MODELO DE APRENDIZAJE AUTOMÁTICO PARA LA PREDICCIÓN DEL COMPORTAMIENTO EPIDEMIOLÓGICO DEL DENGUE EN UN HOSPITAL PEDIATRICO DE CARTAGENA DE INDIAS.

Joel Joel Doria Atencia
Código 90.157.63

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

Director(a)
Diego Linares Ospina

Codirector(a)
Gloria Álvarez Vargas

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, ENERO 1 DE 2025

Tabla de contenido

INTRODUCCIÓN 1

1 DEFINICIÓN DEL PROBLEMA 2

1.1 PLANTEAMIENTO DEL PROBLEMA 2

1.2 FORMULACIÓN DEL PROBLEMA 4

2 OBJETIVOS DEL PROYECTO 5

2.1 OBJETIVO GENERAL 5

2.2 OBJETIVOS ESPECÍFICOS 5

3 MARCO TEÓRICO Y ANTECEDENTES 6

3.1 MARCO TEÓRICO 6

3.1.1 Dengue 6
3.1.2 Definiciones epidemiológicas 7
3.1.3 Modelos Predictivos 7
3.1.4 Modelos de aprendizaje automático 7
3.1.5 Series de Tiempo 8
3.1.6 Redes neuronales artificiales 8
3.1.7 Arboles de Decisión 8
3.1.8 Random Forest 9
3.1.9 Support Vector Machine 9
3.1.10 Regresión Lineal 9
3.1.11 Modelo ingenuo 9
3.1.12 Error absoluto medio 9
3.1.13 Error cuadrático medio 10
3.1.14 Raíz cuadrada del error cuadrático medio 10
3.1.15 R^2 10

3.2 ANTECEDENTES 10

4 ESTRATEGIAS DE PREPARACIÓN DE DATOS 13

5 ENTRENAMIENTO Y SELECCIÓN DEL MODELO PREDICTIVO 16

5.1 PROCEDIMIENTO DE ENTRENAMIENTO Y ESTRATEGIAS DE SELECCIÓN DEL MODELO 16

5.1.1 Modelos base 16
5.1.2 Perceptrón multicapa 17
5.1.3 Random Forest 17
5.1.4 Support vector machine 17
5.1.5 Conteo de aciertos exactos 18
5.1.6 Conteo de tendencias 19

5.2 RESULTADOS DEL MODELO 20

6 INTERPRETACIÓN DE RESULTADOS Y APLICACIONES PRÁCTICAS 26

7 CONCLUSIONES Y TRABAJOS FUTUROS 28

7.1 CONCLUSIONES 28

7.2 TRABAJOS FUTUROS 29
8 REFERENCIAS BIBLIOGRÁFICAS 30
9 Anexos 36

LISTA DE FIGURAS

Figura 1. Tendencia temporal de los tres subtipos clínicos de dengue 14
Figura 2. Tendencias de cada subtipo clínico de dengue. A. Dengue sin signos de alarma, B. Dengue con signos de alarma, C. Dengue grave. 15
Figura 3: Predicciones de la regresión lineal. A: Dengue sin signos de alarma, B: Dengue con signos de alarma, C: Dengue grave 21
Figura 4. Comparativa entre las predicciones entre cada modelo optimizado vs modelo ingenuo 24
Figura 5. Perceptrón multicapa optimizado. A dengue sin signos de alarma, B dengue con signos de alarma, C dengue grave 27

LISTA DE TABLAS

Tabla 1. Hiperparámetros de los modelos base 17
Tabla 2. Búsqueda de hiperparámetros 18
Tabla 3: Métricas de evaluación para modelo de regresión lineal 20
Tabla 4: Métricas de evaluación de los modelos base 21
Tabla 5. Hiperparámetros óptimos 22
Tabla 6. Desempeño global de todos los algoritmos 22
Tabla 7. Predicción exacta y de tendencias 25

LISTA DE ANEXOS

Anexo 1. Prototipo para la utilización del modelo 36
Anexo 2. Graficas comparativas entre el valor predicho vs el valor final. Modelo Random forest optimizado 36
Anexo 3. Graficas comparativas entre el valor predicho vs el valor final. Modelo SVM optimizado 36
Anexo 4. Permiso de la institución para realizar el estudio 37

INTRODUCCIÓN

El dengue en una infección viral transmitida por vectores, causada por un Arbovirus perteneciente a la familia Flaviviridae, con forma de icosaedro de ARN en sentido positivo con siete proteínas no estructurales y tres estructurales, las proteínas no estructurales son las encargadas de la replicación viral y por consiguiente de la patogenia de la enfermedad. El periodo de incubación dura entre tres y catorce días, luego de los cuales comienzan los síntomas, normalmente durante la primera infección el individuo es asintomático, o la enfermedad es leve y autolimitada [1].

Según su sintomatología los pacientes se clasifican en dengue sin signos de alarma, dengue con signos de alarma o dengue grave [1], [2]. La incidencia de esta enfermedad ha ido en aumento, siendo el año con mayor número de casos reportados 2023 con aproximadamente 6.5 millones de casos y 7.300 muertes en más de ochenta países, se estima que 3.9 billones de personas se encuentran en riesgo de presentar infección por Dengue [3]. En Colombia se reportaron 343 casos por cada 100.000 habitantes, y ocupamos el quinto puesto entre los países con mayor letalidad por Dengue [3].

El virus del dengue supone diferentes retos para los profesionales de la salud, el primero y más importante es el colapso de las instituciones, por la alta incidencia los centros de atención ante brotes se colapsan o se declaran en emergencia, además que la atención supone costos por paciente atendido que influye en la carga al sistema de salud. Teniendo en cuenta los costos que supone un caso de dengue y la gravedad de una muerte o complicación, sería pertinente predecir la incidencia de la enfermedad basándose en los registros existentes con el fin de idear y mejorar políticas que se anticipen a los picos de contagio y disminuya la carga a los entes en salud.

Este trabajo tiene como objetivo desarrollar un algoritmo que permita la predicción del comportamiento epidemiológico del dengue en un hospital pediátrico de la ciudad de Cartagena.

Para esto los modelos de aprendizaje automático seleccionados fueron: regresión lineal, redes neuronales artificiales,“Random forest” (RF) y “support vector machine” (SVM).

La regresión lineal no pudo comprender por completo las dinámicas de la enfermedad, al usar las otras arquitectura sin búsqueda activa de sus hiperparámetros encontramos resultados similares a la regresión lineal pero con ligeras mejorías. Posteriormente se realizó un ajuste de hiperparámetros utilizando *“gridsearch*” y *“earlystopping”*, por ultimo, se decidió crear un modelo ingenuo, en donde la predicción se basa simplemente en repetir el último valor.

En nuestro caso no existe una arquitectura universalmente superior a la hora de predecir los casos de dengue, el ajuste de hiperparametros no mejoró significativamente el desempeño de ningún modelo. La red neuronal fue la que mejores resultados presentó, sin embargo, sus predicciones no deben interpretarse de forma aislada, ya que la enfermedad tiene un patrón errático y ruidoso, para intentar predecir su incidencia es necesario incluir otras variables predictoras, aumentar la muestra y explorar otras metodologias de analisis predictivo y estadistico.

1 DEFINICIÓN DEL PROBLEMA

1.1 PLANTEAMIENTO DEL PROBLEMA

El dengue en una infección viral transmitida por vectores, causada por el virus del dengue (DENV) [4]. Se trata de una enfermedad altamente prevalente en países tropicales como Colombia, es trasmitida por los mosquitos del género Aedes y causada por un virus de la familia Flaviviridae con cuatro serotipos: DENV-1, DENV-2, DENV-3, DENV-4, la infección por un serotipo solo proporciona inmunidad homotipica, es decir contra el serotipo que causó la infección, sin embargo, no existe inmunidad cruzada entre serotipos [4], [5]. Ante su alta prevalencia al nivel mundial y su potencial letalidad se considera como un evento de interés en salud pública [6].

No todos los pacientes presentarán dengue grave con shock, existen tres fases en su historia natural, una fase febril la cual presenta síntomas prodrómicos de una infección viral, con fiebre alta y malestar general que puede o no estar acompañado de exantema, seguida por una fase crítica normalmente al quinto día de fiebre que se caracteriza por un cese de la fiebre, esta fase se denomina critica porque aquí el paciente puede tomar dos caminos, o ingresa a la fase de recuperación, con desaparición de cuadro clínico y normalización de paraclínicos, o empeora considerablemente hasta el desarrollo de dengue con signos de alarma o dengue grave [7], [8]. Lo que dificulta la fisiopatología del dengue, son los mecanismos del por qué la enfermedad grave no ocurre en todos los pacientes, que aún son controvertidos y no existe una manera de predecir que paciente presentará síntomas graves y cuáles no.

En los últimos años la incidencia de esta enfermedad ha ido en aumento, siendo el año con mayor número de casos reportados 2023 con aproximadamente 6.5 millones de casos y 7.300 muertes en más de 80 países [3].

La infección por el virus del Dengue produce un aumento en la mortalidad así como en la morbilidad de la población, además que supone costos para el sistema de salud, con un promedio de $88 dólares estadounidenses por paciente ambulatorio y entre $671 y $ 6.532 dólares estadounidenses por cada paciente intrahospitalario [9], además, un aumento en la tasa de mortalidad por dengue significa problemas en la atención oportuna de los pacientes ya que estos eventos son evitables en un 98% de los casos [10], por lo que, si bien la mortalidad por dengue ha disminuido, continúa siendo de vital importancia el monitoreo para la prevención o el tratamiento adecuado.

El panorama epidemiológico en Colombia también es preocupante, se han reportado en promedio 343 casos por cada 100.000 habitantes, y ocupamos el quinto puesto entre los países con mayor letalidad por Dengue, siendo el grupo etario más vulnerable los jóvenes entre 15-19 años [11], [12]. El comportamiento de los brotes por Dengue es fluctuante, con aumento en los contagios cada 3 años, lo cual concuerda con los informes epidemiológicos de 2019, 2023

y 2024, es importante destacar que en los últimos dos años los casos han superado los periodos anteriores, no obstante, han sido los años con mortalidad más baja por Dengue, lo cual puede deberse a un aumento en las estrategias de vigilancia epidemiológica y manejo oportuno de los centros de atención en salud [13], demostrando la importancia que tiene la vigilancia en el manejo de las enfermedades transmitidas por vectores.

Es importante recalcar la importancia del correcto uso de la información del epidemiológica del dengue, al considerarse un evento de interés en salud pública todo caso probable o confirmado por dengue debe notificarse al Sistema Nacional de Vigilancia en Salud Pública (SIVIGILA), que junto al instituto nacional de salud (INS) tienen fichas de notificación obligatoria que cada institución debe diligenciar para luego ser enviadas a su respectivo ente territorial, las fichas aparte de indicar que pacientes tienen o no dengue, son un registro útil de la procedencia, el lugar del caso e incluso los signos y síntomas clínicos que manifiesta [2].

La información normalmente se utiliza para los reportes del instituto nacional de salud (INS), así como del SIVIGILA, con el fin de vigilar la enfermedad [10], [11], al ser el Dengue una patología altamente prevalente las notificaciones son numerosas, por medio de la ciencia de datos se pueden generar algoritmos que aprovechen mejor la información, no solamente para visualizar el comportamiento actual de los eventos en salud, si no, además, para identificar posibles patrones de las enfermedades, con el objetivo de diseñar estrategias de promoción de la salud y prevención de la enfermedad, así como medidas de contingencia para preparar a los centros de atención ante un posible brote.

Actualmente, aunque existen estudios que aplican modelos de aprendizaje automático para la identificación temprana de signos de alarma, en Colombia, pocos se centran en la epidemiología del dengue. El aprovechamiento de la gran cantidad de datos generados por los centros de sanitarios y que son reportados al SIVILIGA e INS a través de las fichas de notificación junto con la integración de la ciencia de datos para la vigilancia del evento, puede transformar la manera en que se monitorea y controla esta enfermedad y de tal modo contribuir a la optimización de la salud pública del país.

1.2 FORMULACIÓN DEL PROBLEMA

Desde el contexto clínico el virus del dengue supone diferentes retos: primero, nos encontramos en una zona endémica, en dónde se encuentra el vector responsable de la transmisión, que si bien previamente se asociaba solo a zonas que no superaran los 2.200 metros sobre el nivel del mar, estudios recientes han demostrado la presencia de mosquitos adultos en lugares con hasta 2.550 metros sobre el nivel del mar [14], partiendo de la premisa de la presencia del vector en la mayor parte del territorio, el segundo reto es la prevención de la mortalidad, el ingreso de un paciente con síntomas sospechosos o con la enfermedad confirmada se maneja de forma rigurosa por los centros de salud, teniendo en cuenta además que por diversos motivos ya sea nivel socioeconómico, origen de zona de difícil acceso, entre otros, los pacientes suelen clasificarse como alto riesgo social, lo que intensifica y justifica las estrategias de control y vigilancia en salud [15], el ultimo reto es el colapso de los centros de salud ante el flujo tan alto de pacientes que manejan, si bien el ministerio de salud establece estado de alerta epidemiológica, dicha alerta se activa teniendo en cuenta los casos reportados hasta el momento y las medidas de contingencia son implementadas luego de la activación del estado de alerta [16], por lo que teniendo en cuenta los costos que supone un caso de dengue y la gravedad de una muerte o complicación, sería pertinente predecir el comportamiento de la enfermedad basándose en los registros existentes con el fin de idear y mejorar políticas que se anticipen a los picos de contagio y disminuya la carga a los entes en salud.

Pregunta de Investigación:
¿Cómo predecir el comportamiento epidemiológico de la infección por virus del Dengue en un hospital pediátrico de la ciudad de Cartagena?

