PROYECTO DE GRADO

MODELO DE INTELIGENCIA ARTIFICIAL PARA LA CLASIFICACIÓN DE TIPOS DE DENGUE BASADO EN DATOS DEL SIVIGILA EN EL DEPARTAMENTO DEL CESAR

DIRECTOR: AMILKAR SIERRA ROMANO

INTEGRANTES:

GERMAN ENRIQUE ARDILA MENDEZ

CC: 1.003.234.282 geardila@unicesar.edu.co

DANIEL EDUARDO ESPAÑA CARPIO

CC: 1.193.397.483 deespana@unicesar.edu.co

2026

TABLA DE CONTENIDO

ESQUEMA DEL PROYECTO DE GRADO

PRELIMINARES

SECCIÓN I: Descripción General

1.1 Titulo del Proyecto de Grado
1.2 Dirección de Ejecución
1.3 Lapso de Ejecución
1.4 Organismo o Institución Responsable del Proyecto
1.5 Información de contacto de los estudiantes
1.6 Línea, sublínea y grupo de investigación del Proyecto

SECCIÓN II: Descripción Situacional

2.1 Identificación del Problema
2.2 Justificación del Proyecto
2.3 Objetivos del Proyecto
2.4 Bases Teóricas
- 2.4.1 Antecedentes
  - 2.4.1.1 Antecedentes históricos.
  - 2.4.1.2 Antecedentes investigativos.
  - 2.4.1.3 Antecedentes legales.
- 2.4.2 Marco Teórico
- 2.4.3 Marco Conceptual

2.5 Marco Metodológico

SECCIÓN III: Desarrollo Científico-Tecnológico

3.1 Desarrollo de las fases de la metodología de sistemas propuesta
3.2 Análisis de Resultados y Discusión
3.3 Conclusiones
3.4 Recomendaciones
3.5 Bibliografía

SECCIÓN I: DESCRIPCIÓN GENERAL

1.1.- TÍTULO DEL PROYECTO

Modelo de Inteligencia Artificial para la Clasificación de Tipos de Dengue basado en datos del SIVIGILA en el Departamento del Cesar

1.2.- DIRECCIÓN DE EJECUCIÓN DEL PROYECTO

Valledupar - Cesar

1.3.- LAPSO DE EJECUCIÓN DEL PROYECTO

La totalidad del proyecto en cuestión, abarcó el tiempo de tres (3) meses

Del 10 de enero de 2026 al 10 de abril de 2026

1.4.- ORGANISMO Y SECCIÓN RESPONSABLE

Prototipo de investigación fundamentado en el estudio del virus transmitido por el mosquito Aedes aegypti que se llevará a cabo en la ciudad de Valledupar, departamento del Cesar.

1.5.- INFORMACION DE CONTACTO DE LOS ESTUDIANTES

Nombre	Apellido	Cédula	Teléfono	Correo
Daniel Eduardo	España Carpio	1193397483	3025683271	deespana@unicesar.edu.co
German Enrique	Ardila Mendez	1003234282	3187442084	geardila@unicesar.edu.co

1.6.- LÍNEA, SUBLINEA Y GRUPO DE INVESTIGACIÓN AL QUE SE SUSCRIBE EL PROYECTO

Línea de Investigación	Transformación Digital
Sub-línea de Investigación	Ciencia e Ingeniería de los Datos
Área Temática	Machine Learning
Grupo de Investigación	

SECCIÓN II. DESCRIPCIÓN SITUACIONAL

2.1.- IDENTIFICACIÓN DEL PROBLEMA

El dengue, causado por el virus del dengue y transmitido principalmente por el mosquito Aedes aegypti, es una amenaza constante en regiones tropicales y subtropicales. En ciudades como Valledupar, Colombia, la proliferación de este vector es favorecida por condiciones ambientales específicas, lo que ha llevado a una elevada tasa de casos reportados.

La vigilancia y control del dengue dependen en gran medida del Sistema Nacional de Vigilancia en Salud Pública (SIVIGILA), el cual recopila datos clínicos y epidemiológicos para la identificación y clasificación de los casos. Sin embargo, los métodos tradicionales de análisis y clasificación manual de estos datos presentan limitaciones significativas, especialmente en la capacidad de predecir brotes y clasificar los tipos de dengue de manera oportuna y precisa.

FORMULACION DEL PROBLEMA:

El dengue, como una de las enfermedades transmitidas por vectores más significativas a nivel mundial, continúa representando una amenaza creciente para la salud pública. Según la Organización Mundial de la Salud (OMS), se reportan más de 390 millones de infecciones anuales, de las cuales aproximadamente 96 millones manifiestan síntomas clínicamente reconocibles (Roster y Rodriguez, 2021). En América Latina, este problema ha alcanzado niveles epidémicos, y en Colombia, solo en los últimos cinco años, se han reportado más de 500,000 casos, lo cual ha sobrecargado los sistemas de salud pública.

Desde 2019 hasta la actualidad, Valledupar ha experimentado variaciones en los reportes de dengue, con un promedio anual cercano a los 5,000 casos, registrando un aumento especialmente marcado en 2022, cuando los reportes superaron los 6,500 casos, cifra que representa un aumento del 30% con respecto al año anterior (SIVIGILA, 2024). Durante 2023, a pesar de las campañas de prevención, los casos mantuvieron un nivel elevado, con más de

5,800 pacientes atendidos, concentrándose principalmente en la población infantil y juvenil. Hasta el primer trimestre de 2025, los datos preliminares indican una tendencia al alza, lo que alerta a las autoridades sanitarias sobre la posibilidad de un nuevo pico epidémico si no se intensifican las medidas preventivas.

En Valledupar, el incremento del 30 % en los casos durante los últimos dos años ha afectado principalmente a niños y jóvenes menores de 20 años con un porcentaje del 78,35 % y adultos con un porcentaje del 13,93% de los casos, quienes presentan mayor vulnerabilidad a desarrollar formas graves de la enfermedad (SIVIGILA, 2024). Este incremento ha derivado en una mayor saturación de los servicios de salud, con hospitales reportando ocupaciones superiores al 90 % en sus áreas de atención a enfermedades infecciosas durante las temporadas de mayor incidencia.

El contexto social es profundamente afectado. Las familias enfrentan sufrimientos en el aspecto físico, emocional, además del impacto económico asociado al tratamiento médico. Las limitaciones económicas y de infraestructura en comunidades de bajos recursos han favorecido la proliferación del mosquito Aedes aegypti, responsable de la transmisión del dengue, con los estratos 1 y 2 concentrando el 68,81 % y 26,92 % de los casos, respectivamente.

En Valledupar, las campañas de fumigación y sensibilización han sido una estrategia constante liderada por la Secretaría Local de Salud para combatir el dengue. Estas intervenciones incluyen acciones diarias en diferentes barrios y comunas de la ciudad, como Arizona, Los Fundadores, El Rocío, Bello Horizonte, y más (SIVIGILA, 2024). La campaña “Juntos contra el dengue” ha sido una de las principales estrategias, abarcando visitas casa a casa para educar a la población sobre medidas preventivas, como el lavado de albercas y la eliminación de recipientes con agua acumulada. Durante estas jornadas, se han identificado criaderos en hasta el 80% de las viviendas visitadas en sectores como La Victoria, lo que refuerza la importancia del compromiso comunitario para complementar los esfuerzos institucionales (Alcaldía de Valledupar, 2024).

Este escenario evidencia la necesidad de un enfoque integral y sostenido que no solo atienda los casos clínicos, sino también se centre en la prevención y educación comunitaria, el fortalecimiento de la infraestructura de salud y la integración de herramientas tecnológicas, modelos predictivos basados en inteligencia artificial, para anticipar y mitigar brotes futuros.

Para este proyecto, se plantea entonces la siguiente pregunta de investigación:

Pregunta Principal:

¿Cómo anticiparse a brotes futuros del virus del dengue en el departamento del Cesar mediante el uso de modelos predictivos basados en inteligencia artificial (IA) que automáticamente clasifiquen los diferentes tipos de dengue?

Sub-Preguntas

¿Qué criterios se deben tener en cuenta para clasificar los diferentes tipos de dengue?

¿Qué modelos de inteligencia artificial son adecuados para evaluar los diferentes resultados y criterios de clasificación de tipos de dengue?

¿Cómo puede integrarse el resultado de estos modelos a un sistema de predicción basado en la sintomatología de un paciente infectado por el virus del dengue?

¿Cómo calificar la eficacia, fiabilidad y precisión brindadas por las diferentes métricas de evaluación de modelos de inteligencia artificial?

La respuesta a estas preguntas permitirá construir y validar los diferentes modelos empleados para la solución de este proyecto.

2.2.- JUSTIFICACIÓN DEL PROYECTO

El dengue, enfermedad viral transmitida por el mosquito Aedes aegypti, representa un importante problema de salud pública a nivel mundial, con especial impacto en regiones tropicales y subtropicales. En ciudades como Valledupar, la incidencia del dengue ha mostrado un crecimiento sostenido, afectando la salud de la población y generando una sobrecarga significativa en el sistema de salud local.

A pesar de los esfuerzos implementados en vigilancia epidemiológica, la clasificación precisa y oportuna de los diferentes tipos de dengue continúa siendo un reto, debido a la complejidad de sus manifestaciones clínicas y a la heterogeneidad en la evolución de la enfermedad entre los pacientes. Los métodos tradicionales de análisis y clasificación, actualmente utilizados, resultan insuficientes para gestionar de manera eficiente el creciente volumen de datos provenientes del Sistema de Vigilancia en Salud Pública (SIVIGILA). Esta limitación repercute negativamente en el diagnóstico temprano y en la toma de decisiones clínicas, afectando la calidad de la atención médica y la efectividad de las estrategias de control de brotes.

Las causas de este problema pueden estar relacionadas con la falta de herramientas tecnológicas avanzadas en los sistemas de vigilancia epidemiológica y la creciente complejidad de los datos clínicos y epidemiológicos asociados al dengue. La diversidad en la presentación clínica de la enfermedad y su progresión varían significativamente entre los pacientes, lo que dificulta su clasificación basada únicamente en criterios manuales y estadísticos tradicionales.

En este contexto, la presente propuesta se justifica en la necesidad de desarrollar un mecanismo automatizado, basado en técnicas de inteligencia artificial (IA), que permita la clasificación precisa de los diferentes tipos de dengue a partir de datos epidemiológicos.

La implementación de un modelo de IA no solo mejoraría el análisis de los datos del SIVIGILA, sino que también facilitaría la detección de patrones y tendencias de forma rápida y precisa, optimizando la respuesta médica ante los brotes. Esto tendría un profundo impacto en las

poblaciones más vulnerables porque se podrían tratar más rápido y más eficientes este tipo de patología y brindarle atención prioritaria a los pacientes que presentan casos graves. Además, beneficiaría significativamente a los centros de salud y hospitales, ya que, con la ayuda de esta herramienta se pueden hacer predicciones a años futuros teniendo en cuenta datos históricos y pueden prepararse tanto en instalaciones como en personal médico para mitigar o contrarrestar los casos que se puedan presentar.

Este trabajo es relevante tanto para la comunidad científica como para la educativa porque propone el uso de inteligencia artificial (IA) para la clasificación de tipos de dengue, lo que representa un avance significativo en la aplicación de técnicas modernas para la mejora de la salud pública. El desarrollo de un modelo de IA permitirá un análisis más eficiente y preciso de los datos del SIVIGILA, ayudando a los profesionales de salud a detectar patrones y tendencias más rápidamente, mejorando la respuesta ante brotes de dengue. Además, este enfoque puede servir de referencia para futuros estudios en la aplicación de inteligencia artificial en el campo de la epidemiología y la salud pública.

La elección de este tema responde a la necesidad urgente de optimizar las herramientas de vigilancia y análisis de enfermedades en entornos de alto riesgo, como Valledupar, donde el dengue constituye una amenaza constante. El desarrollo de soluciones basadas en inteligencia artificial representa una oportunidad para mejorar el manejo de la enfermedad, así como para aportar al avance científico y educativo en el campo de la epidemiología moderna.

2.3.- OBJETIVOS DEL PROYECTO

- **Objetivo General:** Diseñar un modelo de inteligencia artificial empleando la metodología CRISP-DM que automáticamente clasifique los tipos de dengue, basado en datos epidemiológicos del sistema de vigilancia SIVIGILA.
- **Objetivos Específicos**

  - ✓ Analizar los datos del sistema de vigilancia epidemiológica de enfermedades de interés

en salud pública transmitidas por el mosquito del dengue (vector Aedes aegypti) con registros que datan desde el año 2018 hasta el año actual en el departamento del Cesar.

- ✓ Identificar patrones de incidencia del dengue en la base de datos del sistema de vigilancia epidemiológica SIVIGILA en conjunto con las tres (3) primeras etapas de la metodología CRISP-DM.
- ✓ Entrenar modelos de machine learning para la correcta clasificación de los tipos de dengue, utilizando los datos epidemiológicos proporcionados por el sistema SIVIGILA.
- ✓ Evaluar la efectividad de diversas técnicas de machine learning en función de los criterios de rendimiento de minería de datos (descubrimientos), comparándolas con métodos estadísticos tradicionales para la clasificación de casos de dengue.
- ✓ Diseñar un tablero de control (CMI) en la nube como sistema de apoyo a toma de decisiones en centros de salud.

