C. P. Rodríguez et al. "Malaria epidemics in Colombia, 1970–2019". En: *Revista da Sociedade Brasileira de Medicina Tropical* 55 (2022), e0559-2021.

[18] S. M. Mandal, R. R. Sarkar y S. S. Somdatta Sinha. "Mathematical models of malaria: a review". En: *Malaria Journal* (2011).

[19] W. O. Kermack y A. G. McKendrick. "A contribution to the mathematical theory of epidemics". En: *Proceedings of the Royal Society of London. Series A* 115 (1922), págs. 100-121.

[20] N. Mideo, T. Day y A. F. Read. "Modelling malaria pathogenesis". En: *Cellular Microbiology* 10 (2008), págs. 1947-1955. DOI: 10.1111/j.1462-5822.2008.01208.x.

[21] L. F. Molineros Gallón y otros. "Aplicaciones de un modelo integral para el estudio de la malaria urbana en San Andrés de Tumaco, Colombia". En: *Revista Cubana de Medicina Tropical* 66.1 (2014), págs. 3-19.

[22] M. Susser. *Conceptos y estrategias en epidemiología. El pensamiento causal en las ciencias de salud*. México: Cultura Económica, 1991.

[23] A. S. Alzate. "Modelo para el Control de las Enfermedades Transmisibles". En: *Rev Colombiana Med.* 12.3 (1965), págs. 134-138.

[24] R. Briceño-León. "Las ciencias sociales y la salud: un diverso y mutante campo teórico". En: *Cienc Saúde Colect.* 8.1 (2003), págs. 33-45.

[25] H. Vargas. "Prevención y control de la malaria y otras enfermedades transmitidas por vectores en el Perú". En: *Rev Peruana Epidemiol.* 11.1 (2003), e4.

[26] M. G. Basáñez y D. Rodríguez. "Dinámica de transmisión y modelos matemáticos en enfermedades transmitidas por vectores". En: *Entomotropica* 19.3 (2004), págs. 113-124.

[27] D. Ruiz y otros. "Modelación sistémica para el diagnóstico de la interacción clima-malaria en Colombia. Aplicación durante El Niño 1997–1998 y La Niña 1998–2000". En: *Meteorología Colombiana* 5 (2002), págs. 41-48.

[28] J. Mosquera Rentería. *Modelamiento de casos de malaria en la región de Ashanti-Ghana usando regresión logística, Machine Learning y discriminante lineal de Fisher*. Tesis, Universidad Nacional de Colombia. 2024. URL: https://repositorio.unal.edu.co/handle/unal/85962.

[29] Monica Golumbeanu y otros. "Leveraging mathematical models of disease dynamics and machine learning to improve the development of novel malaria interventions". En: *Infectious Diseases of Poverty* 11 (2022), pág. 94. DOI: 10.1186/s40249-022-00981-1.

[30] Rohit Muralidhar, Michelle L. Demory y Marc M. Kesselman. "Exploring the Impact of Batch Size on Deep Learning Artificial Intelligence Models for Malaria Detection". En: *Cureus* 16.2 (2024), e55504. DOI: 10.7759/cureus.55504. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC11167577/.

[31] Mahdieh Poostchi y otros. "Image analysis and machine learning for detecting malaria". En: *Translational Research* 194 (abr. de 2018), págs. 36-55. ISSN: 1931-5244. DOI: 10.1016/j.trsl.2017.12.004. URL: https://www.sciencedirect.com/science/article/pii/S193152441730333X.

[32] S. Rajab, J. Nakatumba-Nabende y G. Marvin. "Interpretable machine learning models for predicting malaria". En: *IEEE Conference Proceedings: Smart Technologies and Systems for Next Generation Computing (ICSTSN)*. 2023, págs. 1-6. DOI: 10.1109/ICSTSN57401.2023.10153152.

[33] Y. A. Adamu. "Malaria prediction model using machine learning algorithms". En: *Turkish Journal of Computer and Mathematics Education (TURCOMAT)* 12.10 (2021), págs. 7488-7496.

[34] Nkiruka Odu, Rajesh Prasad y Clement Onime. "Prediction of malaria incidence using climate variability and machine learning". En: *Informatics in Medicine Unlocked* 22 (2021), pág. 100489. URL: https://www.sciencedirect.com/science/article/pii/S2352914820306596.

[35] Y. Kim y otros. "Malaria predictions based on seasonal climate forecasts in South Africa: A time series distributed lag nonlinear model". En: *Scientific Reports* 9.1 (2019), pág. 17882.

[36] O. Nkiruka, R. Prasad y O. Clement. "Prediction of malaria incidence using climate variability and machine learning". En: *Informatics in Medicine Unlocked* 22 (2021), pág. 100508.

[37] E. Tuta-Quintero et al. "Application of artificial intelligence in the Prediction of Complications in patients with Malaria". En: *Infectio* 28.4 (2024), págs. 235-240.