Preguntas complementarias:
- ¿Cuál es la forma de preparación de datos epidemiológicos para la creación de un modelo de aprendizaje automático?
- ¿Cuáles son las características sociodemográficas y clínicas de los pacientes?
- ¿Qué modelo automático para aplicar en esta metodología?
- ¿Cómo se pueden evaluar los resultados del modelo de aprendizaje automático?
- ¿Cómo se muestran los resultados del modelo?

2 OBJETIVOS DEL PROYECTO

2.1 OBJETIVO GENERAL

Desarrollar un algoritmo que permita la predicción del comportamiento epidemiológico del dengue en un hospital pediátrico de la ciudad de Cartagena

2.2 OBJETIVOS ESPECÍFICOS

• Determinar los atributos que deben tener el modelo predictivo.
• Describir las características sociodemográficas y clínicas de los pacientes.
• Entrenar los modelos para el análisis predictivo con los casos recolectados para identificar patrones en los datos y generar predicciones apropiadas.
• Evaluar los resultados de los modelos en términos de métricas de regresión para seleccionar el de mejores resultados.
• Desarrollar un prototipo para presentar los resultados.

3 MARCO TEÓRICO Y ANTECEDENTES

3.1 MARCO TEÓRICO

3.1.1 **Dengue**

El dengue es una enfermedad viral, transmitida por vectores, principalmente mosquitos de la especie Aedes aegypti, es hiperendemica en zonas tropicales y subtropicales, como Colombia, además, en los últimos años debido al aumento en el registro de casos se considera aparte de un evento de interés en salud pública, una de las infecciones transmitidas por mosquitos de más rápido crecimiento a nivel mundial [17], [18].

La infección es causada por un Arbovirus perteneciente a la familia Flaviviridae, es un virus con forma de icosaedro de ARN en sentido positivo que tiene siete proteínas no estructurales y 3 estructurales, que en conjunto se encargan de formar la estructura del virus y su posterior sintomatología [17]. Existen cuatro serotipos infectantes en humanos del dengue, sin inmunidad cruzada, es decir que la infección por una cepa del virus solo protegerá al individuo de una reinfección contra el mismo serotipo, más no de la infección por otro [1]. Al producir anticuerpos contra el serotipo inicial del virus y generarse infección por un nuevo serotipo, los anticuerpos existentes no logran neutralizarlo, convirtiéndose en anticuerpos facilitadores, que funcionan como una vía de unión entre la célula inmune permitiendo la entrada del virus y por consiguiente su ciclo vital [1]. Para que ocurra la infección es necesario un mosquito hembra que pique a alguien que cuente con la enfermedad durante el periodo de viremia (después de 7 días de incubación), dicho mosquito debe picar a otra persona para transmitirle la enfermedad. El periodo de incubación dura entre 3-14 días, luego de los cuales comienzan los síntomas, normalmente durante la primera infección el individuo es asintomático, o la enfermedad es leve y autolimitada. [1], [2].

Con el dengue sin signos de alarma la sintomatología es inespecífica, se caracteriza por fiebre de inicio abrupto y malestar general, otros síntomas incluyen dolor articular, dolor retro orbital, y la aparición de un exantema generalizado o rubor facial. Este periodo tiende a durar entre 3 a 5 días para dar paso a la fase crítica de la enfermedad, en dónde puede ocurrir o resolución del cuadro clínico o aparición de signos de alarma con deterioro. Los signos de alarma son diversos, los más comunes son la aparición de vómitos, dolor abdominal continuo e intenso, sangrado por mucosas, visceromegalias y hemoconcentración con aumento del hematocrito. Por último, el dengue grave se define por la presencia de: hemorragia masiva, shock por extravasación plasmática, disfunción hepática con elevación de enzimas hepáticas >1.000 U/L y alteración de sistema nervioso central o de dos o más órganos [2], [19].

El diagnóstico es clínico y paraclínico, como nuestro territorio es zona endémica la sospecha existe por episodios febriles de varios días. Durante los 3 primeros días de fiebre se recomienda la prueba de detección de NS1 [17]. NS1 es una proteína no estructural del dengue, su resultado negativo no descarta diagnóstico, a los 5 días se elevan los niveles de anticuerpos IgM, por lo que a partir del quinto día de fiebre se solicita la serología, la IgG se eleva a partir del día 14 de la enfermedad, ya

para este momento se solicita una prueba de ELISA, la cual es un inmunoanálisis de absorción para detectar anticuerpos específicos IgM e IgG para dengue [1], [2], [16]. Al ser una enfermedad catalogada como caso de interés en salud pública a nivel mundial y en Colombia como un problema prioritario en salud pública, la notificación al sistema de vigilancia en salud de un caso probable, confirmado por laboratorio o por nexo epidemiológico, debe realizarse de manera inmediata para los reportes epidemiológicos nacionales [2].

3.1.2 Definiciones epidemiológicas

Brote: Episodios de varios casos de una misma enfermedad que guardan relación entre sí, ya sea por la distribución geográfica, el inicio de los síntomas o las características de los individuos afectados [20].

Epidemia: La aparición alta e inusual de brotes de una enfermedad en una población delimitada, para catalogarse como epidemia los casos nuevos deben superar los previstos. Existen las epidemias por contagio causadas por agentes infecciosos que causan la aparición de la enfermedad de manera relativamente lenta y las epidemias puntuales, en donde un grupo de individuos son expuestos de forma abrupta a un foco lesivo y la elevación en la incidencia de los casos se presenta en horas [20]. Además, según el tiempo de aparición podemos clasificarlas en explosivas si los casos aumentan y disminuyen drásticamente, en lentas si la incidencia se ve multiplicada en un periodo más largo de tiempo y las epidemias en aguja con cola que inician como una epidemia explosiva para continuar como una lenta. Si la epidemia se extiende por varias regiones geográficas hasta propagarse de forma mundial se denomina pandemia [20].

3.1.3 Modelos Predictivos

Los modelos de análisis predictivo son algoritmos que permiten predecir la ocurrencia de un evento teniendo en cuenta información ya existente, con el objetivo de tomar decisiones. Siempre es necesaria la presencia de un algoritmo matemático o estadístico para que realmente se considere un modelo predictivo, [21]. Cuando se realizan estos algoritmos se tiene como objetivo una de dos opciones, la primera es generar una predicción con el menor margen de error posible, la segunda es la comprensión de la relación entre variables más que la optimización del resultado de las mismas [22]. Los tipos de modelos más utilizados actualmente son los de clasificación, agrupación y series temporales. Los modelos de clasificación producen una estimación de una variable categórica, los de agrupación o clustering son juntan datos que comparten similaridades entre sí, aquí como tal no hay una variable que trate de ser predicha, el objetivo es agrupar la información en subgrupos homogéneos. Por ultimo las series temporales utilizan datos con marcadores temporales para visualizar y predecir el comportamiento a lo largo de un periodo de tiempo [22], [23], [24].

3.1.4 Modelos de aprendizaje automático

Son un conjunto de métodos que detectan patrones en información existente, los patrones

luego son utilizados para la predicción de información y toma de decisiones, son inherentemente modelos predictivos [23]. Ahora, se subdividen según sus características en dos grupos, están los modelos predictivos supervisados en dónde se entrena al modelo ingresado valores X para obtener un valor Y [23]. Los modelos descriptivos, también llamados modelos de aprendizaje no supervisado, por otra parte, intentan descubrir los patrones de la información, aquí no hay inputs como tal, sino que trabajamos con los outputs para conocer su comportamiento, con ellos no hay un error o parámetro que indique invalidez ya que buscan identificar conocimiento. Las aplicaciones en el mundo real son variadas y se usan principalmente para el reconocimiento de imágenes [23].

3.1.5 Series de Tiempo

Las series de tiempo son conjuntos de variables que se obtienen y se organizan por un periodo delimitado de tiempo, la información de observa en diferentes momentos para luego por medio de procedimientos estadísticos realizar inferencias. Se asume que todas las muestras son independientes y que tienen la misma distribución. El análisis puede hacerse con una sola serie de tiempo, o podemos analizar varias a la vez, de tal modo podemos observar una o más variables y a grandes rasgos evaluar si existe o no relación entre ambas, eso sí, para determinar relación entre dos o más variables se necesitan más estudios estadísticos [24], [25], [26].

Las series de tiempo pueden ser estacionarias cuando el comportamiento probabilístico de cada colección de valores como la media, la varianza y la autocorrelación, son constantes a lo largo del tiempo, es decir, cada valor debe ser igual a su contraparte representada con el cambio en el tiempo, si no se cumple este enunciado, la serie de tiempo no será estacionaria [26].

3.1.6 Redes neuronales artificiales

Se definen como la suma de muchas unidades simples de procesamiento de información, estos algoritmos tratan de imitar el funcionamiento del cerebro humano. Aquí se enfatiza en los elementos de procesamiento que interpretan datos generalmente no lineales interconectándose a otros elementos de procesamiento, e incluso a ellos mismos, la forma en la que la información se reparte y procesa debe ser definida, los parámetros ajustables que condicionan a las redes neuronales se denominan “Weights” , una vez que se termina el proceso de interpretación de la información de cada una de las unidades de procesamiento, sus resultados se suman para generar el “output” que bien puede tomarse como un resultado en concreto o puede pasarse nuevamente por la red neuronal [27], [28], [29], [30], [31].

3.1.7 Arboles de Decisión

Son modelos de aprendizaje supervisado, se denominan arboles de decisión por su estructura de un árbol invertido, tienen ramificaciones (llamadas esquinas) y vértices (llamados nodos). Parten desde un nodo raíz, cada nodo puede tener una o más esquinas, si un nodo tiene esquinas (ramas) saliendo de él se llama nodo de decisión, si no las tiene recibe el nombre de hoja. Como cada nodo puede tener cero o dos ramificaciones también se les denomina árbol binario, la profundidad de un árbol de decisión corresponde al número de niveles que se encuentren debajo del nodo raíz y entre

el nodo hoja. Cada nodo de decisión es una pregunta, la cual debe ser debidamente escogida, las métricas para correcta separación de las ramas el árbol dependerán del objetivo final del algoritmo, si se trata de un modelo de clasificación utilizaremos "accuracy", "Gini index" o "Entropy", mientras que si buscamos realizar una regresión se utiliza el error cuadrático medio (MSE) [32].

3.1.8 Random Forest

Los modelos de random forest son el siguiente paso para la construcción de un algoritmo robusto, acá se cuenta con múltiples arboles de decisión que realizan predicciones individualmente, la predicción final es el conjunto de cada predicción individual, esto se llama "esemble learning", y funciona de la misma forma, las predicciones en los árboles ocurren por que el algoritmo busca el mejor resultado con cada división de un nodo [33].

3.1.9 Support Vector Machine

Poderoso y versátil modelo de machine learning, ya que permite clasificación linear y no linear, regresión e incluso detección de novedad. Cuando se usa para clasificación el modelo busca hiperplanos óptimos para separar los datos lo mejor posible asegurando distancia entre grupos, estos hiperplanos o márgenes tienden a ser un poco flexibles para evitar confusión con "ouliers". En el caso de la regresión el algoritmo busca encajar los datos dentro de los hiperparámetros, para así predecir su tendencia, esta medida sin embargo, hace al modelo sensible al valor épsilon [34].

3.1.10 Regresión Lineal

Es un modelo simple pero robusto para realizar predicciones, su objetivo en un dataframe es realizar una línea que cruce de la forma más cercana por todas y cada una de las entradas de la base de datos [35]. Para realizar una regresión lineal nos basamos en una ecuación linear:

$$y = mx + b$$

"m" equivale a la pendiente de la línea, "*x"* corresponde a valor de cada atributo y "*b*" es el error o *"bias"*. En un modelo predictivo de regresión lineal se suma cada variable con su respectivo peso y el algoritmo ajusta estos parámetros múltiples veces hasta obtener la mejor línea que divida los datos uniformemente [35].

3.1.11 Modelo ingenuo

Los modelos ingenuos son definidos como la forma más básica de realizar una predicción, en ellos simplemente se toma el ultimo valor conocido como la predicción futura, si las predicciones del modelo ingenuo son muy cercanas a los valores reales, la serie de tiempo es auto correlacionada. Lo ideal es que los otros modelos de IA superen a la predicción ingenua [36].

3.1.12 Error absoluto medio

Promedio de las distancias absolutas entre las predicciones y los valores reales, indica cuanto se equivoca el modelo en promedio y se encuentra en las mismas unidades que la variable objetivo. Su valor ideal debe ser el menor posible [37].

3.1.13 Error cuadrático medio

Similar al error absoluto medio solo que las diferencias entre las predicciones y el valor real se encuentran al cuadrado para obtener siempre un valor positivo [37].

3.1.14 Raíz cuadrada del error cuadrático medio

La raíz cuadrada del error cuadrático medio, permite una interpretación en las unidades de la variable objetivo, pero es más sensible a valores extremos [37].

3.1.15 R^2

También llamado coeficiente de determinación, indica la proporción de la varianza de la variable dependiente que es explicada por la variable independiente, tiene un rango de valores entre cero y uno, mientras más cercano esté a uno quiere decir que los predictores logran explicar la variable objetivo [38].

3.2 ANTECEDENTES

La enfermedad por virus del dengue genera bastante impacto en nuestro medio y en el mundo, con el surgir de nuevas tecnologías cada día aumenta la cantidad de información a la que se puede tener acceso, el uso de sensores, otros dispositivos y el manejo de historias clínicas digitales permiten colocar datos valiosos al alcance de nuestros dedos, información que al analizarse nos ofrece nuevas soluciones a nuestras problemáticas. La ciencia de datos permite incluso predecir futuros brotes de una enfermedad, lo cual en la práctica clínica supone utilidad para la ideación de estrategias tanto para mitigarlos, cómo para reducir su impacto. Múltiples estudios enfocados en la aplicación de modelos de aprendizaje automático con el fin de predecir el comportamiento epidemiológico de una patología se han realizado con diversos resultados.