2.4.- BASES TEÓRICAS.

2. 4.1 ANTECEDENTES

2.4.1.1. Antecedentes históricos.

El dengue desde hace siglos ha sido de las problemáticas en crecimiento del sector salud en Colombia, teniendo en cuenta esta y muchas otras enfermedades fue indispensable el desarrollo de un sistema que permitiera recolectar, analizar y dar previo aviso de presencia de enfermedades a nivel nacional, para lo cual se estructuro el Sistema Nacional de Vigilancia en Salud Pública (SIVIGILA). Con las nuevas tecnologías y el desarrollo de distintos modelos de machine learning, se han desarrollado algoritmos y métodos eficaces en lo relacionado con diagnósticos y predicciones que pueden ayudar a dar tratamientos más oportunos.

Se hizo un estudio al sistema de vigilancia nacional del dengue en un municipio en el Valle del Cauca del año 2008 por Zea & Osorio (2011) realizando entrevistas y revisando datos clínicos de dengue en el departamento. Se hallaron falencias en el conocimiento que tiene el personal respecto al uso del sistema.

En su investigación Alcalá, Quintero, González & Brochero (2015) evaluaron la productividad de pupas de Aedes aegypti (vector del dengue) en épocas de lluvia y sequía en un conglomerado de 100 viviendas y sus espacios públicos en Girardot, lo cual, dio como resultado la presencia de este vector tanto en épocas de lluvia, como de sequía un porcentaje de 94% y 98% al interior de las viviendas. A su vez, con la ayuda de un modelo matemático se representó la dinámica del virus del dengue basado en los casos epidemiológicos desde 1997 en Colombia por Camargo (2012), utilizando muestras de historias clínicas. Con esta simulación real del virus se pudo predecir un brote del periodo 2012 a 2014.

El mejoramiento de la medicina con procesos optimizados como el diagnostico diferencial fue una propuesta de Lugo, Maldonado & Murata (2014) para facilitar diagnósticos clínicos de diversos padecimientos utilizando aprendizaje automático de la inteligencia artificial además de otras técnicas como procesamiento de datos con redes neuronales, clasificadores bayesianos, regresión logística o máquinas de soporte vectorial.

Un estudio realizado por Cardona (2018) demuestra que factores como los sociales, económicos y espaciales favorecen el crecimiento, incidencia y permanencia del dengue en una población, para esto ayudó a la construcción de modelos explicativos para obtener pronósticos precisos y oportunos de la incidencia del dengue y con la nueva tecnología propuesta por Rodríguez, Prieto, Pérez, Pardo, Correa, Mendoza, Bravo, Morales, Rojas & Flórez (2018) basada en la teoría de probabilidad obtuvieron como resultado una predicción acertada para el número de infectados de dengue grave en el total de cada departamento.

Castrillón, Castaño, & Urcuqui (2015) analizaron incidencia del dengue en Colombia de 2004 hasta el 2013, mediante la recopilación de datos en bases de datos como la del INS, IDEAM y SIVIGILA. Con la utilización de herramientas de análisis y síntesis que proporciona la explotación de la información Lujan, Pytel & Pollo (2014), se puede administrar datos relevantes para el tratamiento y/o diagnóstico de pacientes y con ayuda del modelo hibrido que proponen Dávila & Sanchez (2012) contribuir a la mejora de los Sistemas Clínico de Soporte para la Toma de Decisiones (CDSS), mediante el uso de la minería de datos para

clasificar y estudiar grandes cantidades de datos, la metodología CRISP-DM y la combinación de dos modelos matemáticos.

2.4.1.2. Antecedentes investigativos.

Los avances recientes en inteligencia artificial (IA) y machine learning han demostrado ser prometedores para mejorar la clasificación de enfermedades como el dengue. Diversos estudios han explorado la aplicación de algoritmos como los árboles de decisión, las redes neuronales artificiales (ANN) y las máquinas de vectores de soporte (SVM) para el análisis de datos clínicos y epidemiológicos.

Un estudio clave de Santos et al. (2021) utiliza varios algoritmos de machine learning para evaluar el riesgo de dengue en diferentes regiones de Brasil, destacando la utilidad de los árboles de decisión y las ANN para la predicción de casos graves.

Otro estudio realizado por Ramírez y Martínez (2021) evaluó la efectividad de las ANN y random forest para predecir la carga de dengue en Colombia, basándose en datos del SIVIGILA. Los resultados mostraron que los modelos basados en machine learning superaron a los métodos tradicionales de predicción, como los modelos ARIMA.

A nivel internacional, López et al. (2020) también han investigado la aplicación de modelos supervisados para la clasificación de los diferentes tipos de dengue. Este estudio destaca cómo los datos clínicos y ambientales, combinados con técnicas de machine learning, pueden mejorar significativamente la precisión en la clasificación de casos graves y no graves.

No obstante, a pesar de estos avances, uno de los desafíos más recurrentes en la implementación de estos modelos es la calidad y disponibilidad de los datos, ya que los sistemas de vigilancia como SIVIGILA a menudo cuentan con datos incompletos o mal estructurados, lo que impacta negativamente la precisión de los modelos predictivos.

Otra de las investigaciones relevantes en este estudio es la desarrollada por Quiroz et al. (2022), en la cual se resalta la capacidad existente en los algoritmos de inteligencia artificial para la identificación de patrones, lo cual les permite generar una identificación temprana de ciertos

resultados bajo una similitud de patrones dada, factor que ha sido la base para la implementación de estos algoritmos de inteligencia artificial en la medicina, en el caso puntual de la investigación desarrollada por Quiroz et al. (2022) se refiere a la implementación de estos algoritmos en la identificación de pacientes que pudieron estar infectados por el virus COVID-19 que se encuentran en alto riesgo para con ello contribuir a la gestión hospitalaria en el marco de la emergencia dada por la pandemia COVID-19.

La técnica machine learning empleada en este modelo es la de redes neuronales, es preciso aclarar que de acuerdo con Quiroz et al. (2022) “esta rama de la inteligencia artificial se encarga de dotar a sistemas informáticos de la capacidad de “aprender” a partir de un conjunto de datos conocidos” (p. 93) en este caso particular, los datos conocidos son las bases de datos públicas presentadas por el gobierno mexicano, con las que se hicieron las pruebas pertinentes en el proceso de validación del modelo. En dicho modelo, se busca identificar si el paciente basado en sus síntomas, pertenece al grupo de personas que tiene mayores posibilidades de sobrevivir que de fallecer o si, por lo contrario, pertenece al grupo de personas que tiene mayores posibilidades de fallecer que de sobrevivir a la enfermedad, esto permite brindar una atención oportuna a los pacientes y poder planificar y gestionar de una forma óptima los recursos clínicos.

Ruiz y Velásquez (2023) hacen un aporte significativo en el análisis del uso de la inteligencia artificial en el campo de la salud, basándose en el avance generado con la obtención de datos sobre la salud de los pacientes, lo cual facilita el recopilar información sobre sintomatología y el comportamiento de esta en el avance de una enfermedad, a su vez, permite a los algoritmos identificar patrones y poder realizar una predicción sobre el posible avance de la enfermedad. Adicionalmente, esta relación creciente entre la medicina y la tecnología, permite llevar el desarrollo médico hacia la prevención de las enfermedades y disminuir el accionar paliativo que tiene actualmente a medicina, ya que los algoritmos de IA (inteligencia artificial) permite tener un pronóstico o predicción de lo que podría pasar con un análisis de los síntomas presentados por el paciente.

Buñay Mendez et al (2024) expresa en su estudio el desarrollo de un sistema de diagnóstico temprano del dengue mediante técnicas de machine learning. Para ello, se utilizaron datos históricos recolectados en el Centro de Salud de la ciudad de Tena (Ecuador). Esta investigación

busca responder a la necesidad de contar con métodos diagnósticos más rápidos, accesibles y menos invasivos para el dengue, especialmente en regiones endémicas como la nuestra. Se siguió una metodología basada en la Ciencia del Diseño y un enfoque particular en la reducción de dimensionalidad de los datos. Además, se implementaron métodos de ensamble como Bagging y Boosting para mejorar la robustez y precisión de los modelos

Así mismo, de acuerdo con Belloso et al (2023) la medicina se ha visto beneficiada con el incremento de disponibilidad de información contribuyendo a la generación de conocimiento biológico apoyado en herramientas tecnológicas vinculadas con el análisis de datos y con ello la inteligencia artificial también, aclarando que esta última hace referencia a una rama computacional orientada más allá del solo entender, sino que busca construir unidades inteligentes realizando un proceso de razonamiento y racionalidad es decir, como recreación de la conducta humana y el proceder lógico ideal.

las técnicas de machine learning (ML) también conocido como aprendizaje automático, de acuerdo con Belloso et al. (2023), no se queda en solo la generación y análisis de datos, sino que pretende que a través de un algoritmo se genere un modelo basado en la identificación de patrones en un conjunto de datos que pueden usarse para llevar a cabo predicciones, es decir, posibles escenarios futuros que le permiten en el campo médico tomar decisiones y actuar previniendo escenarios críticos para los pacientes. Ahora bien, ese proceso de identificación de patrones, es realizado mediante actividades de prueba por así decirlo en donde se van aplicando pruebas para avanzar se identifican los errores y se lleva a cabo su corrección hasta encontrar el resultado óptimo, el más certero de todos.

Belloso et al. (2023) también manifiesta que actividades desarrolladas en el campo de la salud como “la digitalización de historias clínicas, la anotación de datos a gran escala, la disponibilidad de datos en diversos formatos digitales, el refinamiento en los métodos de ML para analizar múltiples variables complejas y su disponibilidad como código abierto, entre otras.” (p. 79) han permitido que los algoritmos de machine learning tengan un mayor nivel de efectividad y los modelos generados con base a ello tengan una mayor confiabilidad.

De acuerdo con Belloso et al. (2023) entre las técnicas de aprendizaje automático más utilizados en el campo de la salud, se encuentra el clustering en el cual se busca agrupar aquellos datos o

casos con mayor relación o similitud; otra de las técnicas es la de redes neuronales artificiales, en la cual se lleva a cabo el análisis de datos por medio de una relación o conexión entre diversos nodos neuronales similar al proceso de análisis realizado por las neuronas en el cerebro humano, de allí su nombre.

Ahora bien, los autores Belloso et al. (2023) realizan un análisis de la aplicabilidad de la inteligencia artificial en la toma de decisiones e identificación de casos epidemiológicos en una etapa temprana, y destaca que para una adecuada predicción epidemiológica es indispensable contar con una base de datos clínicos y diagnósticos amplia y segura, es decir, que los datos allí depositados sean altamente confiables, esto permitirá que la generación del modelo de predicción epidemiológica sea más exacto en sus resultados.

Esta vinculación entre la inteligencia artificial y la salud, tuvo un mayor crecimiento en el periodo de tiempo relacionado con la pandemia COVID-19, debido a la búsqueda de soluciones en el monitoreo, diagnóstico y tratamiento de las personas infectadas por el virus. Márquez (2020) desarrolla un estudio en el cual se analiza el funcionamiento de las diversas técnicas de inteligencia artificial y el avance tecnológico en el abordaje de las enfermedades infecciosas y de rápido crecimiento como lo fue el COVID-19. En el cual manifiesta que el uso de la inteligencia artificial de la mano de otras herramientas como el análisis de datos avanzado, pueden contribuir significativamente a la detección temprana de los casos de contagio por COVID-19, así como su desarrollo y nivel de criticidad, toda vez que la inteligencia artificial ofrece la generación de softwares que permite la identificación, diagnóstico y tratamiento de casos de contagio con el análisis de datos, haciendo mención como otros investigadores en lo importante que es tener una base de datos de calidad, confiable para el desarrollo de estos software con resultados seguros.

Márquez (2020) también destaca el aporte significativo del uso del aprendizaje automático (Machine learning) en el abordaje de estos casos de contagio por COVID-19, en donde menciona que esta técnica es empleada en el diagnóstico oportuno de estos casos de contagio, así como la predicción sobre el riesgo de contagio por parte de una persona, basado en ciertas características como la edad, peso, hábitos, ubicación geográfica, entre otros. Esto permite no solo identificar el riesgo de contagio por parte de una persona, sino también, la probabilidad de presentar complicaciones en el proceso de contagio por el virus.

Márquez (2020) destaca que “El aprendizaje automático abre un sinnúmero de posibilidades de investigación en diversos campos clínicos… Esto involucra desde los escáneres faciales para identificación de síntomas como la fiebre, wearables para medición y detección de anomalías cardiacas o respiratorias, hasta chatbots que evalúan a un paciente cuando este menciona sus síntomas y, basado en las respuestas dadas, el sistema le indica si debe permanecer en casa, llamar al médico o ir al hospital.” (p. 321)

En este mismo sentido, López et al. (2020) desarrollaron una investigación relacionada con la identificación del COVID-19 usando técnicas de inteligencia artificial para el análisis de radiografías, en la cual se corrobora la importancia de una recolección de datos de calidad en el proceso de análisis de datos, identificación de patrones y elaboración del software de predicción.