[38] O. A. Abisoye. "A computer based approach for the prediction of multiclass symptomatic malaria infection". En: (2019).

[39] H. Chiroma et al. "Malaria severity classification through Jordan-Elman neural network based on features extracted from thick blood smear". En: *Neural Network World* 25.5 (2015), pág. 565.

[40] R. Parveen et al. "Prediction of malaria using artificial neural network". En: *Int J Comput Sci Netw Secur* 17.12 (2017), págs. 79-86.

[41] S. Ruban, A. Naresh y S. Rai. "A noninvasive model to detect malaria based on symptoms using machine learning". En: *Advances in Parallel Computing Technologies and Applications*. IOS Press, 2021, págs. 23-30.

[42] S. S. Yadav et al. "Machine learning based malaria prediction using clinical findings". En: *2021 International Conference on Emerging Smart Computing and Informatics (ESCI)*. IEEE, mar. de 2021, págs. 216-222.

[43] P. E. Fleitas et al. "Machine learning approach to identify malaria risk in travelers using real-world evidence". En: *Heliyon* 10.7 (2024).

[44] H. I. Okagbue et al. "Diagnosing malaria from some symptoms: a machine learning approach and public health implications". En: *Health and Technology* 11.1 (2021), págs. 23-37.

[45] J. D. Gutiérrez. "Stochastic treatment regimes in climate-health research: Reassessing malaria risk under warming scenarios in Colombia". En: *PLOS Global Public Health* 5.9 (2025), e0005252.

[46] J. Mosquera Renteria. "Modelamiento de casos de malaria en la región de Ashanti-Ghana usando regresión logística, Machine Learning y discriminante lineal de Fisher". Tesis doct. Universidad Nacional de Colombia, 2024.

[47] J. Carmona-Fonseca. "La malaria en Colombia, Antioquia y las zonas de Urabá y Bajo Cauca: panorama para interpretar la falla terapéutica antimalárica. Parte 2". En: *Iatreia* 17.1 (2004), págs. 34-53.

[48] V. A. J. Mora, L. F. J. Cañarte y A. N. Zavala-Hoppe. "Malaria: demografía, prevalencia y factores de riesgo en niños a nivel mundial". En: *Revista Científica de Salud BIOSANA* 4.4 (2024), págs. 306-316.

[49] A. Knudson-Ospina y otros. "Perfil clínico y parasitológico de la malaria por Plasmodium falciparum y Plasmodium vivax no complicada en Córdoba, Colombia". En: *Revista de la Facultad de Medicina* 63.4 (2015), págs. 595-607.

[50] L. Hilarión-Gaitán et al. "Desigualdades en salud según régimen de afiliación y eventos notificados al Sistema de Vigilancia (Sivigila) en Colombia, 2015". En: *Biomédica* 39.4 (2019), págs. 737-747.

[51] C. Laborde-Cárdenas et al. "Caracterización epidemiológica de pacientes con malaria, notificados por un asegurador en salud en Colombia, 2016–2017". En: *Revista Cubana de Medicina Tropical* 72.1 (2020), págs. 1-15.

[52] J. A. Cardona-Arias, W. A. Salas-Zapata y J. Carmona-Fonseca. "Determinación y determinantes sociales de la malaria: revisión sistemática, 1980–2018". En: *Revista Panamericana de Salud Pública* 43 (2019), e39.

[53] United Nations. *World Population Prospects 2019*. United Nations, 2019.

[54] World Health Organization. *World Malaria Report 2023*. World Health Organization, 2023.

[55] Robert W. Snow y John A. Omumbo. "Malaria". En: *Disease and Mortality in Sub-Saharan Africa*. 2.ª ed. World Bank, 2006.

[56] Departamento Administrativo Nacional de Estadística. *Proyecciones de población nacional y departamental 2018–2050*. DANE, 2022.

[57] E. Arango et al. "Occupational risk factors and malaria transmission in endemic regions of Colombia". En: *American Journal of Tropical Medicine and Hygiene* 87.6 (2012), págs. 1103-1110.

[58] Gavin Yamey y Marco Schäferhoff. "Global malaria control: epidemiological and demographic perspectives". En: *The Lancet Global Health* 2.9 (2014), e450-e451.

[59] Gareth James et al. *An Introduction to Statistical Learning: with Applications in R*. 2.ª ed. Springer, 2021.

[60] Erin A. Mordecai et al. "Optimal temperature for malaria transmission is dramatically lower than previously predicted". En: *Ecology Letters* 16.1 (2013), págs. 22-30.

[61] Rachel Lowe et al. "Environmental predictors of malaria risk: a review and case study in Colombia". En: *Scientific Reports* 7.1 (2017), págs. 1-11.

[62] Gareth James et al. *An Introduction to Statistical Learning: With Applications in R*. 2.ª ed. New York: Springer, 2021.