En cuanto al uso de modelos de predicción para Dengue, Asia lleva la delantera. En India, se ha abordado la problemática por medio de modelos híbridos con el modelo autorregresivo integrado de media móvil (ARIMA) añadiendo además procesamiento de “neural network” (NNAR) para los datos no lineares que se pudieran encontrar presentes en la base de datos. ARIMA se usa para el análisis inicial con datos lineales, para luego procesar los residuos de la regresión con NNAR, los autores concluyen que su modelo hibrido es útil para series de tiempo que trabajen con grandes cantidades de datos [39]. Otras metodologías también realizadas en India, incluyen la aplicación de un algoritmo “Random Forest” para un sistema de clasificación, monitoreo y estimación de riesgo que tiene una persona de padecer dengue, los resultados de este estudio además fueron comparados con otras técnicas de predicción, superándolas en especificidad y con menores valores de error estándar. Si bien en este articulo no utilizan el aprendizaje automático para identificar posibles picos de contagios, si los utilizan para llevar un control de las características clínicas de los pacientes, información que para los entes de control en salud también es muy útil a la hora de la creación de estrategias en salud pública [40].

China también ha desarrollado investigaciones en este campo. El artículo "The Diagnosis of Dengue in Patients Presenting With Acute Febrile Illness Supervised Machine Learning and Impact of Seasonality" describe la creación de un algoritmo para predecir el diagnóstico final de un paciente con un síndrome febril agudo, usando aprendizaje automático de potencia de gradiente, presentando un valor predictivo negativo mayor al 90%, lo cual quiere decir que pudo predecir que pacientes con síndrome febril, no tenían Dengue [41]. También del mismo país, otros autores mediante "Deep learning" con memoria a largo-corto plazo y redes neuronales predijeron la incidencia de la enfermedad en 20 ciudades, satisfactoriamente obtuvieron menor desviación cuadrática media que otros modelos de "Machine Learning" [42].

Otros países asiáticos que han publicado sobre el tema incluyen Indonesia, Taiwán, Tailandia y Singapur. Sus metodologías incluyen modelos de regresión linear para trabajar con la incidencia de la enfermedad correlacionando los brotes con variables climáticas. Las predicciones con este modelo fueron acertadas solo en presencia de un aumento en tiempo real de contagios y no en momentos con menos reportes de casos [43]. Regresión multivariada de Poisson para el análisis de la incidencia de dengue incluyendo, también, otros determinantes como la temperatura o las lluvias registradas, pronosticando brotes durante 2011 con un margen de error inferior al 3% y una proyección de hasta 16 semanas, solo con datos ambientales y reportes epidemiológicos pasados, siendo la temperatura el predictor más fuerte en su estudio [43].

Debido al avance que los entes de vigilancia de enfermedades infectocontagiosas, nuevos algoritmos más complejos se han creado con el paso del tiempo, como por ejemplo el modelo de operador de selección y contracción mínima absoluta (LASSO en inglés), el cual es utilizado para la predicción de brotes de diferentes enfermedades infecciosas y supone una mejoría del 20% en las predicciones ya que integra patrones medioambientales indispensables para un resultado más específico, sin embargo, no es efectivo si requiere ser usado a largo plazo [50]. La tendencia es utilizar modelos más convencionales como ARIMA o SARIMA , aunque no quiere decir que sean los únicos eficientes, los modelos aditivos (GAM), las redes neuronales artificiales (ANN), "Rain Forest", entre otros cumplen su objetivo correctamente, su implementación más bien depende del tipo de predictores que se quieran evaluar, principalmente si se desea incluir factores externos [44], [45], [46], [47], [48].

No solo la epidemiologia del Dengue ha sido objeto de estudio en Asia, por la alta prevalencia de la patología en el territorio, también se ha utilizado ciencia de datos para mejorar los protocolos de atención sanitaria con la creación de un algoritmo de estratificación del riesgo de dengue grave, el cual, obtuvo una sensibilidad de 0.66 y una especificidad de 0.84 con un valor predictivo negativo de 0.98, este último valor es el más prometedor ya que si el algoritmo clasifica a un paciente como de bajo riesgo, la probabilidad de que desarrolle dengue grave sería mínima, pudiendo mejorar la toma de decisiones en los pacientes y brindar un manejo ambulatorio que disminuiría los gastos al sistema de salud [49].

Otro ejemplo es “Assessing the risk of dengue severity using demographic information and laboratory test results with machine learning”, aquí también usan redes neuronales artificiales, sin embargo, su aplicación es para detectar predictores de riesgo, dando como resultado que los niveles de NS1 y la presencia de anticuerpos IgG e IgM simultáneamente aumentan el riesgo de sufrir dengue grave [50], en general Asia tiene muchos trabajos que evidencian el poder de los modelos predictivos [51], [52].

En América también existen trabajos publicados con resultados similares. Los autores de la referencia 60 utilizaron el modelo SARIMA, ajustado con Box-Jenkins, para crear tres diferentes métodos que permitieran predecir la actividad del virus, al año, a los tres meses y al mes para luego comparar las estimaciones con los datos reportados oficialmente. En este estudio las predicciones estadísticamente significativas fueron las realizadas para el año siguiente, sin embargo, el modelo fue ajustado encontrando valores más satisfactorios para las predicciones a partir de los tres meses [53]. SARIMA no siempre crea los cálculos más acertados, autores de la ciudad de Campinas en el estado de Sao Paulo, en Brasil recolectaron datos en un periodo desde 1998 hasta 2008, luego crearon un modelo para la predicción de los contagios durante 2009, y de forma similar contrastaron el resultado del modelo SARIMA con los casos reales que hubo en la zona. Concluyeron que si bien el modelo predijo con precisión los casos casos para 2009 no es preciso cuando se trata de meses o años en dónde aumenta la transmisión de la infección, ya que en su estudio observaron muchos más casos in vivo [54].

“Dengue forecasting in São Paulo city with generalized additive models, artificial neural networks and seasonal autoregressive integrated moving average models” Describe la creación de un modelo complejo, ya que mezclan diferentes metodologías de aprendizaje automatico, modelo de autorregresión integrado de media móvil por estaciones (SARIMA), modelos aditivos generalizados (GAM), redes neuronales artificiales (ANN) y modelos bayesianos. En este estudio el modelo con mejores resultados y predicciones fue GAM [55].

En general el país de Latinoamérica con más artículos publicados con este eje temático es Brasil, no obstante, México también tiene publicaciones relacionadas. Ahora, las metodologías presentan el mismo comportamiento que en Asia, los estudios concuerdan con que para evaluar solo factores de incidencia análisis autorregresivos o “Random Forest” pueden ser utilizados, sin embargo, los determinantes ambientales favorecen la especificidad del modelo por lo que no deben ser pasados por alto. Para incluir las variables ambientales se modifican los modelos de auto regresión (SARIMA) o directamente se usan redes neuronales [56], [57], [58], [59], [60], [61], [62].

Colombia no ha incursionado mucho en el uso de la inteligencia artificial para el estudio de enfermedades infecciosas, sin embargo, un par de artículos si se han realizado en el territorio, en colaboración con universidades alrededor del mundo. La Universidad Cooperativa de Colombia y SIVIGILA construyeron un modelo “rain Forest” para la predicción de casos de dengue en un contexto nacional y departamental, los resultados de su modelo luego fueron comparados con el clásico ARIMA y ANN superándolos. Sin embargo, las predicciones realizadas a escala nacional

tuvieron mayor precisión que aquellas limitadas a un departamento, probablemente por la mayor cantidad de datos a la que el modelo tenía acceso. Algo importante a resaltar de este estudio es la descripción de las variables predictoras respecto a la precisión del resultado, las variables sociodemográficas juegan un rol importante a nivel nacional, no tanto a nivel departamental, en dónde hubo mayor correlación con variables ambientales y climáticas [63].

Otra publicación en donde se utiliza el modelo ARIMA realizado en Colombia se llama "Intra- and Interseasonal Autoregressive Prediction of Dengue Outbreaks Using Local Weather and Regional Climate for a Tropical Environment in Colombia", aquí el algoritmo fue más útil al predecir la incidencia de 2 semanas a 6 meses [64]. En conjunto con la secretaría de salud de Santander y la universidad de Valencia se implementaron simulaciones "Markov Chain Monte Carlo" encontrando correlación positiva entre la temperatura de la ciudad y el aumento en los contagios tal como está descrito en otros países [65].

Si bien en el país no se ha implementado de forma autóctona el uso de "Machine Learning" con dengue, si se ha hecho con enfermedades infecciosas, el estudio "INFEKTA—An agent-based model for transmission of infectious diseases: The COVID19 case in Bogotá, Colombia" por la universidad nacional combina datos demográficos, mecanismo de transmisión y regiones muy transitadas de Bogotá para predecir el comportamiento y la transmisión de una patología, en este caso COVID-19. INFEKTA funciona creando un espacio simulado euclidiano que permite representar la transmisión dinámica de la COVID-19 teniendo en cuenta las rutinas de los habitantes, puntos de interés o aglomeraciones, restricciones en zonas específicas y medidas de aislamiento personal [66].

Actualmente no existen investigaciones realizadas en la región Caribe en dónde se aplique "Machine Learning" para la vigilancia, control u observación del virus del Dengue y en general las investigaciones llevadas a cabo en el país fueron realizadas a través de organizaciones externas, no nacionales, lo cual resulta preocupante teniendo en cuenta la alta prevalencia de la enfermedad en la zona.

4 ESTRATEGIAS DE PREPARACIÓN DE DATOS

Los datos utilizados en el presente trabajo pertenecen a la base de datos del area de epidemiología clinica de una institución de atención pediatrica de IV nivel de complejidad en salud, que además recoge pacientes pediatricos de toda la costa Caribe. La obtención de la base de datos fue realizada mediente la aprobación del director cientifico del hospital luego de evaluar el proyecto en el comité de etica institucional, los datos proporcionados pertenencen al periodo 2019 hasta 2022. De todas las variables para el análisis predictivo solo se seleccionaron la fecha de notificación del caso y la clasificación clinica final del paciente, ya que los modelos predictivos se entrenarían netamente con datos de incidencia, no se incluyeron otras variables para la formulación de las predicciones debido a problemas con la adquisición de la información, de tal modo que se trabajó con 3346 casos de dengue (las entradas del *"dataframe"*) divididos en 3 subtipos clinicos (número de atributos).

Para asegurar una correcta interpretación cronológica de los registros se convirtió la variable de la fecha en un objeto *"datetime"* con la librería pandas [67], las tres clasificaciones que se incluían en la base de datos fueron: dengue sin signos de alarma (DSSA), dengue con signos de alarma (SA) y dengue grave (DG). Tomando las clasificaciones y las fechas de notificación se creó un nuevo dataframe (DF) con el conteo semanal de cada subtipo de dengue y con las fechas de rango semanales y el numero de la semana correspondiente con el año, se tomó el valor de predicción semanal ya que el ente al que se presentan los resultados (SIVIGILA) trabaja con semanas epidemiológicas [68], por lo que cuando hablamos de incidencia de casos en medicina el estandar es semanal.

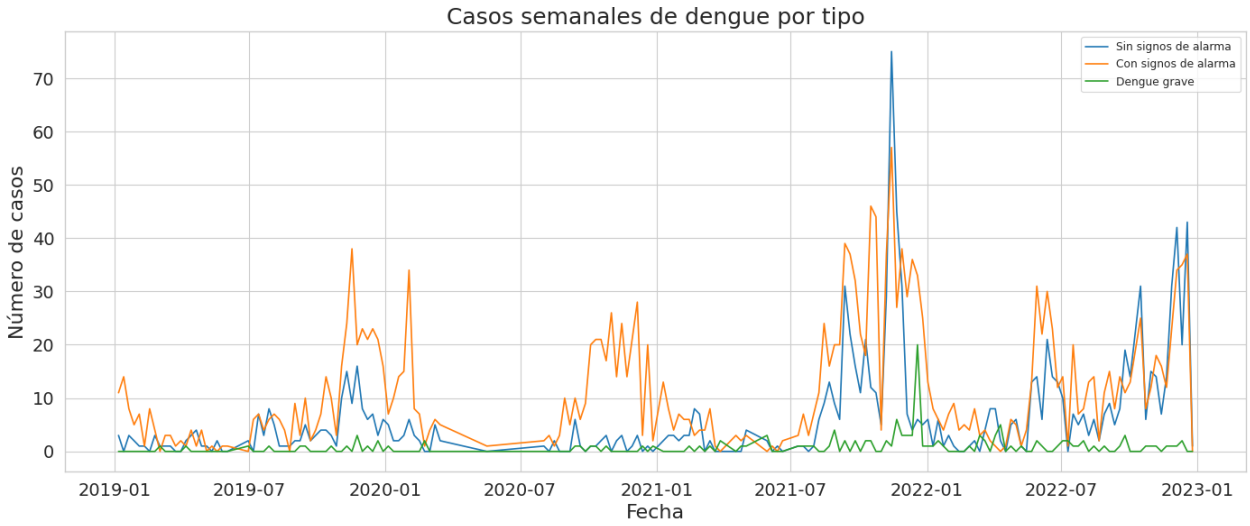


Figura 1. Tendencia temporal de los tres subtipos clínicos de dengue

Una vez construido el nuevo "*dataframe"* se realizó por medio de matplotlib una grafica de serie de tiempo para cada tipo de dengue, de tal modo que se pudiera visualizar la tendencia de la enfermedad durante el periodo estudiado (Figura 1). Con cada uno de los tipos de dengue se confirma lo reportado por el instituto nacional de salud en sus boletines epidemiológicos, los picos de la enfermedad son ciclicos [2], con brotes que ocurren aproximadamente cada tres años con predominio en los ultimos meses del año, correspondiendo a la segunda temprada de lluvias del país (octubre-diciembre) según el calendario climatico de la unidad nacional para la gestión de riesgos y desastres (UNGRD) [69].