Los autores analizan el caso puntual de la detección temprana y manejo de los pacientes tomando como insumo las imágenes radiológicas de tórax tomadas a los pacientes por lo cual en el proceso cotidiano de estas radiografías es necesario contar con radiólogos expertos en la interpretación de estas imágenes radiológicas y realizar un diagnóstico de estos pacientes, sin embargo, los autores mencionan que con el desarrollo de softwares por medio de inteligencia artificial en el cual se analicen las imágenes diagnósticas, se puede llegar a tener un diagnóstico oportuno y agilizar el proceso de diagnóstico al permitirle a los médicos tomar decisiones con mayor prontitud al generar el análisis de dichas imágenes.

López et al. (2020) destacan que la existencia de bases de datos robustas en la actualidad permite realizar mayores pruebas de rigurosidad a los algoritmos y así tener resultados con mayor certeza, esto acompañado del avance técnico de los equipos de cómputo, han permitido que se pueda avanzar en la aplicabilidad de estas técnicas de inteligencia artificial y obtener resultados positivos en su vinculación con el campo médico.

En esta misma línea, Ortega (2022) desarrolló una investigación direccionada al uso de la inteligencia artificial para detectar y clasificar enfermedades respiratorias, pero en este caso tomando como base de análisis los sonidos pulmonares, siendo la auscultación torácica uno de los aspectos empleados en el estudio clínico para la emisión de diagnóstico, por lo cual para llevar a cabo la investigación el autor hizo uso de una base de datos abiertas para el análisis y clasificación de los sonidos resultado de la auscultación, para lo cual se estudia la frecuencia de

dichos sonidos mediante el uso de una herramienta matemática para establecer con base en la frecuencia de estos sonidos, su clasificación tomando como aspectos esenciales características como “sibilancia, roncus, crepitantes, frote pleural entre otros o la ausencia de estos” (p. 11).

Con los resultados clasificados por estas características en las frecuencias de sonido, se procederá a la asociación de estos resultados con la enfermedad respiratoria correspondiente, posteriormente con esta información el autor manifiesta que se generan espectrogramas de wavelet y este es el insumo para la aplicación de la técnica de inteligencia artificial, que, en el caso puntual de la investigación, son las redes neuronales convencionales.

Por medio de estas redes neuronales convencionales, se genera una clasificación de las enfermedades respiratorias y con la aplicabilidad de los datos de sonidos existentes en la base de datos, se llevó a cabo las pruebas de entrenamiento del modelo diseñado y generación del modelo final para con ello poder tener una herramienta que identifique y clasifique oportunamente dichos sonidos producto de la auscultación y genere un diagnóstico certero de la enfermedad respiratoria que padece el paciente.

Otro de los casos de aplicación analizados, es el estudio realizado por Hoyos et al. (2023) en el cual se desarrolla un modelo de inteligencia artificial para la toma de decisiones en relación a la detección temprana de diabetes, para lo cual hicieron uso de la inteligencia artificial para el diseño de un modelo que permitiera generar un concepto de apoyo para la toma de decisiones por parte de los médicos en la detección temprana de la diabetes, tomando en cuenta la afectación generada por este enfermedad en la calidad de vida de las personas.

Para dicho modelo, los autores hicieron uso de la técnica de mapas cognitivos difusos para ello, tomaron como base de datos para el análisis la presentada por el Hospital para diabéticos de Sylhet en Bangladés, la cual es de acceso libre y les ofrece información médica de 520 individuos de los cuales 320 han sido diagnosticados con diabetes y 200 aún no son diagnosticados, aunque para el momento de análisis estos valores fueron igualados (320 con diabetes y 320 sin diabetes) generando una muestra total de 640 pacientes.

Ahora bien, en lo referente a la técnica de inteligencia artificial empleada, Hoyos et al. (2023) establece que “Los mapas cognitivos difusos son técnicas computacionales que buscan simular el

razonamiento humano, en forma similar a como lo hacen los expertos o cualquier persona con conocimientos sobre un tema en particular" (p. 113) el modelo de mapa cognitivo difuso generado en el trabajo de estos investigadores, logró obtener un 95% de exactitud, 96% de sensibilidad y 94% de especificidad, variables claves en la optimización del modelo.

Llevando a cabo una investigación aplicada sobre inteligencia artificial en el campo de la salud, se encuentra la investigación desarrollada por Bedoya et al. (2023) enfocada en la identificación de tuberculosis pulmonar por medio del uso de herramientas de la inteligencia artificial, realizando el procedimiento de análisis de datos por medio de técnicas como árboles de decisión, redes neuronales y dos métodos de ensamble. Posterior a la intervención de los datos y la fase de entrenamiento de cada modelo, se evalúa la efectividad de estos, manifestando así que con un 95.36% el método generado con el uso de la técnica de ensamble Extra Trees.

2.4.1.3. Antecedentes legales.

En Colombia, el tratamiento de datos personales está regulada por la Ley 1581 de 2012 y su Decreto Reglamentario 1377 de 2013, especialmente aquellos datos sensibles directamente relacionados con la salud, se regulan bajo esta Ley de Protección de Carácter Personal. Esta normativa garantiza la protección de los derechos fundamentales de privacidad y confidencialidad de las personas, garantizando que la utilización de información personal se realice dentro de un marco legal y ético.

La implementación de este proyecto debe incorporar un enfoque legal y ético que cumpla con los principios de protección de datos, asegurando que el modelo de inteligencia artificial no genere desigualdades de los datos ni reproducción de sesgos. Se implementarán algunas técnicas y practicas recomendadas, como la anonimización de los datos antes de su análisis para la reducción de riesgos. En conclusión, la construcción de estas herramientas debe mantener un equilibrio entre el avance tecnológico y el compromiso social, teniendo en cuenta los aspectos éticos y legales que involucra el manejo de datos sensibles en el ámbito de la salud pública.

Algunos de los principios aplicados fueron:

Legalidad: Los datos de salud del SIVIGILA serán exclusivamente utilizados para el análisis epidemiológico y la predicción de brotes de dengue. Se excluirá cualquier otro tipo de usos ajenos al propósito del proyecto.

Acceso y circulación restringida: Para el procesamiento de los datos recolectados, solo se utilizaron variables clave, como rango de edad, géneros, estratos y síntomas. Se eliminaron identificadores únicos y campos que no son relevantes para el entrenamiento de los modelos y así reducir al mínimo la cantidad de información necesaria para construir una herramienta predictiva efectiva. Los datos se anonimizaban para evitar cualquier relación directa con los titulares reforzando la privacidad.

Veracidad o calidad: A través de validaciones automáticas y manuales se garantiza que los datos e información proveniente del Sistema Nacional de Vigilancia en Salud Pública (SIVIGILA) estuviera libre de errores y actualizada. Se hizo corrección de datos para las inconsistencias o duplicaciones encontradas en los datos históricos antes del entrenamiento de los modelos de predicción.

Limitación de la finalidad: Durante la construcción del sistema se establecieron los objetivos específicos donde se define estrictamente el uso de los datos personales. Se aplica este principio para dar garantía que los datos recopilados solo serán utilizados con el propósito de análisis y predicciones de brotes de dengue en el municipio de Valledupar. Dentro del proyecto se maneja un gran volumen de datos personales y de casos clínicos, es necesario que se cumplan todos los requerimientos para la protección de los datos personales.

2.4.2 MARCO TEÓRICO

Dengue.

Es una enfermedad tipo viral que se transmite principalmente por la picadura del mosquito *Aedes Aegypti*, afectando a personas de distintas edades. Hasta la fecha se tiene conocimiento de 4 serotipos (DENV-1, DENV-2, DENV-3, DENV-4) responsables de la aparición de esta patología, su presencia actual se extiende en más de 100 países, afectando en su mayoría a poblaciones más vulnerables, lo cuál ha hecho que este sea un reto difícil para los países con más incidencia de esta enfermedad.

Clasificación del dengue.

Dentro de la clasificación podemos encontrar tres clasificaciones según los síntomas que presente el paciente.

a) **Dengue sin signos de alarma:** Esta se presenta cuando en primera instancia cuando paciente presenta fiebre, a su vez, puede incluir dolor de cabeza, náuseas y vómitos entre otros síntomas leves que se pueden manifestar

b) **Dengue con signos de alarma:** Cuando el paciente presenta síntomas mas fuertes, pueden ser señal de alarma, algunos de los síntomas que se presentan en este tipo son, vomito constante, dolor abdominal fuerte, sangrado de mucosas, disminución de plaquetas, además de otros síntomas fuertes

c) **Dengue grave:** Este tipo se manifiesta cuando el virus se presenta de forma agresiva y el paciente tiende a presentar dificultad para respirar por acumulación de liquidos, perdida de la conciencia, sangrado profuso, compromete gravemente algunos órganos.

Machine learning.

Rama de la inteligencia artificial enfocada en desarrollar o construir algoritmos y modelos capaces de aprender a partir de un conjunto de datos procesados

Modelos de Machine Learning.

a) **Regresión Lineal.**

Modelo matemático que estudia la relación entre dos o más variables para predecir un comportamiento, aplicado en diversos campos como sociales o científicos e incluso algunos que no tiene relación alguna con la tecnología

b) **Redes neuronales.**

Modelo de inteligencia artificial inspirado en el comportamiento del cerebro humano, constituida por nodos interconectados que en conjunto son conocidos como redes neuronales, capaces de aprender, reconocer y clasificar patrones de un conjunto de datos etiquetados para entrenamiento

c) **Random forest.**

Algoritmo de árboles predictores, entrenado a partir de un conjunto de muestras aleatorias, es un modelo de machine learning robusto y bastante preciso cuando se trata de una gran cantidad de datos, manejando cientos de variables de entrada

Evaluación de modelos.

Este proceso permite hacer una comparación y evaluación de los modelos de machine learning implementados mediante el uso de las métricas, que son las que nos indican que modelo cuenta con mayor precisión con respecto a los datos de entrenamiento y los datos de prueba o datos que el modelo no conoce.

SIVIGILA.

El sistema nacional de vigilancia en salud pública es el encargado de hacer revisiones sobre el comportamiento de enfermedades o patologías que se pueden presentar en el territorio nacional que puedan afectar a la población en general. Con la información recopilada en el sistema es que los entes de control pueden tomar decisiones.

2.4.3 **MARCO CONCEPTUAL**

Dengue: Enfermedad transmitida por el mosquito Aedes Aegypti, con 3 tipos de clasificación (Dengue sin signos de alarma, Dengue con signos de alarma y Dengue Grave).

SIVIGILA: Herramienta utilizada para recolectar y procesar la información sobre epidemias o enfermedades.

Inteligencia Artificial: Campo de la informática que puede realizar actividades sin la necesidad de requerir a la inteligencia humana, se utilizará para el procesamiento de datos clínicos.

Machine Learning: Rama de la inteligencia artificial con la característica de construir algoritmos para aprendizaje a partir de un conjunto de datos.

Scikit-learn: Biblioteca del lenguaje Python que incluye diversos algoritmos necesarios para el entrenamiento de modelos de aprendizaje automatico

Numpy: Biblioteca del lenguaje Python para la creación de matrices, vectores y útil para realizar operaciones matemáticas robusta.

Pandas: Biblioteca de Python usada principalmente para manipulación y análisis de datos

Dataset: Conjunto de información organizada en un formato específico, se utilizará en este caso para extraer, procesar y entrenar los diferentes modelos con la información contenida en este.

Redes Neuronales: Algoritmo adaptativo capaz de replicar el comportamiento del cerebro humano, basando su aprendizaje de un conjunto de datos.

Regresión Lineal: Modelo matemático que se puede aplicar para predecir el padecimiento de patologías mediante factores de riesgos que ya se conocen.

Random forest: Algoritmo ideal para solucionar errores que se encuentran en la clasificación y regresión, mejorando la precisión de las predicciones.

Evaluación de modelos: Se utiliza para verificar que tan bueno se comportan los modelos con los datos usados para su entrenamiento, mediante las métricas error absoluto medio (MAE), el error cuadrático medio (MSE), el error cuadrático medio raíz (RMSE) y el coeficiente de determinación (R^2).

2.5 **MARCO METODOLÓGICO**

El presente proyecto adopta un enfoque analítico sustentado en el uso de la metodología CRISP-

DM (Cross Industry Standard Process for Data Mining), ampliamente reconocida en proyectos relacionados a ciencia y minería de datos. Esta metodología permite organizar de una forma estructurada todo el compendio de etapas que involucra el desarrollo de modelos de inteligencia Artificial, desde la comprensión del problema hasta su implementación y evaluación.

Métodos y Técnicas de recopilación de información

Se utilizará el método analítico-deductivo, apoyado en técnicas de minería de datos y aprendizaje automatico con el fin de explorar, procesar y analizar datos epidemiológicos del SIVIGILA. Las principales técnicas a emplear son:

- **Análisis exploratorio de datos (EDA):** útil para examinar patrones de incidencia del virus del dengue y su comportamiento a través del tiempo.
- **Limpieza y transformación de datos:** estandarización, normalización y codificación de variables.
- **Modelado predictivo supervisado:** junto con algoritmos de clasificación como Random Forest y redes neuronales
- **Evaluación de Modelos:** mediante métricas como accuracy, F1-score.

Estas técnicas serán aplicadas progresivamente en las distintas etapas de la metodología CRISP-DM.

El enfoque metodológico se alinea con las siguientes etapas de CRISP-DM:

- **Comprensión del Negocio:** análisis del contexto epidemiológico del virus del dengue en el departamento del Cesar.
- **Comprensión de los datos:** exploración inicial del (los) dataset extraidos del SIVIGILA, identificación de variables relevantes y análisis estadístico descriptivo.
- **Preparación de los datos:** limpieza de datos, tratamiento de valores nulos, casteo de variables categóricas y selección de atributos para modelos.
- **Modelado:** entrenamiento de modelos de clasificación junto con diferentes técnicas de machine learning.

- **Evaluación:** validación cruzada, análisis de desempeño y comparación de modelos a través de métricas
- **Implementación:** desarrollo de un aplicativo funcional de predicción del tipo de dengue a través de sintomatología del paciente**.**

Como instrumentos de recolección de información y desarrollo del aplicativo se tiene lo siguiente:

- **Notebooks de Google Colab (Python):** útiles para limpieza, transformación y análisis de los datos iniciales (librerías: pandas, scikit-learn, seaborn, matplotlib)
- Documentos técnicos del SIVIGILA y reportes estadísticos
- **Herramientas de visualización:** Tableros diseñados en Power BI
- **Plataformas de procesamiento en la nube**: Power BI Groups

El estudio relevante para el proyecto se llevará a cabo en el municipio de Valledupar, departamento del Cesar, una de las regiones con mayor incidencia de brotes del virus del dengue en los últimos años. La población objetivo se compone de casos de dengue reportados entre los años 2019 y 2024 registrados en el sistema nacional de vigilancia epidemiológica.

Variables consideradas

Las variables/categorías clave que se analizarán incluyen:

- **Demográficas:** edad, sexo, estrato, municipio.
- **Clínicas:** sintomatología del paciente (dolor de cabeza, plaquetas, entre otras).
- **Temporales:** fecha de ingreso, fecha notificación, entre otras.
- **Etiqueta/Objetivo:** tipo de dengue confirmado por diagnóstico clínico.

SECCIÓN III: Desarrollo Científico-Tecnológico

3.1 DESARROLLO DE LAS FASES DE LA METODOLOGÍA DE SISTEMAS PROPUESTA

La metodología implementada para este proyecto fue CRISP-DM (Cross-Industry Standard Process for Data Mining), una metodología cíclica y estructurada para proyectos de ciencia de datos, compuesta por seis fases o etapas principales: comprensión del negocio, comprensión de los datos, preparación, modelado, evaluación y despliegue.

A continuación, se detalla cada una de las etapas y actividades realizadas a lo largo del proyecto.

3.1.1 Fase 1: Comprensión del Negocio

La fase de comprensión del negocio tuvo como propósito definir el problema a abordar, los objetivos del proyecto y el impacto esperado en el contexto de la salud pública.

El dengue, como enfermedad viral transmitida por el mosquito *Aedes aegypti*, representa una problemática creciente en el departamento del Cesar, especialmente en la ciudad de Valledupar, donde se han evidenciado incrementos significativos en el número de casos reportados en los últimos años. La clasificación oportuna de los tipos de dengue (sin signos de alarma, con signos de alarma y dengue grave) resulta fundamental para la toma de decisiones clínicas y la priorización de la atención médica en centros de salud

Sin embargo, los métodos tradicionales de análisis presentan limitaciones frente al volumen y complejidad de los datos epidemiológicos disponibles en el Sistema Nacional de Vigilancia en Salud Pública (SIVIGILA), lo cual dificulta la identificación oportuna de patrones y la anticipación de brotes del virus. Bajo este contexto, el objetivo del proyecto consistió en diseñar un modelo de inteligencia artificial basado en técnicas de *machine learning*, capaz de clasificar automáticamente los tipos de dengue a partir de variables clínicas y epidemiológicas.

El problema fue abordado como una tarea de clasificación supervisada, donde la variable objetivo corresponde al tipo de dengue diagnosticado. Como criterios de éxito del modelo, se definieron métricas de desempeño como la precisión, el recall y el F1-score, dando especial importancia a la

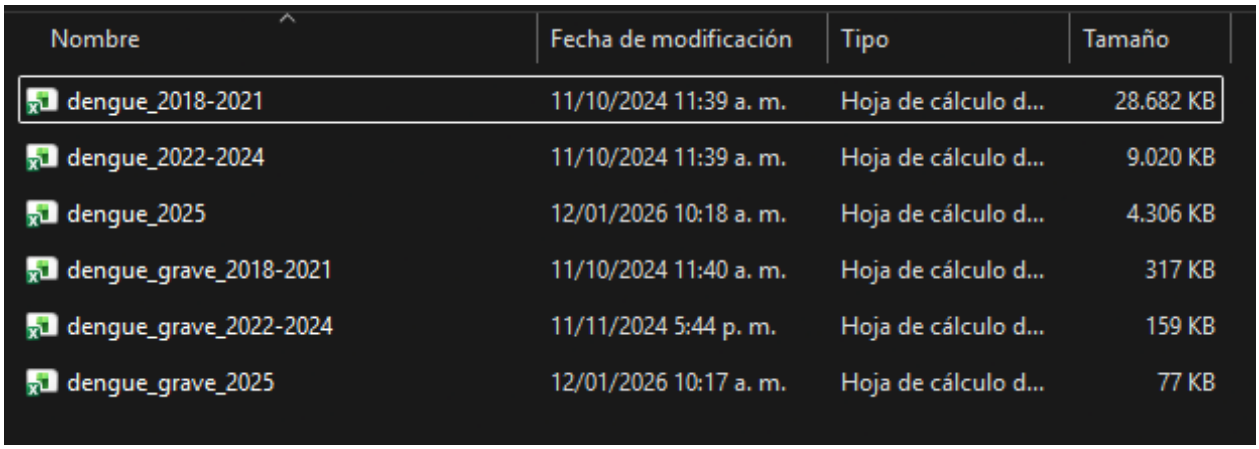
correcta identificación de los casos de dengue grave, esto debido a su alto impacto en la salud y el bienestar de los pacientes infectados.

3.1.2 Fase 2: Comprensión de los Datos

En esta fase se realizó el análisis inicial del conjunto de datos proveniente del Sistema Nacional de Vigilancia en Salud Pública (SIVIGILA), el cual contiene registros epidemiológicos de casos de dengue en el departamento del Cesar, abarcando un periodo comprendido entre los años 2018 y 2025.

Origen de los datos: Sistema Nacional de Vigilancia en Salud Pública (SIVIGILA).
Datos fuente: Seis (6) archivos en formato excel (.xls, .xlsx)

- **dengue_ 2018-2021.xls** (15489 registros)
- **dengue_ 2022-2024.xlsx** (13476 registros)
- **dengue_ 2025.xlsx** (7376 registros)
- **dengue grave 2018-2021.xls** (168 registros)
- **dengue grave 2022-2024.xls** (314 registros)
- **dengue grave 2025.xlsx** (124 registros)

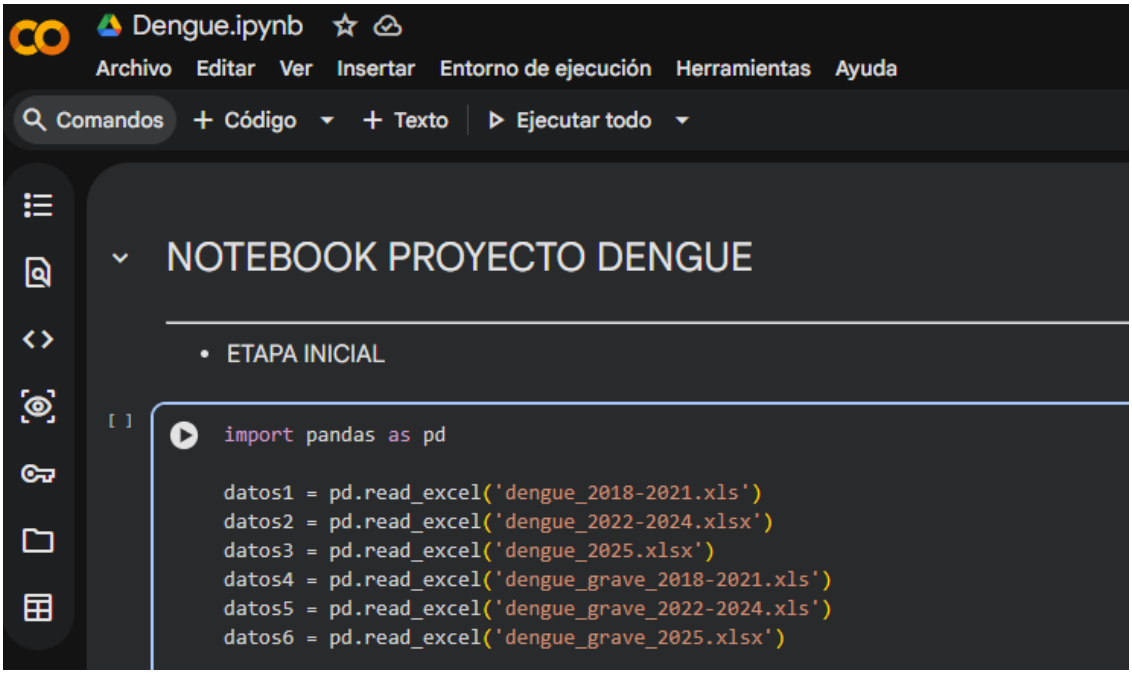


Nombre	Fecha de modificación	Tipo	Tamaño
dengue_2018-2021	11/10/2024 11:39 a. m.	Hoja de cálculo d...	28.682 KB
dengue_2022-2024	11/10/2024 11:39 a. m.	Hoja de cálculo d...	9.020 KB
dengue_2025	12/01/2026 10:18 a. m.	Hoja de cálculo d...	4.306 KB
dengue_grave_2018-2021	11/10/2024 11:40 a. m.	Hoja de cálculo d...	317 KB
dengue_grave_2022-2024	11/11/2024 5:44 p. m.	Hoja de cálculo d...	159 KB
dengue_grave_2025	12/01/2026 10:17 a. m.	Hoja de cálculo d...	77 KB

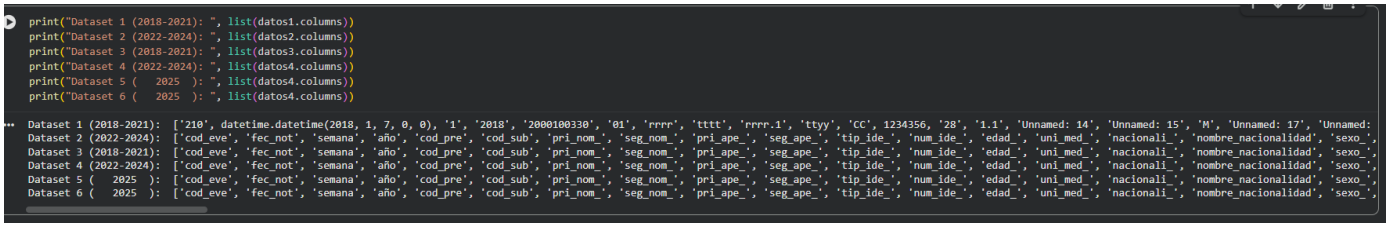
Se identificaron variables relevantes para el análisis, tales como edad, género, estrato socioeconómico, sintomatología, clasificación del caso, ubicación geográfica y fechas de notificación.

Posteriormente, se llevó a cabo un análisis exploratorio de datos (EDA), con el fin de comprender la distribución y comportamiento de las variables, esto se realizó en conjunto a través de una herramienta colaborativa de la suite de Google, Google Colab, que funciona como un entorno de ejecución por celdas como un notebook de Python.