Además del comportamiento ciclico es evidente el aumento progresivo de los casos de cada subtipo de dengue con el paso del tiempo, en general dengue con signos de alarma es la manifestación de la enfermedad más frecuente en nuestro centro de atención (Figura 2A).

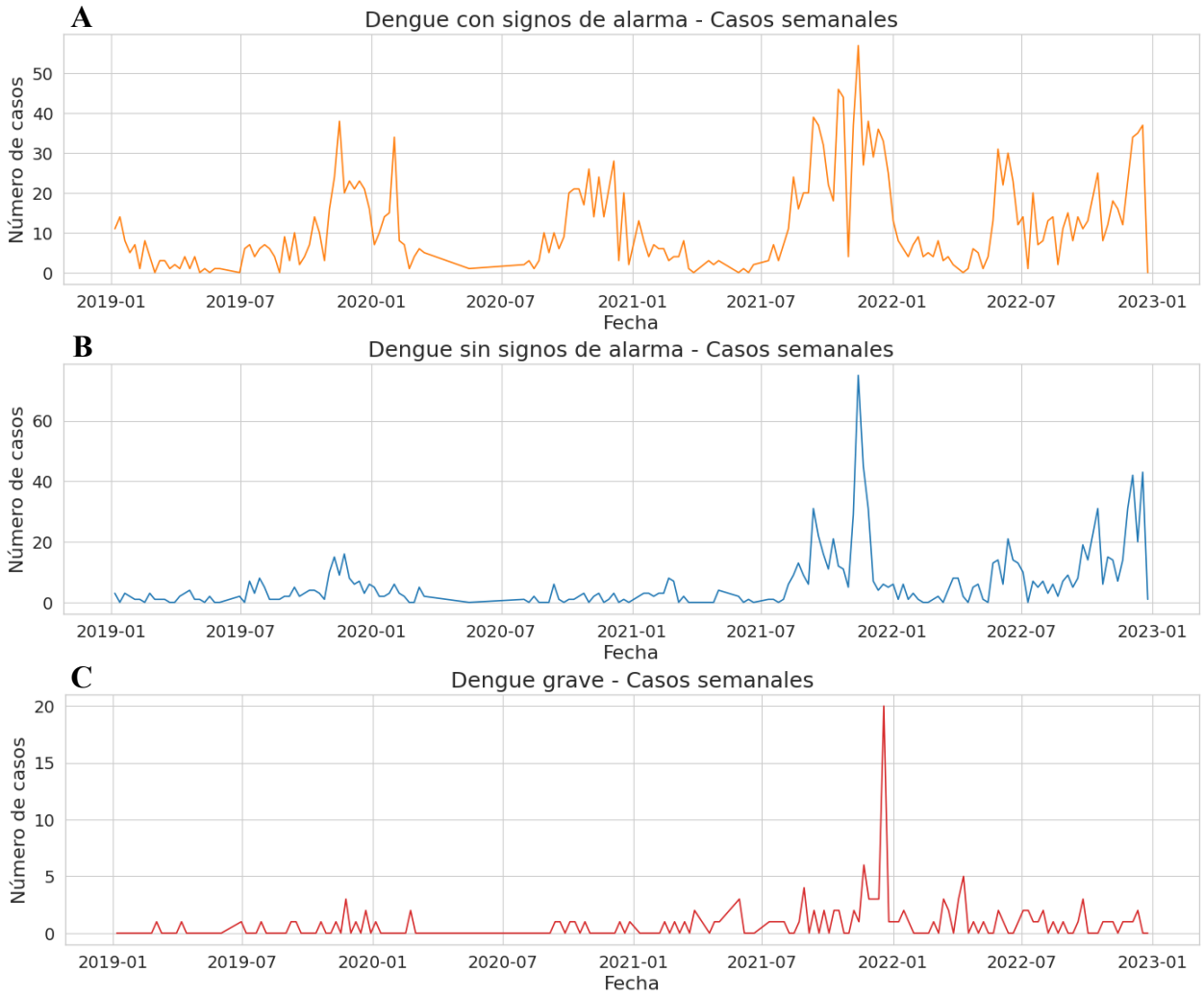


Figura 2. Tendencias de cada subtipo clínico de dengue. A. Dengue sin signos de alarma, B. Dengue con signos de alarma, C. Dengue grave.

Para realizar correctamente las pruebas con los modelos predictivos se dividió toda la base de datos en dos conjuntos, uno grupo de entrenamiento (train) y uno de prueba (test), la selección fue a traves de una división temporal no aleatoria, respetando el estandar 80-20 [70], [71], es decir, 80% de los datos se colocaron en el conjunto de entrenamiento (casos de 2019-2021) y el 20% restante al grupo de prueba (2022), no se realizó conjunto de validación ya que luego de la agrupación semanal no se contaban con los suficientes datos que permitieran una división satisfactoria para los tres conjuntos.

Posteriormente se aplicó un escalado tipo *"Min-max"* para cada subtipo de dengue usando *"MinMaxScaler",* para así convertir los datos en valores entre cero y uno, el escalador se ajustó

únicamente sobre el conjunto de entrenamiento y luego se aplicó al conjunto de prueba, con el fin de evitar información del conjunto de prueba en la transformación del entrenamiento [72], [73].
Para convertir la serie temporal en un conjunto de datos apto para aprendizaje supervisado se realizó una función de ventanas deslizantes [74] con tamaño dos (ventana = 2), esto quiere decir que la predicción de los casos se obtendría al analizar las dos semanas previas, las entradas al modelo serían las semanas -2 y -1 y la salida la semana cero. Como último paso antes de la implmentación de los algoritmos de machine learning se verificó que las matrices generadas no contienen valores nulos (NaN), que no existan ventanas en ceros que no correspondan a casos reales y que todos los valores del escalado correspondan correctamente a su valor sin escalar.

5 ENTRENAMIENTO Y SELECCIÓN DEL MODELO PREDICTIVO

5.1 PROCEDIMIENTO DE ENTRENAMIENTO Y ESTRATEGIAS DE SELECCIÓN DEL MODELO

Los modelos seleccionados para el trabajo fueron: regresion lineal, redes neuronales artificiales, *"Random forest* (RF) y *"support vector machine"* (SVM). La regresión lineal fue usada como modelo base, para luego comparar la red neuronal, RF y SVM.

5.1.1 **Modelos base**

En primer lugar como base utilizamos una regresión lineal múltiple, asumiendo que los casos de una tercera semana se pueden predecir con una combinación lineal de las dos semanas previas. La regresión se ejecutó para cada uno de los subtipos de dengue, es decir, dengue sin signos de alarma (SSA), dengue con signos de alarma (SA) y dengue grave (DG).
Para comparar resultados utilizamos diferentes arquitecturas frecuentemente elegidas en trabajos similares [40], [42], [75], como punto de partida con hiperparámetros muy básicos y sin ajustar. Los algoritmos aplicados junto a sus hiperparámetros básicos se encuentran descritos en la tabla 1.

Tabla 1. Hiperparámetros de los modelos base

Modelo	*Subgrupos aplicados*	*Hiperparámetros fijados manualmente*
Random Forest	SSA, SA, Grave	n_estimators = 100 ; random_state = 42
SVM	SSA, SA, Grave	kernel = 'rbf' (el resto de hiperparámetros en valores por defecto)
MLP	SSA, SA, Grave	hidden_layer_sizes = (64, 32) ; max_iter = 500 ; random_state = 42

Los algoritmos sin optimización fueron el punto de partida, para luego realizar un ajuste de hiperparámetros utilizando *"gridsearch*" y *"earlystopping"*, como métricas de evaluación se incluyó el error absoluto medio (MAE) para obtener la diferencia actual entre las predicciones y los valores reales de manera consistente, expresados en la misma unidad (número de pacientes con el tipo de dengue), el error cuadrático medio (MSE) ya que penaliza errores muy grandes, la raíz del error cuadrático medio (RMSE) para convertir el MSE a la unidad deseada y el valor R^2 para comunicar la calidad de los modelos [76], [77].

5.1.2 Perceptrón multicapa

Se implementaron 24 posibles combinaciones para cada uno de los subtipos clínicos, tres arquitecturas de la red neuronal, dos con solo una capa oculta de dieciséis o treinta y dos neuronas, y una con dos capas ocultas, la primera de treinta y dos neuronas seguida por una de dieciséis. Se contaba también con dos posibles funciones de activación la cuales fueron: unidad lineal rectificada (ReLU) o tangente hiperbólico (tanh). Solo se incluyó el optimizador *"ADAM"* con dos posibles valores de *"learning rate"* de. 0,001 y 0,0005. Se realizó validación cruzada específica para series temporales mediante la clase *"TimeSeriesSplit"* de la biblioteca *"scikit-learn",* la cual preserva el orden temporal de las observaciones y evita fugas de información. El mejor modelo se obtiene usando el error cuadrático medio para determinar la mejor combinación de hiperparámetros. Los mejores hiperparámetros del perceptrón multicapa (MLP) se resumen en la tabla 2.

5.1.3 Random Forest

La grilla del modelor de *"Random forest"* también fue reducida, para evitar un sobreajuste con los pocos datos de entrenamiento y prueba, la complejidad del bosque y la profundidad de cada árbol fueron los hiperparametros a explorar, se mantuvo el mismo método de validación cruzada *"TimeSeriesSplit"* con tres particiones y la misma métrica para la selección de los mejores hiperparámetros MSE. Tabla 2

5.1.4 Support vector machine

El *"kernel"* escogido fue función básica radial (RBF) para capturar patrones no lineales en el

"dataframe", además, se mantuvo un valor "C" bajo para permitir una mejor generalización de datos nuevos. Los valores "*gamma"* y *"épsilon"* también se ajustaron teniendo en cuenta el tamaño de la base de datos, usando la misma clase para validación cruzada y manteniendo la métrica de evaluación para seleccionar los mejores hiperparámetros. Tabla 2

Tabla 2. Búsqueda de hiperparámetros

Modelo	***Hiperparámetros***	***Valores evaluados***
MLP	Número de capas ocultas	(16), (32), (32, 16)
	Función de activación	relu, tanh
	Learning rate	0.001, 0.0005
	Alfa	0.0001, 0.001
	Optimizador	adam
	Número de iteraciones	early stopping=True, max_iter=500
Random Forest	Número de estimaciones	5, 10, 15
	Profundidad máxima	3, 5, 7
	Observaciones por hoja	1, 2, 4
SVM	C	0.1, 1, 10
	Gamma	0.01, 0.1, 1
	Épsilon	0.01, 0.1, 0.2
	Kernel	rbf

5.1.5 Conteo de aciertos exactos

Dado que la variable objetivo es el número de casos de dengue, es discreta, mientras que las predicciones generadas por los modelos son continuas, todas las predicciones fueron transformadas mediante redondeo. Con el fin realizar una comparación directa entre los valores observados y estimados en términos de unidades fácilmente interpretables.

Sea y_i el número real de casos observado y $\hat{y}_i$ la predicción del modelo para la observación i. Se definieron los valores redondeados como:

$$y_i^{(r)} = \text{round}(y_i)$$
$$\hat{y}_i^{(r)} = \text{round}(\hat{y}_i)$$

El número total de aciertos se calculó como:

$$A = \sum_{i \in \mathcal{V}} \mathbf{1}(y_i^{(r)} = \hat{y}_i^{(r)})$$

$\mathbf{1}(\cdot)$ es la función indicadora, que toma el valor de 1 cuando la predicción coincide exactamente con el valor observado, y 0 cuando no. El conjunto $\mathcal{V}$ representa las observaciones válidas.

Finalmente, la proporción de aciertos se estimó mediante:

$$P = \frac{A}{|\mathcal{V}|}$$

$|\mathcal{V}|$ corresponde al número total de observaciones consideradas en la evaluación.

5.1.6 Conteo de tendencias

Además de las métricas clásicas de error, se evaluó la capacidad de los modelos para reproducir correctamente la tendencia temporal de los casos de dengue, muy útil para la vigilancia epidemiológica, independientemente del número exacto de aciertos.

Se calculó definiendo y_t como el número de casos observados en la semana t. La tendencia se modeló a partir de la variación entre semanas consecutivas:

$$\Delta y_t = y_t - y_{t-1}$$

Con base en esta diferencia, se estableció la función de tendencia:

$$T(y_t) = \begin{cases} +1 & \text{si } \Delta y_t > \tau \\ -1 & \text{si } \Delta y_t < -\tau \\ 0 & \text{si } |\Delta y_t| \leq \tau \end{cases}$$

donde τ representa un umbral de tolerancia que define cambios irrelevantes (en este estudio, $\tau = 1$ caso).

De manera análoga, se calcularon las tendencias para las predicciones del modelo $\hat{y}_t$:

$$T(\hat{y}_t)$$

La exactitud en la predicción de tendencias se definió como:

$$A = \sum_{t=2}^{n} \mathbf{1}(T(y_t) = T(\hat{y}_t))$$

$\mathbf{1}(\cdot)$es la función indicadora. Finalmente, la proporción de coincidencias se calculó como:

$$P = \frac{A}{n-1}$$

5.2 RESULTADOS DEL MODELO

La optimización de hiperparámetros se llevó a cabo mediante búsqueda en cuadrícula *("GridSearch"),* utilizando como función objetivo el error cuadrático medio negativo. Este procedimiento permitió seleccionar la combinación de parámetros que minimizara el error de predicción durante la validación. Como la variable objetivo fue previamente escalada, las predicciones se transformaron nuevamente a la escala original antes de calcular las métricas de desempeño, asegurando una interpretación clínica adecuada.

Con la regresión lineal obtuvimos un error absoluto medio (MAE) y una raíz cuadrada del error absoluto medio (RMSE) altos, 8,25 y 11,64 respectivamente para dengue sin signos de alarma y 7,07 y 9,61 para dengue con signos de alarma (Tabla 3), el valor R^2 demuestra que el modelo falló en capturar la variación de los casos, presentando un puntaje negativo para SSA y cercanos a cero para SA y DG (Tabla 3). El R^2 negativo nos indica que las predicciones son peores que simplemente tomar la media de los casos como valor futuro, mientras que números cercanos a cero significa que el modelo no logra capturar bien las variaciones reales de los datos [78]. La regresión lineal asume una correlación lineal entre las semanas predictoras y la semana objetivo, este abordaje resulta no ser suficiente para capturar el comportamiento de la enfermedad (Figura 3). Luego de evaluar las métricas se justifica el uso de modelos que funcionan mejor con datos no lineales y con mucho ruido como perceptrón multicapa, "*random forest*" y *"support vector machine".*

Tabla 3: Métricas de evaluación para modelo de regresión lineal

Modelo	*Subgrupo*	*MAE*	*RMSE*	*R²*
Regresión Lineal	SSA	8.25	11.64	**–0.8**
Regresión Lineal	SA	7.07	9.61	**0.06**
Regresión Lineal	Grave	0.70	0.85	**–0.14**

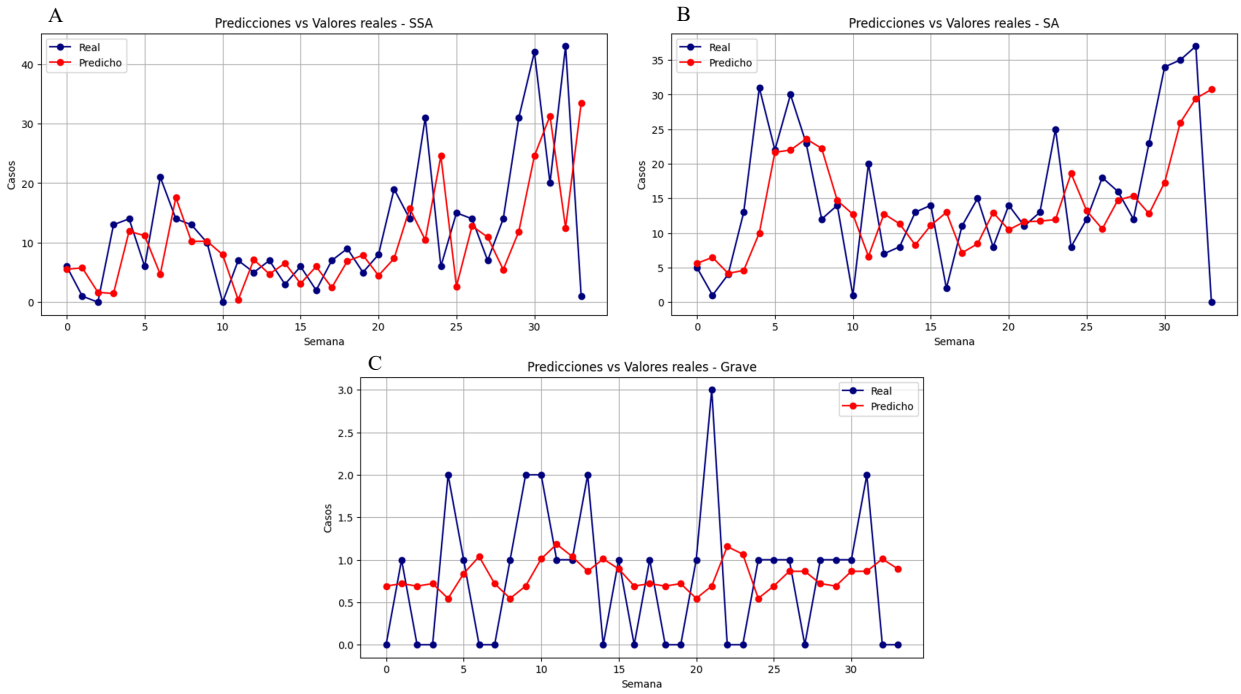


Figura 3: Predicciones de la regresión lineal. A: Dengue sin signos de alarma, B: Dengue con signos de alarma, C: Dengue grave

En la Tabla 4 se resumen las métricas de desempeño de los modelos base para cada subgrupo clínico. Los errores (MAE y RMSE) no variaron mucho a los obtenidos con la regresión lineal, con un R^2 bajo o incluso negativo en algunos casos.

Tabla 4: Métricas de evaluación de los modelos base

Modelo	*Subgrupo*	*MAE*	*RMSE*	*R2*
MLP	SSA	6.81	9.88	0.15
Random Forest	SSA	7.67	10.91	-0.04
SVR	SSA	6.86	9.71	0.18
MLP	SA	7.41	9.68	0.04
Random Forest	SA	7.95	10.31	-0.09
SVR	SA	7.79	9.99	-0.02
MLP	Grave	0.75	0.91	-0.32
Random Forest	Grave	0.87	1.1	-0.92
SVR	Grave	1.23	1.39	-2

Luego de la optimización de hiperparámetros, las arquitecturas con mejores MSE fueron seleccionadas (Tabla 5).

Tabla 5. Hiperparámetros óptimos

Modelo	*Subgrupo*	*Hiperparámetros óptimos*
MLP	SSA	Activación = *tanh*; α = 0.001; Capas ocultas = (32, 16); Tasa de aprendizaje = 0.001; Optimizador = *adam*
MLP	SA	Activación = *tanh*; α = 0.0001; Capas ocultas = (32, 16); Tasa de aprendizaje = 0.001; Optimizador = *adam*
MLP	Grave	Activación = *relu*; α = 0.001; Capas ocultas = (32, 16); Tasa de aprendizaje = 0.001; Optimizador = *adam*
Random Forest	SSA	n_estimators = 5; max_depth = 5; min_samples_leaf = 1
Random Forest	SA	n_estimators = 5; max_depth = 5; min_samples_leaf = 4
Random Forest	Grave	n_estimators = 15; max_depth = 3; min_samples_leaf = 4
SVR	SSA	C = 10; ε = 0.01; γ = 0.1; kernel = RBF
SVR	SA	C = 10; ε = 0.01; γ = 0.1; kernel = RBF
SVR	Grave	C = 0.1; ε = 0.1; γ = *scale*; kernel = RBF

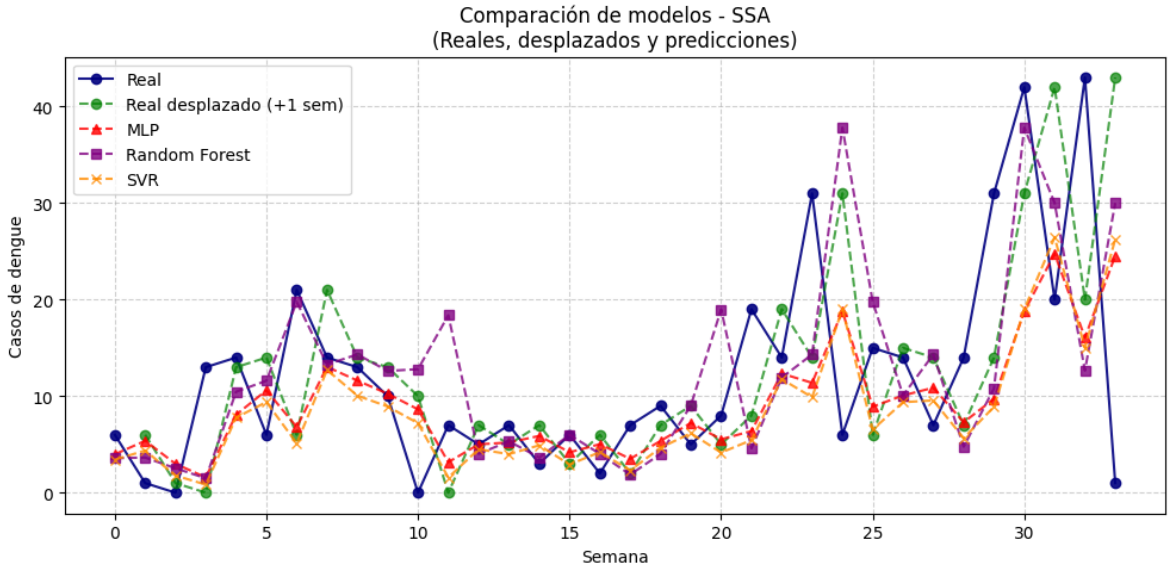
En general, para cada subtipo de dengue los hiperparámetros ideales son bastante similares, presentando cambios principalmente con dengue grave, con el MLP cambia solamente la función de activación de tangente hiperbólica a unidad lineal rectificada, el modelo *"random forest"* y *"support vector machine"* fueron los que mayor variación presentaron para cada variante clínica. Una vez terminada la optimización se obtuvieron resultados parecidos a los presentados con los algoritmos sin ajuste. Tabla 6

Tabla 6. Desempeño global de todos los algoritmos

Subtipo de dengue	*Enfoque de modelación*	*MAE*	*MSE*	*RMSE*	*R²*
Sin signos de alarma (SSA)	Regresión lineal	8.25	135.40	11.64	−0.18
	Perceptrón multicapa	6.81	97.65	9.88	0.15
	Random forest	7.67	119.00	10.91	−0.04
	Support vector machine	6.86	94.24	9.71	0.18
	MLP con ajuste de hiperparámetros	7.32	108.53	10.42	0.05
	RF con ajuste	8.02	137.11	11.71	−0.20
	SVM con ajuste	7.84	120.61	10.98	−0.05
Con signos de alarma (SA)	Regresión lineal	7.07	92.36	9.61	0.06
	Perceptrón multicapa	7.41	93.75	9.68	0.04

	Random forest	7.95	106.27	10.31	−0.09
	Support vector machine	7.79	99.78	9.99	−0.02
	MLP con ajuste de hiperparámetros	7.10	91.02	9.54	0.07
	RF con ajuste	7.80	98.41	9.92	0.00
	SVM con ajuste	7.40	96.43	9.82	0.02
Grave	Regresión lineal	0.70	0.72	0.85	−0.14
	Perceptrón multicapa	0.75	0.83	0.91	−0.32
	Random forest	0.87	1.22	1.10	−0.92
	Support vector machine	1.23	1.95	1.39	−2.00
	MLP con ajuste de hiperparámetros	0.79	0.90	0.95	−0.42
	RF con ajuste	0.86	1.08	1.04	−0.70
	SVM con ajuste	1.26	2.08	1.44	−2.28

Para evaluar aún más las arquitecturas se decidió crear un modelo ingenuo, en donde la predicción se basa simplemente en repetir el último valor [79], con el fin de comparar si de verdad el uso de estrategias de aprendizaje automático es mejor que simplemente predefinir la tendencia del valor futuro usando exactamente el valor pasado. Figura 4



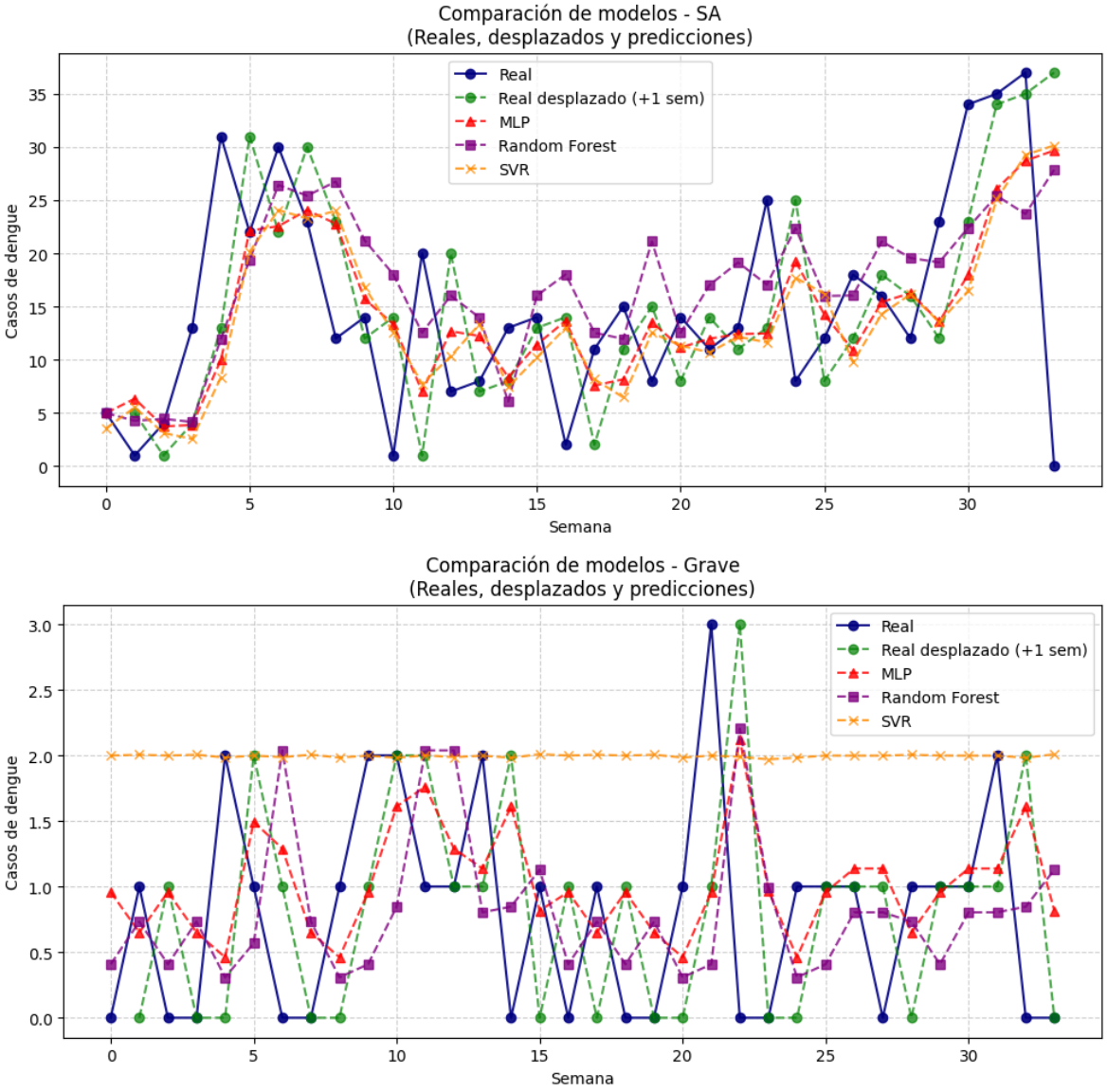


Figura 4. Comparativa entre las predicciones entre cada modelo optimizado vs modelo ingenuo

Una vez listas todas las predicciones cada método se comparó realizando el conteo de aciertos totales, además, también se evaluó la cantidad de veces que los algoritmos lograban predecir la tendencia de los casos de dengue, es decir, si en el futuro los casos aumentarían o disminuirían, esto como una medida más flexible de evaluación.