En este entorno, se cargaron todos los archivos anteriormente mencionados:



Se listan las columnas de los distintos archivos convertidos dataframes de pandas, donde se observa que el archivo del periodo 2018-2021 no contenía encabezados



Debido a esto, se procedió a mapeárselos con base en los demás dataframes, resultando de la siguiente manera:

```
columnas = [
    'cod_eve','fec_not','semana','año','cod_pre','cod_sub','pri_nom_','seg_nom_','pri_ape_','seg_ape_',
    'tip_ide_','num_ide_','edad_','uni_med_','nacionali_','nombre_nacionalidad','sexo_','iden_gener',
    'otra_ident','orient_sex','otra_orien','cod_pais_o','cod_dpto_o','cod_mun_o','area_','localidad_',
    'cen_pobla_','vereda_','bar_ver_','dir_res_','lat_dir','long_dir','ocupacion_','tip_ss_','cod_ase_',
    'per_etn_','nom_grupo_','estrato_','gp_discapa','gp_desplaz','gp_migrant','gp_carcela','gp_gestan',
    'sem_ges_','gp_indigen','gp_pobicbf','gp_mad_com','gp_desmovi','gp_psiquia','gp_vic_vio','gp_otros',
    'fuente_','cod_pais_r','cod_dpto_r','cod_mun_r','fec_con_','ini_sin_','tip_cas_','pac_hos_','fec_hos_',
    'con_fin_','fec_def_','ajuste_','telefono_','fecha_nto_','cer_def_','cbmte_','uni_modif','nuni_modif',
    'fec_arc_xl','nom_dil_f_','tel_dil_f_','fec_aju_','nit_upgd','fm_fuerza','fm_unidad','fm_grado',
    'version','desplazami','cod_pais_d','cod_dep_d','cod_mun_d','famantdngu','direclabor','fiebre',
    'cefalea','dolrretroo','malgias','artralgia','erupcionr','dolor_abdo','vomito','diarrea','somnolenci',
    'hipotensio','hepatomeg','hem_mucosa','hipotermia','aum_hemato','caida_plaq','acum_liqui','extravasac',
    'hemorr_hem','choque','daño_organ','muesttejid','mueshigado','muesbazo','muespulmon','muescerebr',
    'muesmiocar','muesmedula','muesriñon','clasfinal','conducta','nom_eve','nom_upgd','npais_proce',
    'ndep_proce','nmun_proce','npais_resi','ndep_resi','nmun_resi','ndep_notif','nmun_notif','FechaHora'
]

datos1_sin_nombre = pd.read_excel('dengue_2018-2021.xls', header=None)
datos1_sin_nombre.columns = columnas
datos1 = datos1_sin_nombre

print("Columnas de datos1:", list(datos1.columns))
print("Columnas de datos2:", list(datos2.columns))
print("Columnas de datos3:", list(datos3.columns))
print("Columnas de datos4:", list(datos4.columns))
print("Columnas de datos5:", list(datos5.columns))
print("Columnas de datos6:", list(datos6.columns))

# datos1.columns.equals(datos2.columns) and datos1.columns.equals(datos3.columns) and datos1.columns.equals(datos4.columns)
```

```
Columnas de datos1: ['cod_eve', 'fec_not', 'semana', 'año', 'cod_pre', 'cod_sub', 'pri_nom_', 'seg_nom_', 'pri_ape_', 'seg_ape_', 'tip_ide_', 'num_ide_', 'edad_',
Columnas de datos2: ['cod_eve', 'fec_not', 'semana', 'año', 'cod_pre', 'cod_sub', 'pri_nom_', 'seg_nom_', 'pri_ape_', 'seg_ape_', 'tip_ide_', 'num_ide_', 'edad_',
Columnas de datos3: ['cod_eve', 'fec_not', 'semana', 'año', 'cod_pre', 'cod_sub', 'pri_nom_', 'seg_nom_', 'pri_ape_', 'seg_ape_', 'tip_ide_', 'num_ide_', 'edad_',
Columnas de datos4: ['cod_eve', 'fec_not', 'semana', 'año', 'cod_pre', 'cod_sub', 'pri_nom_', 'seg_nom_', 'pri_ape_', 'seg_ape_', 'tip_ide_', 'num_ide_', 'edad_',
Columnas de datos5: ['cod_eve', 'fec_not', 'semana', 'año', 'cod_pre', 'cod_sub', 'pri_nom_', 'seg_nom_', 'pri_ape_', 'seg_ape_', 'tip_ide_', 'num_ide_', 'edad_',
Columnas de datos6: ['cod_eve', 'fec_not', 'semana', 'año', 'cod_pre', 'cod_sub', 'pri_nom_', 'seg_nom_', 'pri_ape_', 'seg_ape_', 'tip_ide_', 'num_ide_', 'edad_',
```

Donde todos los dataframes tienen las mismas columnas.

A partir de este punto, se decidió unificar toda la data en un solo dataframe, al que nombraremos DatasetFinal.xlsx y posteriormente será leído con el objetivo de seleccionar sólo las columnas relevantes para el estudio:

```
datos_unificado = pd.concat([datos1, datos2, datos3, datos4, datos5, datos6], ignore_index=True)
print(datos_unificado.info())

columnas_seleccionadas = datos_unificado[['cod_eve', 'fec_not', 'edad_', 'sexo_', 'cod_dpto_r', 'cod_mun_r',
                                          'fec_con_', 'ini_sin_', 'tip_cas_', 'pac_hos_', 'fec_hos_', 'con_fin_',
                                          'estrato_', 'fiebre', 'cefalea', 'dolrretroo', 'malgias', 'artralgia',
                                          'erupcionr', 'dolor_abdo', 'vomito', 'diarrea', 'somnolenci', 'hipotensio',
                                          'hepatomeg', 'hem_mucosa', 'hipotermia', 'aum_hemato', 'caida_plaq',
                                          'acum_liqui', 'extravasac', 'hemorr_hem', 'choque', 'daño_organ', 'clasfinal',
                                          'nom_eve'
                                          ]]

print(columnas_seleccionadas.info())
dataset_prueba = columnas_seleccionadas.copy()
dataset_prueba.to_excel('DatasetFinal.xlsx', index=False)
```

Información del dataframe

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 36948 entries, 0 to 36947
Data columns (total 36 columns):
 #   Column      Non-Null Count  Dtype  
---  ------      --------------  -----  
 0   cod_eve     36948 non-null  int64  
 1   fec_not     36948 non-null  object 
 2   edad_       36948 non-null  int64  
 3   sexo_       36948 non-null  object 
 4   cod_dpto_r  36948 non-null  int64  
 5   cod_mun_r   36947 non-null  object 
 6   fec_con_    36948 non-null  object 
 7   ini_sin_    36948 non-null  object 
 8   tip_cas_    36948 non-null  int64  
 9   pac_hos_    36948 non-null  int64  
 10  fec_hos_    36948 non-null  object 
 11  con_fin_    36948 non-null  int64  
 12  estrato_    36010 non-null  float64
 13  fiebre      36939 non-null  float64
 14  cefalea     36938 non-null  float64
 15  dolrretroo  36938 non-null  float64
 16  malgias     36939 non-null  float64
 17  artralgia   36939 non-null  float64
 18  erupcionr   36938 non-null  float64
 19  dolor_abdo  36938 non-null  float64
 20  vomito      36937 non-null  float64
 21  diarrea     36937 non-null  float64
 22  somnolenci  36937 non-null  float64
 23  hipotensio  36937 non-null  float64
 24  hepatomeg   36937 non-null  float64
 25  hem_mucosa  36937 non-null  float64
 26  hipotermia  36937 non-null  float64
 27  aum_hemato  36937 non-null  float64
 28  caida_plaq  36937 non-null  float64
 29  acum_liqui  36936 non-null  float64
 30  extravasac  14738 non-null  float64
 31  hemorr_hem  14738 non-null  float64
 32  choque      14738 non-null  float64
 33  daño_organ  14738 non-null  float64
 34  clasfinal   36939 non-null  float64
 35  nom_eve     36948 non-null  object 
dtypes: float64(23), int64(6), object(7)
memory usage: 10.1+ MB
```

Se realiza el filtrado de datos por departamento, dejando solo el código 20 del campo **‘cod_dpto_r’**, este corresponde al departamento del Cesar.

```
dataset_prueba['cod_dpto_r'].value_counts().sort_index()
print(f'Dataset sin filtrar por departamento: {dataset_prueba.shape}')
dataset_prueba = dataset_prueba[dataset_prueba['cod_dpto_r'] == 20]
print(f'Dataset filtrado solo para el departamento del cesar: {dataset_prueba.shape}')
```

```
Dataset sin filtrar por departamento: (36948, 36)
Dataset filtrado solo para el departamento del cesar: (36661, 36)
```

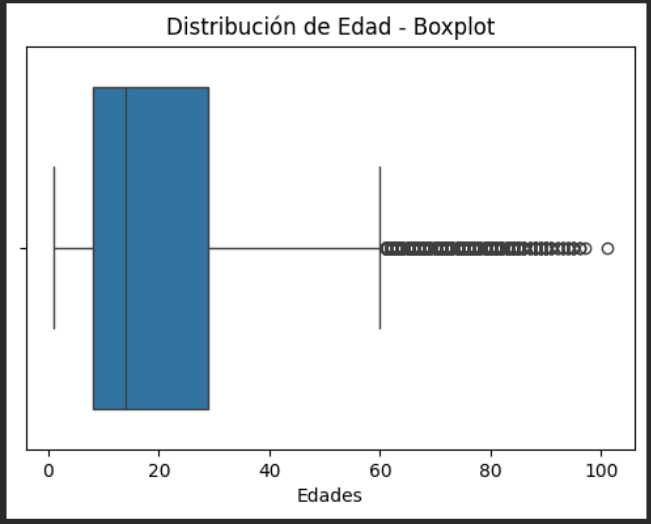
Se identifican outliers (valores atípicos) en variables como la edad, donde luego de los 60 años son casos muy dispersos a lo usual.

Verificar Outliers en variables numericas

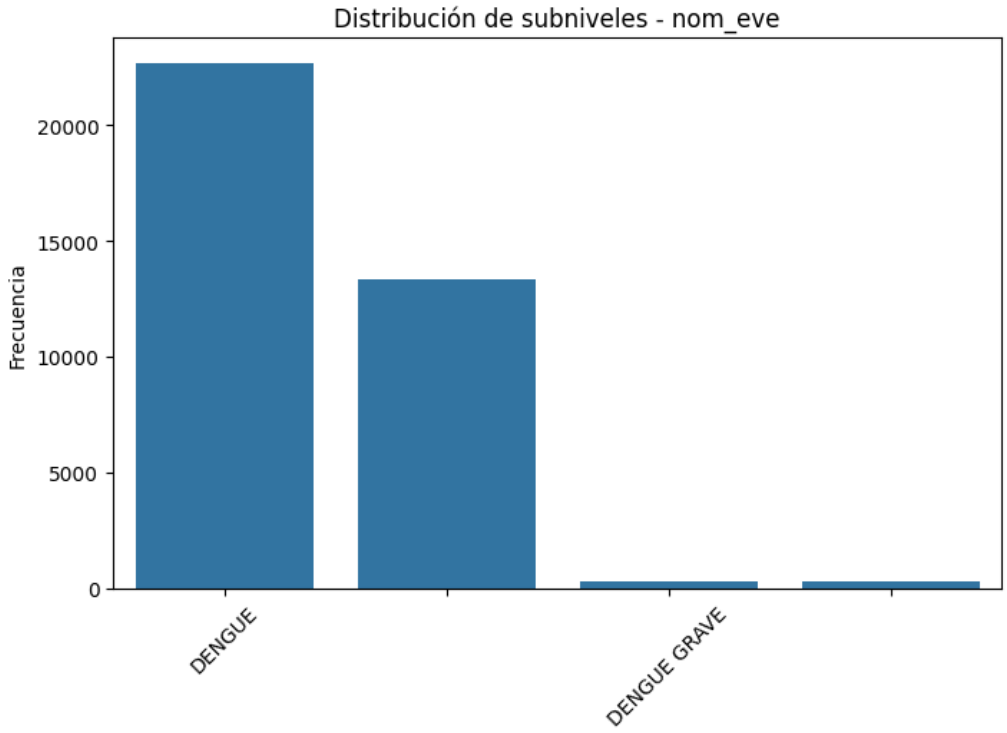
- Edad ('edad_')

```
import seaborn as sns
import matplotlib.pyplot as plt
```

```
plt.figure(figsize=(6,4))
sns.boxplot(x=dataset_prueba['edad_'])
plt.xlabel('Edades')
plt.title('Distribución de Edad - Boxplot')
plt.show()
```



Se identifican inconsistencias en variables categóricas, por ejemplo, en los subniveles de la variable nombre del evento ('nom_eve'), se observó que habían más de dos subniveles, lo cual puede corresponder a un error de digitación o espacios en blanco:




```
dataset_prueba['nom_eve'].value_counts()
```

```
                count
nom_eve
DENGUE          22656
DENGUE          13338
DENGUE GRAVE      309
DENGUE GRAVE      290

dtype: int64
```

Este análisis exploratorio permitió:

- Analizar la incidencia por grupos etarios, destacando una mayor afectación en población infantil y juvenil.
- Examinar la proporción de casos según tipo de dengue.
- Detectar valores atípicos, datos faltantes y posibles inconsistencias en los registros, que serán tratadas en la siguiente fase

Adicionalmente, se evidenció la presencia de un desbalance en la variable objetivo, donde los casos de dengue grave representan una proporción menor en comparación con las otras categorías, lo cual constituyó un desafío para el entrenamiento de modelos de clasificación.

Sin embargo, esta fase permitió establecer una comprensión clara de la calidad, estructura y limitaciones de los datos, sentando las bases para su posterior preparación.

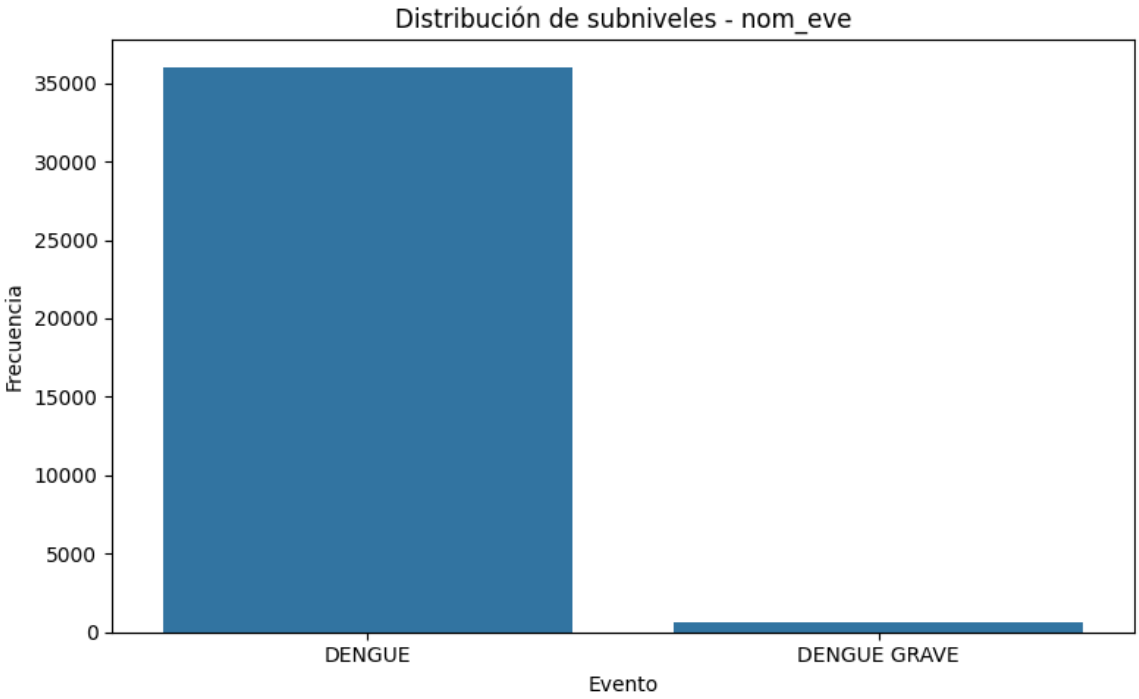
3.1.3 Fase 3: Preparación de los Datos

La fase de preparación de los datos comprendió todas las actividades necesarias para transformar el conjunto de datos en un formato adecuado y legible para el entrenamiento de los modelos de machine learning.

Inicialmente, se realizó un proceso de limpieza de datos que incluyó la eliminación de registros duplicados, el tratamiento de valores nulos y la corrección de inconsistencias en variables categóricas y numéricas, como podemos ver a continuación:

- Se aplicaron métodos de strip y upper que eliminan espacios innecesarios y se convierten mayúscula todos los valores, obteniendo así solo los dos subniveles que necesarios del campo **nom_eve** para el entrenamiento de los modelos

```
dataset_prueba['nom_eve'] = (
    dataset_prueba['nom_eve']
    .str.strip()
    .str.upper()
)
```



- La variable **'estrato'** tiene varios valores vacíos (929), al realizar un conteo de subniveles se determinó que esta variable se puede imputar con la moda, de esta manera esos datos faltantes se llenaron con el valor que más se repite

```
moda = df['estrato_'].mode()[0]
df['estrato_'] = df['estrato_'].fillna(moda)
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 36592 entries, 0 to 36592
Data columns (total 34 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   cod_eve     36592 non-null  int64
 1   fec_not     36592 non-null  datetime64[ns]
 2   edad_       36592 non-null  int64
 3   sexo_       36592 non-null  object
 4   cod_mun_r   36592 non-null  object
 5   fec_con_    36592 non-null  datetime64[ns]
 6   ini_sin_    36592 non-null  datetime64[ns]
 7   tip_cas_    36592 non-null  int64
 8   pac_hos_    36592 non-null  int64
 9   fec_hos_    27423 non-null  datetime64[ns]
 10  estrato_    36592 non-null  float64
 11  fiebre      36583 non-null  float64
 12  cefalea     36582 non-null  float64
 13  dolrretroo  36582 non-null  float64
 14  malgias     36583 non-null  float64
 15  artralgia   36583 non-null  float64
```

- Se binarizaron las variables features o los síntomas, ya que estaban codificados de la siguiente manera:

  1 = sintoma presente, 2 = sintoma no presente

  Esto no es lo más óptimo para machine learning, por lo cual se binarizó de la siguiente manera:

  0 = sintoma no presente, 1 = sintoma presente

```
cols_sintomas = [
    'fiebre','cefalea','dolrretroo','malgias','artralgia','erupcionr',
    'dolor_abdo','vomito','diarrea','somnolenci','hipotensio',
    'hepatomeg','hem_mucosa','hipotermia','aum_hemato',
    'caida_plaq','acum_liqui','extravasac','hemorr_hem',
    'choque','daño_organ'
]
```

```
df[cols_sintomas] = df[cols_sintomas].fillna(2)
print("Imputancion correcta!!")
df[cols_sintomas] = df[cols_sintomas].replace({1:1, 2:0})
print("Recodificación completa!!")
```

```
Imputancion correcta!!
Recodificación completa!!
```

- La variable "fec_hos_", es una variable con fecha cruda, lo cual no se puede emplear en los modelos de machine learning, de igual manera se binarizó la variable con la finalidad de indicar:
  1 = si el registro tiene fecha de hospitalización
  0 = registro sin fecha de hospitalización

```
df["tiene_fec_hos"] = df["fec_hos_"].notna().astype(int)
df["tiene_fec_hos"].value_counts( )
```

```
                count
tiene_fec_hos
1               27410
0                9165

dtype: int64
```

Posteriormente, se aplicaron transformaciones sobre otras variables tales como:

- Codificación de variables categóricas mediante técnicas como *Label Encoding* y/o *One-Hot Encoding*.
- Normalización o estandarización de variables numéricas, cuando fue requerido por los modelos utilizados.
- Selección de variables relevantes, descartando aquellas que no aportaban valor predictivo o que podían generar ruido en el modelo.

Adicional, se realizó un proceso de selección de variables específicas para el entrenamiento de los modelos (feature) y se excluyeron variables que posiblemente serían causantes de data leakage. Estas variables son relevantes, pero al tener alta correlación con la variable objetivo producen alto sesgo en los modelos.

```
LEAKAGE_VARS = [
    "clasfinal", "tiene_fec_hos", "pac_hos_", "fec_hos",
    "hipotensio", "choque", "hipotermia", "hem_mucosa",
    "hemorr_hem", "aum_hemato", "caida_plaq", "acum_liqui",
    "extravasac", "daño_organ"
]

FEATURE_VARS = [
    "fiebre", "cefalea", "dolrretroo", "malgias",
    "artralgia", "erupcionr", "dolor_abdo", "vomito",
    "diarrea", "somnolenci", "hepatomeg"
]

ANALYSIS_VARS = ["edad_", "sexo__M", "estrato_", "cod_mun_r", "fec_not"]

TARGET = "dengue_grave"
COD_VALLEDUPAR = 1

THRESHOLDS = [0.3, 0.4, 0.5]
RANDOM_STATE = 42

print("Configuración global lista")
```

Finalmente, el conjunto de datos fue dividido en subconjuntos de entrenamiento y prueba, generalmente bajo una proporción de 80% para entrenamiento y 20% para prueba, garantizando así la evaluación objetiva del desempeño de los modelos.

3.1.4 Fase 4: Modelado

En esta fase se implementaron diferentes algoritmos de machine learning con el objetivo de construir modelos capaces de clasificar los tipos de dengue a partir de los datos preparados.

Antes de realizar el proceso de entrenamiento de los modelos, se evidenció que el conjunto de datos se encontraba extremadamente desbalanceado, por lo cual, fue necesario aplicar el método smote. Teniendo en cuenta que más del 98% del conjunto de datos corresponde a “dengue” y tan sólo el 1.64% corresponde a “dengue grave”, se aplicó el método smote para que el modelo no solo aprenda a predecir dengue e ignore dengue grave al ser minoría en los datos. Este método se aplicó solo en los datos de entrenamiento, de esta forma se crean datos sintéticos que permiten balancear los datos para el entrenamiento y que el modelo aprenda a distinguir dengue grave

```
available_features = [f for f in FEATURE_VARS if f in df.columns]
print(f" Features disponibles ({len(available_features)}): {available_features}")

X = df[available_features].copy()
y = df[TARGET].copy()

X = X.fillna(X.mode().iloc[0])

X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=RANDOM_STATE, stratify=y
)
print(f" Train: {X_train.shape} | Test: {X_test.shape}")
print(f"   Positivos en train: {y_train.sum()} ({y_train.mean()*100:.2f}%)")
```

```
smote = SMOTE(random_state=RANDOM_STATE)
X_train_res, y_train_res = smote.fit_resample(X_train, y_train)
print(f" Tras SMOTE — Train: {X_train_res.shape} | Positivos: {y_train_res.sum()}")

scaler = StandardScaler()
X_train_sc = scaler.fit_transform(X_train_res)
X_test_sc  = scaler.transform(X_test)

features = X.columns.tolist()
joblib.dump(features, "features.pkl")
joblib.dump(scaler,   "scaler.pkl")
print(" scaler.pkl y features.pkl guardados")
```

```
Tras SMOTE — Train: (57562, 11) | Positivos: 28781
scaler.pkl y features.pkl guardados
```

Se entrenaron diversos modelos de clasificación, entre los cuales se destacan:

- Random Forest
- Redes Neuronales Artificiales
- Regresión Logística

```python
models = {
    "Logistic Regression": LogisticRegression(
        class_weight="balanced", max_iter=1000, random_state=RANDOM_STATE
    ),
    "Random Forest": RandomForestClassifier(
        n_estimators=200, class_weight="balanced",
        max_depth=10, random_state=RANDOM_STATE, n_jobs=-1
    ),
    "MLP Neural Network": MLPClassifier(
        hidden_layer_sizes=(64, 32), max_iter=300,
        early_stopping=True, validation_fraction=0.1,
        random_state=RANDOM_STATE, n_iter_no_change=15
    )
}

trained_models = {}
for name, model in models.items():
    model.fit(X_train_sc, y_train_res)
    trained_models[name] = model
    print(f" {name} entrenado")
```

Cada uno de estos modelos fue seleccionado debido a sus características particulares. Por ejemplo, Random Forest se destaca por su robustez ante datos ruidosos y su capacidad para manejar múltiples variables, mientras que las redes neuronales permiten capturar relaciones no lineales complejas en los datos.

Adicionalmente, se aplicaron técnicas de validación, como la validación cruzada, para garantizar la estabilidad y generalización de los modelos.

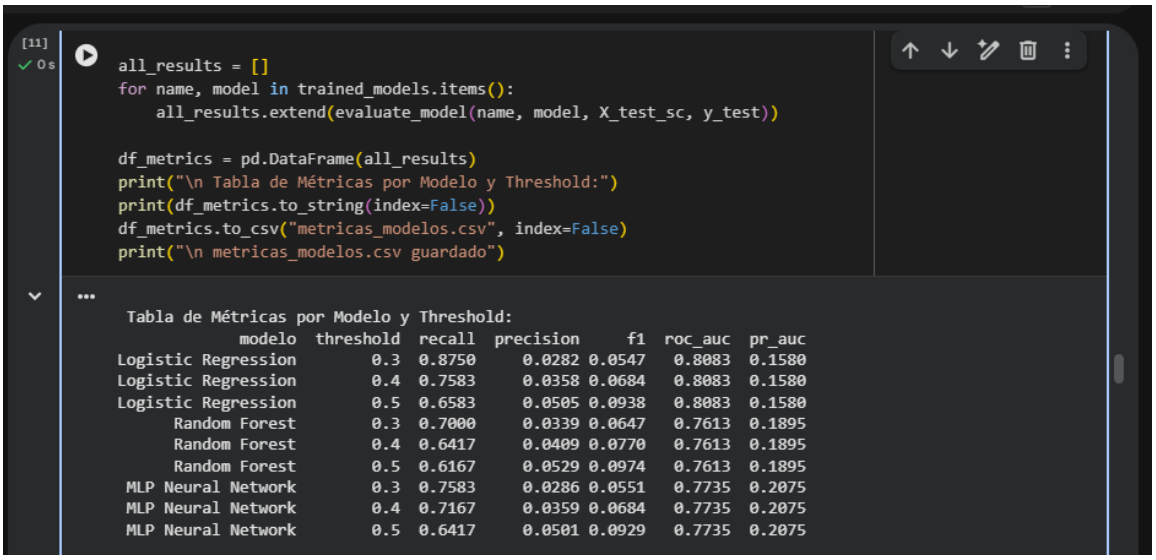
3.1.5 Fase 5: Evaluación

La fase de evaluación se planteó como objetivo medir el desempeño de los modelos entrenados y posteriormente seleccionar el más adecuado para su implementación.
Para ello, se utilizaron diversas métricas de evaluación, entre las cuales se incluyen:

- Precision
- Recall
- F1-score
- ROC_AUC

Estas métricas permitieron realizar un análisis sobre el comportamiento de los distintos modelos tanto de forma general e individual

De igual manera, se llevó a cabo un análisis del desempeño de los distintos modelos utilizando diferentes umbrales de decisión (Threshold) en las probabilidades predichas. Con esto se busca analizar el comportamiento del modelo en términos de Recall, precision y F1-Score, ajustando el umbral de decisión y no sólo fijando el umbral estándar (0.5), a partir de esto se busca una configuración óptima que favorezca la reducción de falsos positivos