Para aciertos exactos el MLP optimizado obtuvo los mejores resultados, sin embargo, para la captura de la tendencia del aumento y disminución de casos, RF se desempeñó mejor. El modelo ingenuo también presentó varios aciertos a la hora de capturar la tendencia, pero solo evidenció predicciones exactas para dengue grave, de todos, el algoritmo con resultados más estables fue el perceptrón multicapa, por lo que al final se seleccionó como modelo definitivo. Tabla 7

Tabla 7. Predicción exacta y de tendencias

Subtipo de dengue	***Enfoque de predicción***	***Aciertos exactos (casos)***	***Total semanas***	***Exactitud exacta (%)***	***Aciertos de tendencia***	***Total comparaciones***	***Exactitud de tendencia (%)***
Sin signos de alarma	Ingenuo (t−1)	0	33	0.00	5	33	15.15
	Red neuronal multicapa (MLP)	2	34	5.88	5	33	15.15
	Bosque aleatorio	1	34	2.94	15	33	45.45
	Máquina de vectores de soporte	1	34	2.94	6	33	18.18
Con signos de alarma	Ingenuo (t−1)	0	33	0.00	8	33	24.24
	Red neuronal multicapa (MLP)	3	34	8.82	6	33	18.18
	Bosque aleatorio	2	34	5.88	7	33	21.21
	Máquina de vectores de soporte	2	34	5.88	6	33	18.18
Dengue grave	Ingenuo (t−1)	11	33	33.33	24	33	72.73
	Red neuronal multicapa (MLP)	11	34	32.35	26	33	78.79
	Bosque aleatorio	11	34	32.35	23	33	69.70
	Máquina de vectores de soporte	5	34	14.71	28	33	84.85

6 INTERPRETACIÓN DE RESULTADOS Y APLICACIONES PRÁCTICAS

Las enfermedades transmitidas por vectores son un problema de salud pública mundial, Colombia cuenta con todas las condiciones climáticas y geográficas para la propagación de tales patologías, actualmente tenemos un robusto sistema de vigilancia de eventos de interés en salud (SIVIGILA) las herramientas que ofrece se quedan con un enfoque descriptivo, muy útil si, para determinar tendencias pero a la hora de la predicción de brotes y epidemias existen mejores alternativas, mucho más sofisticadas para proteger la salud de la población [80].

Si bien bibliográficamente se evidencia un ciclo repetitivo de casos de dengue concentrado en períodos de lluvia, solo con observar la serie de tiempo podemos realizar deducciones (figuras 1 y 2). En primer lugar los casos de los tres subtipos clínicos de dengue aumentaron anualmente durante el periodo observado (2019-2022), pero ese aumento no es homogéneo entre semana y semana, hay brotes explosivos de casos nuevos, seguidos por valles abruptos, incluso dentro de los meses lluviosos, (octubre a diciembre) [69] los reportes de casos no parecen seguir un patrón estable, nuestra serie de tiempo es ruidosa. Otro de los hallazgos es que dengue con signos de alarma es la clasificación más frecuente. Dengue grave tiene frecuencias interesantes, los casos con esta clasificación no fueron muchos, las semanas con cero casos reportados no son infrecuentes y presenta además diferentes picos aislados pero de pocos casos. Dengue sin signos de alarma solo presentó un brote abrupto en 2021.

La regresión lineal no pudo comprender por completo las dinámicas del dengue, por eso los R^2 negativos y valores altos de error absoluto medio y RMSE (Tabla 6), con un error aproximado entre semana de 8 pacientes, sin embargo, la raíz del error cuadrático medio evidencia la presencia de errores mayores al promedio de casos semanales, es de esperarse en series de tiempo para patologías infecciosas, los patrones de contagio, suelen tener un comportamiento errático, que al asumir asociación lineal no se pueden capturar del todo [81].

Los métodos predictivos base sin búsqueda activa de sus hiperparámetros obtuvieron resultados comparables a la regresión lineal pero con ligeras mejorías, teniendo la máquina de soporte vectorial y la red neuronal las mejores métricas, esto concuerda con estudios similares, en dónde para la predicción de enfermedades también obtienen mejores resultados [81].

Las métricas de evaluación tras las optimización no mostraron cambios significativos, de hecho en la mayoría de los casos empeoran a comparación del modelo base, exceptuando a la predicción dengue con signos de alarma de la red neuronal, en todas las métricas de interés. Continuando con el mismo subtipo clínico (DCSA), el bosque aleatorio mejoró ligeramente el MAE, pero empeoró el RMSE que penaliza errores, lo cual en un contexto clínico es muy valioso, además al obtener un R^2 en cero realmente el modelo no se diferencia de usar como valor futuro la media de los casos actuales. Dengue grave presentó las peores métricas, es muy probable que se deba a la gran cantidad de valores en cero que había por semana, con cambios abruptos difíciles de predecir. Tabla 6

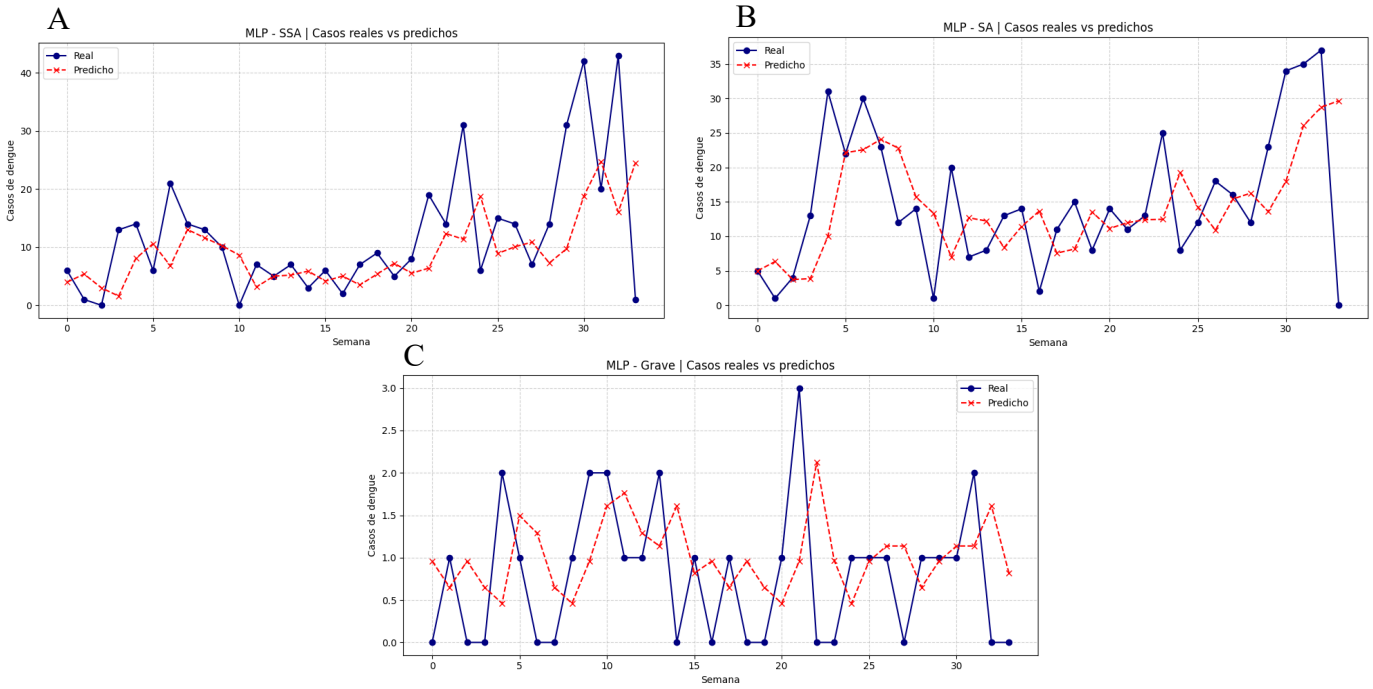


Figura 5. Perceptrón multicapa optimizado. A dengue sin signos de alarma, B dengue con signos de alarma, C dengue grave

Como ultimo tipo de validación se empleó una predicción ingenua *("naïve model")* que simplemente predecirá el número de casos de la semana siguiente como el mismo que la semana actual, con la evaluación de aciertos exactos nuevamente el MLP presentó mejoras con una exactitud del 5,8 y 8,8% en comparación con el modelo ingenuo y las otras arquitecturas, (Tabla 7) RF, sin embargo, fue el modelo que pudo capturar de forma más precisa la tendencia de los datos incluso superando el bosque aleatorio. La red neuronal a grandes rasgos fue la que tuvo mejor desempeño, ya que tiene la mayor cantidad de aciertos exactos, y una predicción de tendencias estables en los diferentes subgrupos clínicos, lo cual está acorde con diversos estudios previos en los que las redes neuronales son los mejores modelos para capturar el comportamiento de enfermedades infecciosas [39], [40].

Con el fin de explorar la aplicabilidad práctica de los modelos desarrollados, se diseñó un prototipo de herramienta interactiva orientada al apoyo a la vigilancia epidemiológica del dengue. Si bien las métricas obtenidas por los modelos no fueron las esperadas, si hubo mejoría con el perceptrón multicapa luego de la optimización, cabe aclarar, que el uso de esta herramienta no pretende desplazar las alertas y recursos propuestos por el instituto nacional de salud, debe ser usada como complemento para la vigilancia de la enfermedad y solo en el contexto de la institución de la que se recibió la información (Anexo 1).

Más allá del desempeño y aplicabilidad computacional ¿qué significan estos resultados en el sector salud?
Cuando creamos un modelo que se aplicará a personas, sea en el contexto sanitario o no, siempre debemos tener en cuenta la cantidad de factores que pueden explicar una variable, en el caso específico de las enfermedades infecciosas la complejidad de las interacciones entre variables es mucha, existen diversos factores que impactan la incidencia de una enfermedad, podemos señalar algunos obvios, como el periodo climático o la presencia del mosquito que la transmite, pero existen dinámicas que no son tan obvias, desde factores genéticos, costumbres sociales, etc. [82]. Las enfermedades no se comportan de una forma perfecta y esta es una de las principales limitaciones para el modelado de predicciones específicamente con patologías infecciosas, evidentemente, solo contar con el número de casos neto no es suficiente para acercarnos a una predicción fiel de la enfermedad, adicionalmente solo se contó con un periodo muy limitado de tiempo que al transformar nuestros datos para el análisis semanal disminuyó drásticamente la cantidad de información disponible para el entrenamiento y las posterior prueba de los algoritmos, la red neuronal optimizada obtuvo mejores métricas para el tipo de dengue con más casos disponibles.

Por último, otro punto importante que pudo afectar a nuestro análisis es el hecho del periodo donde se obtuvieron los datos, el *"dataframe"* incluye un año problemático, el 2020, en donde hubo un descenso en la notificación de diversas enfermedades por la prioridad que establecía la enfermedad por COVID-19 [83], es probable que los casos de 2020 no sean en realidad la cantidad de casos que realmente atendió la institución, y con menos casos, menor robustez de entrenamiento. En general la poca cantidad de datos acompañada de solo el uso de los casos semanales para la predicción del comportamiento epidemiológico del dengue son las principales limitaciones de nuestro estudio.

Todos nuestros resultados deben interpretarse con cautela, en sí, es posible realizar la predicción de la incidencia del Dengue, no obstante, esta enfermedad representa retos para el correcto entrenamiento y aplicabilidad de algoritmos de análisis predictivo.

7 CONCLUSIONES Y TRABAJOS FUTUROS

7.1 CONCLUSIONES

Con nuestros resultados podemos concluir que el uso de *"machine learning"* para la predicción de enfermedades es posible, con el presente trabajo logramos desarrollar y evaluar varios modelos para dengue, sin embargo, la implementación en el sector salud supone algunos retos. Si bien se tiene identificada la alta tasa de dengue en Colombia, hay un crecimiento rápido de la cantidad de casos por año, la enfermedad tiene un comportamiento errático y ruidoso que dificulta la adecuada predicción de los modelos de los fenómenos epidemiológicos intrínsecos de la enfermedad.

El modelo predictivo con mejores resultados fue la red neuronal artificial, lo cual es respaldado por numerosos articulos previamente descritos [40], [45], [51], las redes neuronales son bastante versatiles a la hora de crear relaciones con datos no lineales y tan aleatorios como lo son los brotes de una enfermedad infecciosa, el modelo tiene mejores resultados especificamente en periodos más controlados de la enfermedad y espcificamente con el subtipo clinico de dengue con signos de alarma. Sin embargo, no existe una arquitectura universalmente superior a la hora de predecir los casos de dengue con nuestros datos, el ajuste de hiperparametros no mejoró significativamente el desempeño de ningún modelo, llegando incluso a empeorar metricas de evaluación.