```
[11]
✓ 0s
all_results = []
for name, model in trained_models.items():
    all_results.extend(evaluate_model(name, model, X_test_sc, y_test))

df_metrics = pd.DataFrame(all_results)
print("\n Tabla de Métricas por Modelo y Threshold:")
print(df_metrics.to_string(index=False))
df_metrics.to_csv("metricas_modelos.csv", index=False)
print("\n metricas_modelos.csv guardado")
```

```
 Tabla de Métricas por Modelo y Threshold:
              modelo  threshold  recall  precision     f1  roc_auc  pr_auc
Logistic Regression        0.3  0.8750     0.0282 0.0547   0.8083  0.1580
Logistic Regression        0.4  0.7583     0.0358 0.0684   0.8083  0.1580
Logistic Regression        0.5  0.6583     0.0505 0.0938   0.8083  0.1580
      Random Forest        0.3  0.7000     0.0339 0.0647   0.7613  0.1895
      Random Forest        0.4  0.6417     0.0409 0.0770   0.7613  0.1895
      Random Forest        0.5  0.6167     0.0529 0.0974   0.7613  0.1895
 MLP Neural Network        0.3  0.7583     0.0286 0.0551   0.7735  0.2075
 MLP Neural Network        0.4  0.7167     0.0359 0.0684   0.7735  0.2075
 MLP Neural Network        0.5  0.6417     0.0501 0.0929   0.7735  0.2075
```

A partir de los resultados obtenidos, se realizó una comparación entre los modelos implementados, seleccionando aquel que presentó el mejor equilibrio entre las métricas evaluadas y una mayor capacidad de generalización, obteniendo como resultado que el modelo de **Regresión logística** fue el seleccionado para la implementación de la fase de predicción debido a que cuenta con el recall más alto (0.875) en lo que respecta a la detección de casos de dengue grave.

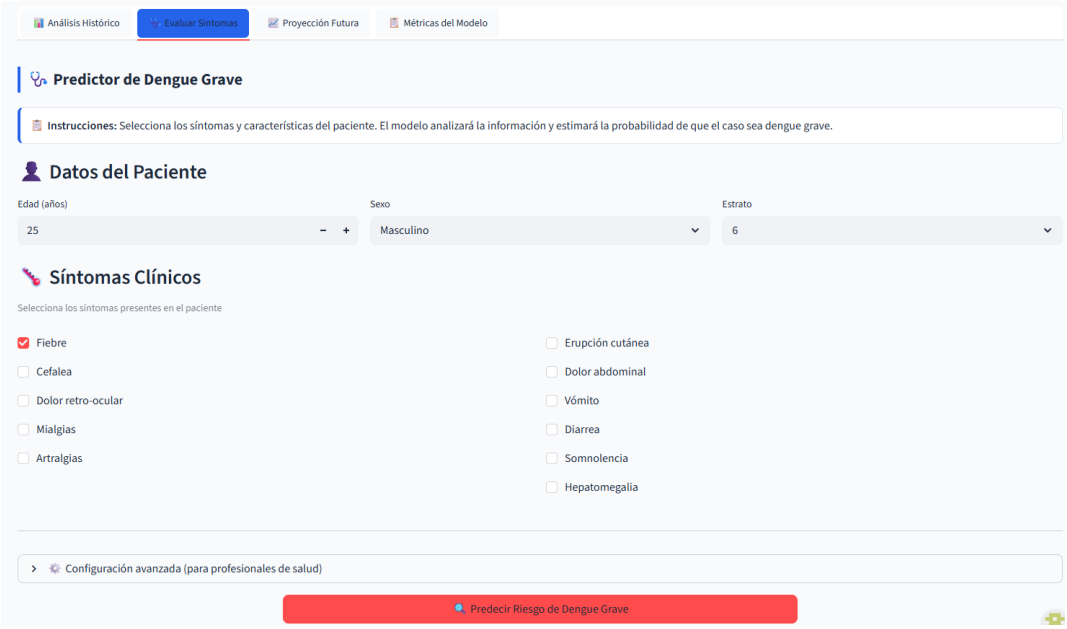
Aunque el modelo **random forest** se comporta de una mejor manera con el dataset extremadamente desbalanceado, este mismo pierde muchos casos denominados **graves positivos** y en lo que respecta a este proyecto, principalmente se busca poder hacer una detección temprana de pacientes con riesgo de dengue grave.

3.1.6 Fase 6: Despliegue

La fase de despliegue consistió en la implementación de los resultados del modelo en herramientas que permitan su uso práctico en el contexto de la salud pública.

Como producto final del proyecto, se desarrolló un **aplicativo de apoyo clínico**, el cual integra el modelo de machine learning seleccionado, junto con una feature que permite ingresar la sintomatología de un paciente, obteniendo como salida la clasificación del tipo de dengue, junto con su probabilidad asociada.

Interfaz Principal del aplicativo



Predicción de Dengue Grave a través de Datos y Sintomatología del Paciente

Este sistema tiene como finalidad apoyar a los profesionales de la salud en la toma de decisiones, mejorar la atención oportuna de los pacientes y contribuir a la prevención de brotes mediante el análisis predictivo.

Adicionalmente, se planteó el desarrollo un tablero de control (dashboard) en **Power BI**, el cual permite visualizar de manera interactiva:

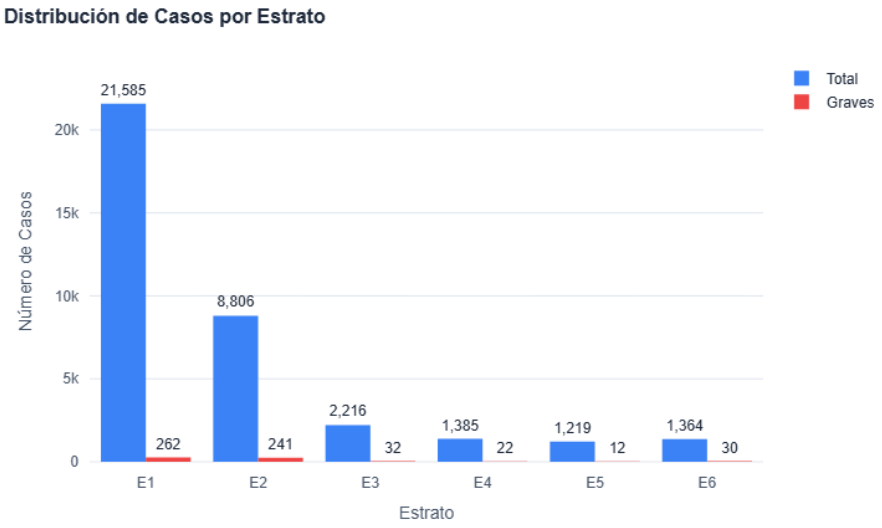
- La evolución de los casos de dengue en el tiempo
- La distribución geográfica de los casos
- La clasificación de los tipos de dengue
- Indicadores clave para la toma de decisiones

Finalmente, el despliegue de estas herramientas representa un avance significativo en la integración de la inteligencia artificial en el ámbito de la salud pública, aportando soluciones innovadoras para el análisis epidemiológico en el departamento del Cesar.

3.2 ANÁLISIS DE RESULTADOS Y DISCUSIÓN

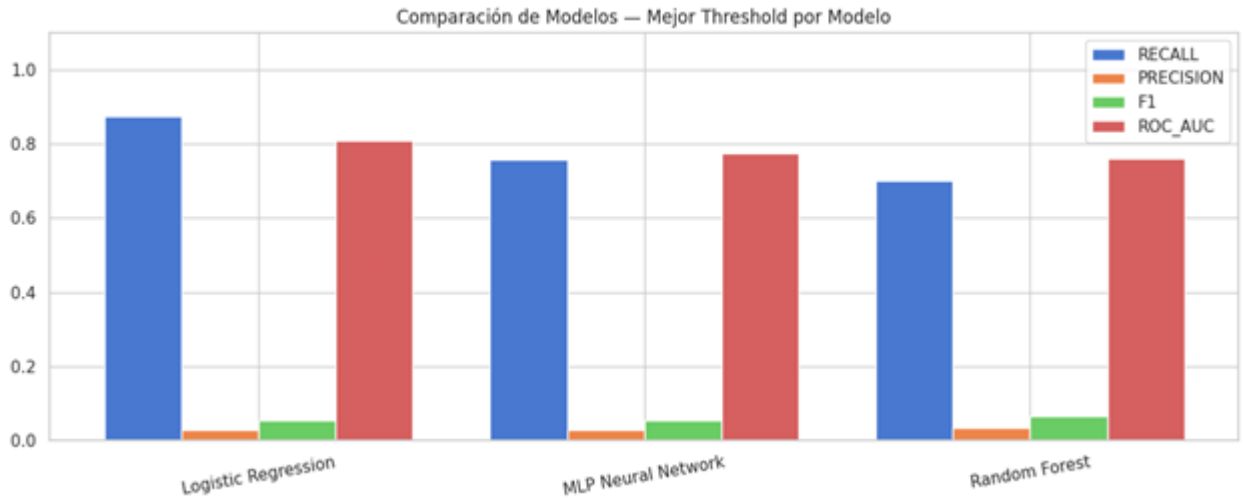
El desarrollo del presente proyecto permitió evidenciar el potencial de las técnicas de inteligencia artificial, particularmente el uso de machine learning, con la finalidad de abordar problemáticas asociadas a la salud pública como puede ser la clasificación de los tipos de dengue a partir de datos epidemiológicos del Sistema Nacional de Vigilancia en Salud Pública (SIVIGILA).

A partir del análisis exploratorio de los datos (EDA), se identificaron patrones relevantes en la distribución de los casos, destacándose una mayor incidencia en población infantil y juvenil, así como una fuerte concentración de casos en estratos socioeconómicos bajos. Estos hallazgos son coherentes con lo planteado en la descripción situacional del problema, donde se evidencia la vulnerabilidad de estos grupos poblacionales frente al dengue.



Uno de los principales desafíos encontrados durante el desarrollo del modelo fue el alto desbalance en la variable objetivo, donde los casos de dengue grave representaban una proporción mínima frente a los casos de dengue sin signos de alarma. Esta situación fue algo crítica, ya que un modelo entrenado sin tratamiento adecuado tendría tendencia a ignorar la clase minoritaria, afectando así negativamente la capacidad de detección de casos graves, los cuales precisamente, son los de mayor relevancia clínica.

Durante la fase de modelado, se evaluaron diferentes algoritmos de clasificación, entre ellos Random Forest, Redes Neuronales Artificiales y Regresión Logística. Los resultados mostraron que, aunque modelos como Random Forest presentan un buen desempeño general, estos tendían a sacrificar la detección de casos graves en favor de una mayor precisión global.



En contraste, el modelo de Regresión Logística presentó un mejor desempeño en términos de recall para la clase de dengue grave (0.875), lo cual resulta fundamental en el contexto del proyecto, ya que permite minimizar los falsos negativos, es decir, reducir la probabilidad de no detectar pacientes con riesgo elevado.

Este resultado refleja una decisión clave en la discusión del modelo: priorizar métricas orientadas a la sensibilidad sobre métricas globales como la precisión, dado el impacto clínico que tiene la correcta identificación de casos de dengue grave. En este sentido, el modelo seleccionado se alinea con un enfoque preventivo en salud pública.

Finalmente, los resultados obtenidos validan la hipótesis de que el uso de modelos de inteligencia artificial puede mejorar significativamente la clasificación de los tipos de dengue, permitiendo un análisis más rápido, preciso y escalable en comparación con los métodos tradicionales.

3.3 CONCLUSIONES

El estudio de los datos provenientes del Sistema Nacional de Vigilancia en Salud Pública (SIVIGILA) permitió consolidar un conjunto de datos depurado y homogéneo, compuesto por más de 36,000 registros relevantes, el cual permitió cumplir con el objetivo general de diseñar un modelo de inteligencia artificial basado en la metodología CRISP-DM para la clasificación automática de los tipos de dengue utilizando datos epidemiológicos del sistema SIVIGILA.

- ✓ Se destaca la importancia de aplicar técnicas ETL (Extract, Transform, Load), que no solo facilitaron la preparación de los datos para los modelos de machine learning, sino que también revelaron la necesidad de mejorar los sistemas de recolección de datos epidemiológicos en Colombia, especialmente en zonas endémicas como Valledupar.

- ✓ Al momento de analizar los patrones se identificó tendencias significativas en la incidencia del dengue. Los resultados evidenciaron que los estratos socioeconómicos bajos, particularmente el estrato 1, concentran el 61.81% de los casos, lo que subraya la relación entre condiciones de vulnerabilidad social y la enfermedad.

- ✓ Los datos indicaron que el grupo etario más afectado es el de 1 a 20 años, reflejando la alta susceptibilidad de los jóvenes ante esta enfermedad. En términos geográficos, Valledupar emergió como el municipio con mayor cantidad de casos, alcanzando un 30.95% del total departamental.

- ✓ Se identificaron indicadores clínicos claves como fiebre, dolor abdominal y vómitos persistentes, los cuales son determinantes para diferenciar casos graves de los no graves. Este conocimiento ofrece una base sólida para diseñar intervenciones de salud pública más focalizadas y efectivas.

- ✓ La comparación entre modelos confirmó la superioridad de los enfoques de machine learning frente a los métodos estadísticos tradicionales, resultando altamente efectivos para la clasificación de los tipos de dengue. Entre los modelos evaluados, la regresión logística demostró ser la más adecuada para el contexto del proyecto, debido a su capacidad para

detectar casos de dengue grave con un alto nivel de recall, lo cual es fundamental para la toma de decisiones clínicas.

- ✓ La implementación de un sistema de apoyo clínico basado en el modelo desarrollado demuestra la aplicabilidad práctica del proyecto, demuestra la forma de integrar la inteligencia artificial en entornos reales de salud pública.

En términos generales, este proyecto confirma que la integración de la inteligencia artificial en los sistemas de vigilancia epidemiológica puede contribuir significativamente a la detección temprana, clasificación y prevención de enfermedades como el dengue.

3.4 RECOMENDACIONES

- ✓ Fortalecer los procesos de recolección y estructuración de datos en el Sistema Nacional de Vigilancia en Salud Pública (SIVIGILA), garantizando mayor organización, consistencia y estandarización de la información epidemiológica, especialmente en zonas endémicas como Valledupar.

- ✓ Implementar mecanismos de validación y control de calidad de datos desde su origen, con el fin de reducir inconsistencias, valores faltantes y errores de digitación que afectan el desempeño de los modelos de inteligencia artificial.

- ✓ Ampliar la cobertura de los datos utilizados, integrando información de otras regiones del país, lo que permitiría mejorar la capacidad de generalización del modelo y su aplicabilidad a nivel nacional.

- ✓ Incorporar variables adicionales de tipo ambiental y climático (temperatura, humedad, precipitaciones), debido a su relación directa con la proliferación del vector *Aedes aegypti* y su impacto en la dinámica del dengue.

- ✓ Explorar técnicas más avanzadas de machine learning y deep learning, así como métodos de ensamble (boosting, stacking), que puedan incrementar la precisión y robustez del modelo desarrollado.

- ✓ Promover la adopción de herramientas basadas en inteligencia artificial dentro del sistema de salud pública, facilitando la detección temprana de patrones epidemiológicos y mejorando la respuesta ante posibles brotes de dengue.

- ✓ Utilizar este proyecto como base para futuras investigaciones en el campo de la epidemiología, fomentando el desarrollo de soluciones tecnológicas que contribuyan a la prevención, diagnóstico y control de enfermedades transmisibles.

3.5 BIBLIOGRAFIA

REFERENCIAS

A. Publicaciones periódicas

Alcalá, L., Quintero, J., González-Uribe, C., & Brochero, H. (2015). Productividad de Aedes aegypti (L.) (Diptera: Culicidae) en viviendas y espacios públicos en una ciudad endémica para dengue en Colombia. *Biomédica*, *35*(2), 258–268. https://doi.org/10.7705/BIOMEDICA.V35I2.2567

Alvis-Guzmán, N., Zakzuk-Sierra, J., Vargas-Moranth, R., Alcocer-Olaciregui, A., & Parra-Padilla, D. (2017). Dengue, Chikunguña y Zika en Colombia 2015-2016. *Revista MVZ Córdoba, ISSN 0122-0268, ISSN-e 1909-0544, Vol. 22, Nº. Extra 0, 2017*, *22*(0), 2. https://doi.org/10.21897/rmvz.1069

Lugo-Reyes, S. O., Maldonado-Colín, G., & Murata, C. (2014). Inteligencia artificial para asistir el diagnóstico clínico en medicina. *Revista Alergia México*, *61*(2), 110–120. https://doi.org/10.29262/ram.v61i2.33

Rodríguez-Velásquez, J. O., Prieto-Bohórquez, S. E., Pérez-Díaz, C. E., Pardo-Oviedo, J. M., Correa-Herrera, S. C., Mendoza-Beltrán, F. del C., Bravo-Ojeda, J. S., Morales-Pertuz, C. A., Rojas-Avila, N. A., & Flórez-Cárdenas, M. (2018). Predicción espacio-temporal probabilista de la epidemia de dengue total y grave en Colombia. *Revista de Salud Pública*, *20*(3), 354–358. https://doi.org/10.15446/rsap.v20n3.42701

Castrillón, J. C., Castaño, J. C., & Urcuqui, S. (2015). Dengue en Colombia: diez años de evolución. *Revista Chilena de Infectología*, *32*(2), 142–149. https://doi.org/10.4067/S0716-10182015000300002

Dávila Hernández, F., & Sanchez Corales, Y. (2012). Técnicas de minería de datos

aplicadas al diagnóstico de entidades clínicas Data mining techniques aplied to diagnosys of clinical entities. *Revista Cubana de Informática Médica*, *2012*(2), 174–183. http://scielo.sld.cu

Bedoya, O., Guarín, H., & Agudelo, J. (2023). Aplicación de técnicas de inteligencia artificial para la detección de tuberculosis pulmonar en Colombia. Revista EIA, 20(39), 1–23. https://doi.org/10.24050/reia.v20i39.1617

Belloso, W., et al. (2023). Inteligencia artificial en enfermedades infecciosas. ASEI - Actualizaciones en Sida e Infectología, 31(112), 77–90. https://doi.org/10.52226/revista.v31i112.209

Anaya, M., & Rodríguez, C. (2022). ABC de la inteligencia artificial (IA) aplicada en la salud. Medicina, 43(4), 493–496. https://doi.org/10.56050/01205498.1639

Conrad, B., et al. (2024). Utilization of machine learning for dengue case screening. BMC Public Health, 24(1573). https://doi.org/10.1186/s12889-024-19083-8

García, C., & Ramírez, E. (2021). Developing a dengue forecast model using machine learning techniques. PLOS Computational Biology. https://doi.org/10.1371/journal.pcbi.1009337

Gayathri, V. (2022). Data-driven methods for dengue prediction and surveillance using real-world and Big Data: A systematic review. PLOS Neglected Tropical Diseases. https://doi.org/10.1371/journal.pntd.0009646

Hoyos, W., Hoyos, K., & Ruiz, R. (2023). Modelo de inteligencia artificial para la detección temprana de diabetes. Biomédica, 43(3), 110–125.

López, J., & García, D. (2020). Utilization of machine learning for dengue case screening. BMC Public Health, 20(1), 1–10. https://doi.org/10.1186/s12889-020-08653-w

López, J., et al. (2020). Revisión crítica sobre la identificación de covid-19 a partir de imágenes de rayos x de tórax usando técnicas de inteligencia artificial. Revista Cubana de Transformación Digital, 1(3), 67–99. https://rctd.uic.cu/rctd/article/view/103/29

Márquez, J. (2020). Inteligencia artificial y Big Data como soluciones frente a la COVID-19. Revista de Bioética y Derecho, (50), 315–331. https://scielo.isciii.es/pdf/bioetica/n50/1886-5887-bioetica-50-00315.pdf

Messina, J. P., et al. (2021). Dengue models based on machine learning techniques: A systematic review. Journal of the Royal Society Interface, 18(184). https://doi.org/10.1098/rsif.2020.0897

Quiroz, M., U'Ren, A., & León, R. (2022). Inteligencia artificial habilita la identificación de pacientes COVID-19 en riesgo. Boletín de la Sociedad Mexicana de Física, 36(2), 91–100. https://dialnet.unirioja.es/servlet/articulo?codigo=8510779

Ramírez, F., & Martínez, P. (2020). Machine learning and dengue forecasting: Comparing random forests and artificial neural networks for predicting dengue burden. PLOS Neglected Tropical Diseases. https://doi.org/10.1371/journal.pntd.0008995

Rocklöv, J., & Martínez, P. (2022). A reproducible ensemble machine learning approach to forecast dengue outbreaks. Scientific Reports, 12(1). https://doi.org/10.1038/s41598-022-05394-9

Santos, M., & de Lima, R. (2021). Machine learning algorithms for dengue risk assessment: A case study. Computational and Applied Mathematics, 40(1), 1–20. https://doi.org/10.1007/s40314-021-01467-w

Singh, K., et al. (2024). Dengue Fever Outbreak Prediction Using Machine Learning Models: A Comparative Study. Data Science and Applications, 819, 443–455. https://doi.org/10.1007/978-981-99-7820-5_36

Swanson, E., et al. (2020). Predictors of dengue incidence using machine learning algorithms. Journal of Epidemiology and Global Health, 11(2), 130–145. https://doi.org/10.2991/jegh.k.200915.001

Ruiz, R., & Velásquez, J. (2023). Inteligencia artificial al servicio de la salud del futuro. Revista Médica Clínica Las Condes, 34(1), 84–91. https://doi.org/10.1016/j.rmclc.2022.12.001

Buñay Mendez, B. F., & Chango Sailema, W. G. (2024). Predicción Temprana del Dengue mediante Inteligencia Artificial: Un Enfoque basado en Análisis de Química Sanguínea Histórica. Estudios Y Perspectivas Revista Científica Y Académica, 4(3), 2923–2936. https://doi.org/10.61384/r.c.a.v4i3.590

B. Reportes

Instituto Nacional de Salud. (s.f.). Lineamientos para la atención de casos de dengue en Colombia. https://www.ins.gov.co/buscador-eventos/Lineamientos/Pro_Dengue.pdf

Congreso de Colombia. (2012). Ley 1581 de 2012 – Ley de protección de datos personales. https://esdegue.edu.co/sites/default/files/Normatividad/LEY%20TRATAMIENTO%20DE%20DATOS%20-%20LEY%201581%20DE%202012.pdf

C. Tesis de Magister o Disertación Doctoral

Camargo España, G. F. (2012). *Modelamiento de la dinámica del dengue en Colombia*. https://repositorio.unal.edu.co/handle/unal/12128

Cardona Gallego, J. A. (2018). *Influencia de variables sociales, económicas y espaciales en enfermedades transmitidas por vectores usando algoritmos y técnicas de Machine Learning*. Pereira : Universidad Tecnológica de Pereira. https://hdl.handle.net/11059/9237

Sánchez Rojas, J. D. (2023). *Desarrollo de un modelo de Machine Learning para la clasificación de tipos de dengue de acuerdo a su nivel de severidad: Un estudio de caso de Bucaramanga, Colombia*. https://hdl.handle.net/20.500.12495/10797

Ortega, J. (2022). *Detección y clasificación de enfermedades respiratorias mediante sonido pulmonar aplicando inteligencia artificial* [Tesis de pregrado o maestría]. Universidad de Pamplona. http://repositoriodspace.unipamplona.edu.co/jspui/handle/20.500.12744/5516

D. Artículos presentados en conferencias

Zea, D., & Osorio, L. (2011). The status of the dengue surveillance system in a Colombian municipality.

E. Manuales

Lujan, F., Pytel, P., & Pollo Cattaneo, M. F. (2014). *Metodología para la aplicación de procesos de explotación de información en establecimientos de salud*. http://sedici.unlp.edu.ar/handle/10915/42661

F. De internet

https://www.ins.gov.co/buscador-eventos/Lineamientos/Pro_Dengue.pdf

Regresión lineal - Wikipedia, la enciclopedia libre. (n.d.). Retrieved July 14, 2025, from https://es.wikipedia.org/wiki/Regresi%C3%B3n_lineal

Red neuronal artificial - Wikipedia, la enciclopedia libre. (n.d.). Retrieved July 14, 2025, from https://es.wikipedia.org/wiki/Red_neuronal_artificial

Random forest - Wikipedia, la enciclopedia libre. (n.d.). Retrieved July 14, 2025, from https://es.wikipedia.org/wiki/Random_forest

Roster, K., & Rodriguez, F. (2021). Neural Networks for Dengue Prediction: A Systematic Review. arXiv, 1, 16. https://doi.org/10.48550/arXiv.2106.12905
(Repositorio de preprints, sin revisión por pares formal)