Dengue grave fue la clasificación de dengue más dificil de modelar, esto puede deberse a la poca cantidad de casos de dengue grave durante toda la serie de tiempo, existiendo multiples semanas sin reporte de casos, esta cantidad de valores en cero dificulta el procesamiento de las arquitecturas de *"machine learning"*.

La comparación con un modelo ingenuo es indispensable para evaluar de forma integral el desempeño de cada algoritmo, siempre debe realizarse como una metrica adicional.

De este trabajo se puede aprender que la integración del *"machine learning"* a la salud publica es compleja y requiere ser abordada intersectorialmente, las métricas matemáticas por si solas no son el unico criterio de utilidad, siempre se deben tener en cuenta los diferentes factores que acompañan al fenómeno, priorizando la interpretabilidad de los modelos para los profesionales que tomarán decisiones con base a ellos.

7.2 TRABAJOS FUTUROS

Para trabajos futuros se recomendaría en primera instancia expandir el periodo de la serie de tiempo, así como implementar el estudio de manera multicéntrica, es decir, en varios hospitales y clínicas a la vez, esto con el fin de aumentar la cantidad de casos disponibles para el entrenamiento de los algoritmos. Además, añadir variables sociodemográficas como la procedencia de los pacientes, el grupo etario y nivel socioeconómico, así como variables clínicas como presencia de comorbilidades o síntomas de alerta, permitirían enriquecer el modelo en el contexto clínico. Otro grupo de variables que deberían incluirse en estudios futuros son las inherentes a procesos climáticos, como temperatura, humedad y frecuencia de precipitaciones.

Para mejorar la robustez del estudio se podrían integrar más metodologías tanto de aprendizaje automático como estadísticas, incluso la creación de diferentes modelos mixtos para evaluar su efectividad en la predicción de la incidencia de la enfermedad.

Por último, sería útil comprobar el prototipo en un entorno clínico durante un periodo de tiempo controlado, es decir, implementar su uso en la institución durante periodos sin brote y con brotes para medir su efectividad en el mundo real.

8 REFERENCIAS BIBLIOGRÁFICAS

[1] «Dengue virus infection – a review of pathogenesis, vaccines, diagnosis and therapy - PMC». Accedido: 18 de noviembre de 2024. [En línea]. Disponible en: https://pmc.ncbi.nlm.nih.gov/articles/PMC10194131/

[2] Instituto Nacional de Salud, «Protocolo de vigilancia en salud pública Dengue». 15 de julio de 2024. Accedido: 18 de noviembre de 2024. [En línea]. Disponible en: https://www.ins.gov.co/buscador-eventos/Lineamientos/Pro_Dengue.pdf

[3] «Dengue and severe dengue». Accedido: 4 de noviembre de 2024. [En línea]. Disponible en: https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue

[4] «Infecciones víricas emergentes: fiebre amarilla, dengue, chikungunya, zika, fiebre hemorrágica de Crimea-Congo, enfermedad por el virus del Ébola y otras viriasis. Rabia - ClinicalKey». Accedido: 4 de noviembre de 2024. [En línea]. Disponible en: https://www.clinicalkey.es/#!/content/book/3-s2.0-B9788413824864003097?scrollTo=%23hl0000436

[5] «Dengue - ClinicalKey». Accedido: 4 de noviembre de 2024. [En línea]. Disponible en: https://www.clinicalkey.es/#!/topic/dengue?topic=dengue

[6] S. Halstead, «Recent advances in understanding dengue», *F1000Research*, vol. 8, p. F1000 Faculty Rev, jul. 2019, doi: 10.12688/f1000research.19197.1.

[7] N. Endo, A. Goto, T. Suzuki, S. Matsuda, y S. Yasumura, «Factors associated with enrollment and adherence in outpatient cardiac rehabilitation in Japan», *J. Cardiopulm. Rehabil. Prev.*, vol. 35, n.º 3, pp. 186-192, 2015, doi: 10.1097/HCR.0000000000000103.

[8] «Dengue: Guías para el diagnóstico, tratamiento, prevención y control; 2009 - OPS/OMS | Organización Panamericana de la Salud». Accedido: 4 de noviembre de 2024. [En línea]. Disponible en: https://www.paho.org/es/documentos/dengue-guias-para-diagnostico-tratamiento-prevencion-control-2009

[9] «Modelos predictivos y prescriptivos para dengue en Córdoba - Investigación / i+D+I en TIC - Universidad EAFIT». Accedido: 4 de noviembre de 2024. [En línea]. Disponible en: https://www.eafit.edu.co/investigacion/grupos/i-d-i-tic/Paginas/modelos-predictivos-y-prescriptivos-para-dengue-en-cordoba.aspx

[10] Colombia. Ministerio de Salud y Protección Social. Dirección de Desarrollo de Talento Humano en Salud, *Dengue: memorias*. 2013. [En línea]. Disponible en: https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/VS/TH/Memorias_dengue.pdf

[11] «Actualización Epidemiológica - Aumento de casos de dengue en la Región de las Américas - 18 de junio del 2024 - OPS/OMS | Organización Panamericana de la Salud». Accedido: 4 de noviembre de 2024. [En línea]. Disponible en: https://www.paho.org/es/documentos/actualizacion-epidemiologica-aumento-casos-dengue-region-americas-18-junio-2024

[12] «El Ministerio de Salud llama a intensificar las medidas de prevención y control del dengue ante el impacto del fenómeno de El Niño y la Niña». Accedido: 4 de noviembre de 2024. [En línea]. Disponible en: https://www.minsalud.gov.co/Paginas/intensificar-las-

medidas-de-prevencion-y-control-del-dengue.aspx
[13] «26 agosto 2024. Dengue en Colombia. 243,538 casos y 106 muertes», Fundación iO. Accedido: 4 de noviembre de 2024. [En línea]. Disponible en: https://fundacionio.com/dengue-en-colombia-243-538-casos-y-106-muertes/
[14] F. Ruiz-López *et al.*, «Presencia de Aedes (Stegomyia) aegypti (Linnaeus, 1762) y su infección natural con el virus del dengue en alturas no registradas para Colombia», *Biomédica*, vol. 36, n.º 2, pp. 303-308.
[15] R. Holzmann y S. Jorgensen, «Manejo social del riesgo: un nuevo marco conceptual para la protección social y más allá», *Rev Fac Nac Salud Pública*, pp. 73-106, 2003.
[16] «Alerta epidemiológica por dengue en Colombia». Accedido: 25 de noviembre de 2024. [En línea]. Disponible en: https://www.minsalud.gov.co/Paginas/Alerta-epidemiologica-por-dengue-en-Colombia.aspx
[17] «Dengue overview: An updated systemic review - ScienceDirect». Accedido: 18 de noviembre de 2024. [En línea]. Disponible en: https://www.sciencedirect.com/science/article/pii/S1876034123002587?via%3Dihub
[18] S. A. Kularatne y C. Dalugama, «Dengue infection: Global importance, immunopathology and management», *Clin. Med.*, vol. 22, n.º 1, p. 9, ene. 2022, doi: 10.7861/clinmed.2021-0791.
[19] «Dengue - OPS/OMS | Organización Panamericana de la Salud». Accedido: 18 de noviembre de 2024. [En línea]. Disponible en: https://www.paho.org/es/temas/dengue
[20] O. Peláez Sánchez y P. Más Bermejo, «Brotes, epidemias, eventos y otros términos epidemiológicos de uso cotidiano», *Rev. Cuba. Salud Pública*, vol. 46, p. e2358, oct. 2020.
[21] P. Kuri-Morales, «Las pandemias: el COVID-19», *Cir. Cir.*, vol. 88, n.º 3, pp. 249-251, jun. 2020, doi: 10.24875/ciru.20000234.
[22] *2. Introduction to Predictive Analytics and Data Mining*. Accedido: 19 de noviembre de 2024. [En línea]. Disponible en: https://learning.oreilly.com/library/view/predictive-analytics-data/9780135946527/ch02.xhtml
[23] *Chapter 1: An Introduction to Data Mining and Predictive Analytics*. Accedido: 19 de noviembre de 2024. [En línea]. Disponible en: https://learning.oreilly.com/library/view/data-mining-and/9781118868706/9781118868706c01.xhtml
[24] K. Murphy, «Introduction», en *Machine Learning A Probabilistic Perspective*, Primera., MIT Press, pp. 1-24.
[25] R. H. Shumway y D. S. Stoffer, «Time Series Regression and Exploratory Data Analysis», en *Time Series Analysis and Its Applications: With R Examples*, R. H. Shumway y D. S. Stoffer, Eds., Cham: Springer International Publishing, 2017, pp. 45-74. doi: 10.1007/978-3-319-52452-8_2.
[26] R. H. Shumway y D. S. Stoffer, «ARIMA Models», en *Time Series Analysis and Its Applications: With R Examples*, R. H. Shumway y D. S. Stoffer, Eds., Cham: Springer International Publishing, 2017, pp. 75-163. doi: 10.1007/978-3-319-52452-8_3.
[27] S. L. Ho y M. Xie, «The use of ARIMA models for reliability forecasting and analysis», *Comput. Ind. Eng.*, vol. 35, n.º 1, pp. 213-216, oct. 1998, doi: 10.1016/S0360-8352(98)00066-7.
[28] X. Zhang, X. Zhang, y W. Wang, «Artificial Neural Network», en *Intelligent Information Processing with Matlab*, X. Zhang, X. Zhang, y W. Wang, Eds., Singapore: Springer Nature,

2023, pp. 1-37. doi: 10.1007/978-981-99-6449-9_1.
[29] E. Guresen y G. Kayakutlu, «Definition of artificial neural networks with comparison to other networks», *Procedia Comput. Sci.*, vol. 3, pp. 426-433, ene. 2011, doi: 10.1016/j.procs.2010.12.071.
[30] B. Müller, J. Reinhardt, y M. T. Strickland, «Neural Networks Introduced», en *Neural Networks: An Introduction*, B. Müller, J. Reinhardt, y M. T. Strickland, Eds., Berlin, Heidelberg: Springer, 1995, pp. 13-23. doi: 10.1007/978-3-642-57760-4_2.
[31] *1. Gearing Up for Predictive Modeling*. Accedido: 19 de noviembre de 2024. [En línea]. Disponible en: https://learning.oreilly.com/library/view/mastering-predictive-analytics/9781787121393/ch01.html
[32] *Getting Started with Predictive Analytics*. Accedido: 19 de noviembre de 2024. [En línea]. Disponible en: https://learning.oreilly.com/library/view/practical-predictive-analytics/9781785886188/ed7c07a1-cbfd-4ffb-b0b5-2d350c91ccf8.xhtml
[33] *9 Splitting data by asking questions: Decision trees*. Accedido: 3 de noviembre de 2025. [En línea]. Disponible en: https://learning.oreilly.com/library/view/grokking-machine-learning/9781617295911/Text/09.xhtml
[34] *7. Ensemble Learning and Random Forests*. Accedido: 3 de noviembre de 2025. [En línea]. Disponible en: https://learning.oreilly.com/library/view/hands-on-machine-learning/9781098125967/ch07.html
[35] *3 Drawing a line close to our points: Linear regression*. Accedido: 3 de noviembre de 2025. [En línea]. Disponible en: https://learning.oreilly.com/library/view/grokking-machine-learning/9781617295911/Text/03.xhtml
[36] *15. Processing Sequences Using RNNs and CNNs*. Accedido: 15 de febrero de 2026. [En línea]. Disponible en: https://learning.oreilly.com/library/view/hands-on-machine-learning/9781098125967/ch15.html
[37] *3 Drawing a line close to our points: Linear regression*. Accedido: 15 de febrero de 2026. [En línea]. Disponible en: https://learning.oreilly.com/library/view/grokking-machine-learning/9781617295911/Text/03.xhtml
[38] «2. End-to-End Machine Learning Project | Hands-On Machine Learning with Scikit-Learn and PyTorch». Accedido: 15 de febrero de 2026. [En línea]. Disponible en: https://learning.oreilly.com/library/view/hands-on-machine-learning/9798341607972/ch02.html#id49
[39] T. Chakraborty, S. Chattopadhyay, y I. Ghosh, «Forecasting dengue epidemics using a hybrid methodology», *Phys. Stat. Mech. Its Appl.*, vol. 527, p. 121266, ago. 2019, doi: 10.1016/j.physa.2019.121266.
[40] A. Kukkar, Y. Kumar, J. K. Sandhu, M. Kaur, T. S. Walia, y M. Amoon, «DengueFog: A Fog Computing-Enabled Weighted Random Forest-Based Smart Health Monitoring System for Automatic Dengue Prediction», *Diagnostics*, vol. 14, n.º 6, Art. n.º 6, ene. 2024, doi: 10.3390/diagnostics14060624.
[41] D. K. Ming *et al.*, «The Diagnosis of Dengue in Patients Presenting With Acute Febrile Illness Using Supervised Machine Learning and Impact of Seasonality», *Front. Digit. Health*, vol. 4, mar. 2022, doi: 10.3389/fdgth.2022.849641.
[42] J. Xu *et al.*, «Forecast of Dengue Cases in 20 Chinese Cities Based on the Deep Learning Method», *Int. J. Environ. Res. Public. Health*, vol. 17, n.º 2, p. 453, ene. 2020, doi:

10.3390/ijerph17020453.
[43] A. L. Ramadona, L. Lazuardi, Y. L. Hii, Å. Holmner, H. Kusnanto, y J. Rocklöv, «Prediction of Dengue Outbreaks Based on Disease Surveillance and Meteorological Data», *PLOS ONE*, vol. 11, n.º 3, p. e0152688, mar. 2016, doi: 10.1371/journal.pone.0152688.
[44] Y. Chen, C. W. Chu, M. I. C. Chen, y A. R. Cook, «The utility of LASSO-based models for real time forecasts of endemic infectious diseases: A cross country comparison», *J. Biomed. Inform.*, vol. 81, pp. 16-30, may 2018, doi: 10.1016/j.jbi.2018.02.014.
[45] D. Phung *et al.*, «Identification of the prediction model for dengue incidence in Can Tho city, a Mekong Delta area in Vietnam», *Acta Trop.*, vol. 141, pp. 88-96, ene. 2015, doi: 10.1016/j.actatropica.2014.10.005.
[46] S. A. Lauer *et al.*, «Prospective forecasts of annual dengue hemorrhagic fever incidence in Thailand, 2010–2014», *Proc. Natl. Acad. Sci.*, vol. 115, n.º 10, pp. E2175-E2182, mar. 2018, doi: 10.1073/pnas.1714457115.
[47] P. Siriyasatien, A. Phumee, P. Ongruk, K. Jampachaisri, y K. Kesorn, «Analysis of significant factors for dengue fever incidence prediction», *BMC Bioinformatics*, vol. 17, n.º 1, p. 166, abr. 2016, doi: 10.1186/s12859-016-1034-5.
[48] Y. Chen, J. H. Y. Ong, J. Rajarethinam, G. Yap, L. C. Ng, y A. R. Cook, «Neighbourhood level real-time forecasting of dengue cases in tropical urban Singapore», *BMC Med.*, vol. 16, n.º 1, p. 129, ago. 2018, doi: 10.1186/s12916-018-1108-5.
[49] Y. Shi *et al.*, «Three-Month Real-Time Dengue Forecast Models: An Early Warning System for Outbreak Alerts and Policy Decision Support in Singapore», *Environ. Health Perspect.*, vol. 124, n.º 9, pp. 1369-1375, sep. 2016, doi: 10.1289/ehp.1509981.
[50] S.-W. Huang, H.-P. Tsai, S.-J. Hung, W.-C. Ko, y J.-R. Wang, «Assessing the risk of dengue severity using demographic information and laboratory test results with machine learning», *PLoS Negl. Trop. Dis.*, vol. 14, n.º 12, p. e0008960, dic. 2020, doi: 10.1371/journal.pntd.0008960.
[51] D. K. Ming *et al.*, «Applied machine learning for the risk-stratification and clinical decision support of hospitalised patients with dengue in Vietnam», *PLOS Digit. Health*, vol. 1, n.º 1, p. e0000005, ene. 2022, doi: 10.1371/journal.pdig.0000005.
[52] N. Seltenrich, «Singapore Success: New Model Helps Forecast Dengue Outbreaks», *Environ. Health Perspect.*, vol. 124, n.º 9, pp. A167-A167, sep. 2016, doi: 10.1289/ehp.124-A167.
[53] M. Gharbi *et al.*, «Time series analysis of dengue incidence in Guadeloupe, French West Indies: Forecasting models using climate variables as predictors», *BMC Infect. Dis.*, vol. 11, n.º 1, p. 166, jun. 2011, doi: 10.1186/1471-2334-11-166.
[54] E. Z. Martinez, E. A. S. da Silva, y A. L. D. Fabbro, «A SARIMA forecasting model to predict the number of cases of dengue in Campinas, State of São Paulo, Brazil», *Rev. Soc. Bras. Med. Trop.*, vol. 44, n.º 4, pp. 436-440, 2011, doi: 10.1590/s0037-86822011000400007.
[55] O. S. Baquero, L. M. R. Santana, y F. Chiaravalloti-Neto, «Dengue forecasting in São Paulo city with generalized additive models, artificial neural networks and seasonal autoregressive integrated moving average models», *PLOS ONE*, vol. 13, n.º 4, p. e0195065, abr. 2018, doi: 10.1371/journal.pone.0195065.
[56] L. A. Castro *et al.*, «Using heterogeneous data to identify signatures of dengue outbreaks at fine spatio-temporal scales across Brazil», *PLoS Negl. Trop. Dis.*, vol. 15, n.º 5, p.

e0009392, may 2021, doi: 10.1371/journal.pntd.0009392.
[57] K. Roster, C. Connaughton, y F. A. Rodrigues, «Machine-Learning–Based Forecasting of Dengue Fever in Brazilian Cities Using Epidemiologic and Meteorological Variables», *Am. J. Epidemiol.*, vol. 191, n.º 10, pp. 1803-1812, sep. 2022, doi: 10.1093/aje/kwac090.
[58] F. Cortes *et al.*, «Time series analysis of dengue surveillance data in two Brazilian cities», *Acta Trop.*, vol. 182, pp. 190-197, jun. 2018, doi: 10.1016/j.actatropica.2018.03.006.
[59] B. C. Bohm *et al.*, «Utilization of machine learning for dengue case screening», *BMC Public Health*, vol. 24, n.º 1, p. 1573, jun. 2024, doi: 10.1186/s12889-024-19083-8.
[60] M. A. Johansson, N. G. Reich, A. Hota, J. S. Brownstein, y M. Santillana, «Evaluating the performance of infectious disease forecasts: A comparison of climate-driven and seasonal dengue forecasts for Mexico», *Sci. Rep.*, vol. 6, n.º 1, p. 33707, sep. 2016, doi: 10.1038/srep33707.
[61] P. M. Luz, B. V. M. Mendes, C. T. Codeço, C. J. Struchiner, y A. P. Galvani, «Time series analysis of dengue incidence in Rio de Janeiro, Brazil», *Am. J. Trop. Med. Hyg.*, vol. 79, n.º 6, pp. 933-939, dic. 2008.
[62] «Application of Artificial Neural Networks for Dengue Fever Outbreak Predictions in the Northwest Coast of Yucatan, Mexico and San Juan, Puerto Rico». Accedido: 24 de noviembre de 2024. [En línea]. Disponible en: https://www.mdpi.com/2414-6366/3/1/5
[63] N. Zhao *et al.*, «Machine learning and dengue forecasting: Comparing random forests and artificial neural networks for predicting dengue burden at national and sub-national scales in Colombia», *PLoS Negl. Trop. Dis.*, vol. 14, n.º 9, p. e0008056, sep. 2020, doi: 10.1371/journal.pntd.0008056.
[64] «Intra- and interseasonal autoregressive prediction of dengue outbreaks using local weather and regional climate for a tropical environment in Colombia - PubMed». Accedido: 24 de noviembre de 2024. [En línea]. Disponible en: https://pubmed.ncbi.nlm.nih.gov/24957546/
[65] «Bayesian dynamic modeling of time series of dengue disease case counts | PLOS Neglected Tropical Diseases». Accedido: 24 de noviembre de 2024. [En línea]. Disponible en: https://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0005696
[66] J. Gomez, J. Prieto, E. Leon, y A. Rodríguez, «INFEKTA—An agent-based model for transmission of infectious diseases: The COVID-19 case in Bogotá, Colombia», *PLOS ONE*, vol. 16, n.º 2, p. e0245787, feb. 2021, doi: 10.1371/journal.pone.0245787.
[67] *Acquiring and Processing Time Series Data*. Accedido: 25 de noviembre de 2025. [En línea]. Disponible en: https://learning.oreilly.com/library/view/modern-time-series/9781835883181/Text/Chapter_02.xhtml
[68] «Análisis de indicadores para la vigilancia». Accedido: 25 de noviembre de 2025. [En línea]. Disponible en: https://www.ins.gov.co/BibliotecaDigital/analisis-de-indicadores-para-la-vigilancia.pdf
[69] «Calendario Climático 2025-2026». Accedido: 25 de noviembre de 2025. [En línea]. Disponible en: https://portal.gestiondelriesgo.gov.co/Paginas/Slide_home/Calendario-Climatico-2025-2026.aspx
[70] «Splitting the Data for Training and Testing», O'Reilly Online Learning. Accedido: 25 de noviembre de 2025. [En línea]. Disponible en: https://learning.oreilly.com/videos/python-fundamentals-with/9780135917411/9780135917411-PFLL_Lesson14_11/

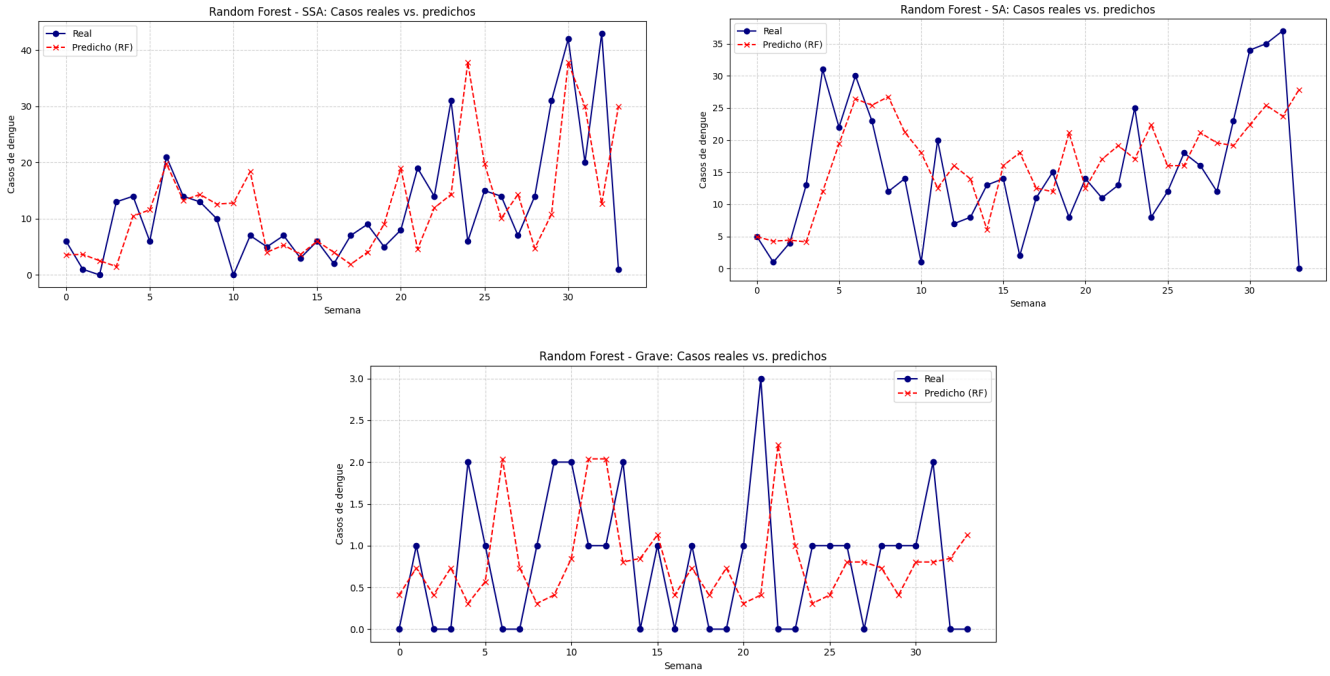
[71] *CHAPTER 4: Introduction to Autoregressive and Automated Methods for Time Series Forecasting*. Accedido: 25 de noviembre de 2025. [En línea]. Disponible en: https://learning.oreilly.com/library/view/machine-learning-for/9781119682363/c04.xhtml
[72] *2. End-to-End Machine Learning Project*. Accedido: 25 de noviembre de 2025. [En línea]. Disponible en: https://learning.oreilly.com/library/view/hands-on-machine-learning/9798341607972/ch02.html
[73] *12 Introducing deep learning for time series forecasting*. Accedido: 25 de noviembre de 2025. [En línea]. Disponible en: https://learning.oreilly.com/library/view/time-series-forecasting/9781617299889/Text/12.htm
[74] *2. The What, Where, When, and How of Data Processing*. Accedido: 25 de noviembre de 2025. [En línea]. Disponible en: https://learning.oreilly.com/library/view/streaming-systems/9781491983867/ch02.html
[75] I. N. Tanawi, V. Vito, D. Sarwinda, H. Tasman, y G. F. Hertono, «Support Vector Regression for Predicting the Number of Dengue Incidents in DKI Jakarta», *Procedia Comput. Sci.*, vol. 179, pp. 747-753, ene. 2021, doi: 10.1016/j.procs.2021.01.063.
[76] *CHAPTER 14: Time Series*. Accedido: 6 de enero de 2026. [En línea]. Disponible en: https://learning.oreilly.com/library/view/applied-machine-learning/9781394155378/c14.xhtml
[77] *15. Processing Sequences Using RNNs and CNNs*. Accedido: 6 de enero de 2026. [En línea]. Disponible en: https://learning.oreilly.com/library/view/hands-on-machine-learning/9781098125967/ch15.html
[78] *2. Supervised Learning*. Accedido: 25 de noviembre de 2025. [En línea]. Disponible en: https://learning.oreilly.com/library/view/introduction-to-machine/9781449369880/ch02.html
[79] *5.2 Some simple forecasting methods | Forecasting: Principles and Practice (3rd ed)*. Accedido: 7 de enero de 2026. [En línea]. Disponible en: https://otexts.com/fpp3/simple-methods.html
[80] «Vigilancia». Accedido: 15 de febrero de 2026. [En línea]. Disponible en: https://www.ins.gov.co/Direcciones/Vigilancia/Paginas/SIVIGILA.aspx
[81] S. A. Inam, «A review of artificial intelligence for predicting climate driven infectious disease outbreaks to enhance global health resilience», *Discov. Public Health*, vol. 22, n.º 1, p. 738, nov. 2025, doi: 10.1186/s12982-025-01167-4.
[82] A. Z. Al Meslamani, I. Sobrino, y J. de la Fuente, «Machine learning in infectious diseases: potential applications and limitations», *Ann. Med.*, vol. 56, n.º 1, p. 2362869, dic. 2024, doi: 10.1080/07853890.2024.2362869.
[83] A. J. R. Reyes, Y. S. Lizarazo, y L. C. P. Herrera, «INFORME DE EVENTO DENGUE, COLOMBIA 2020», n.º 04, 2019.

9 Anexos

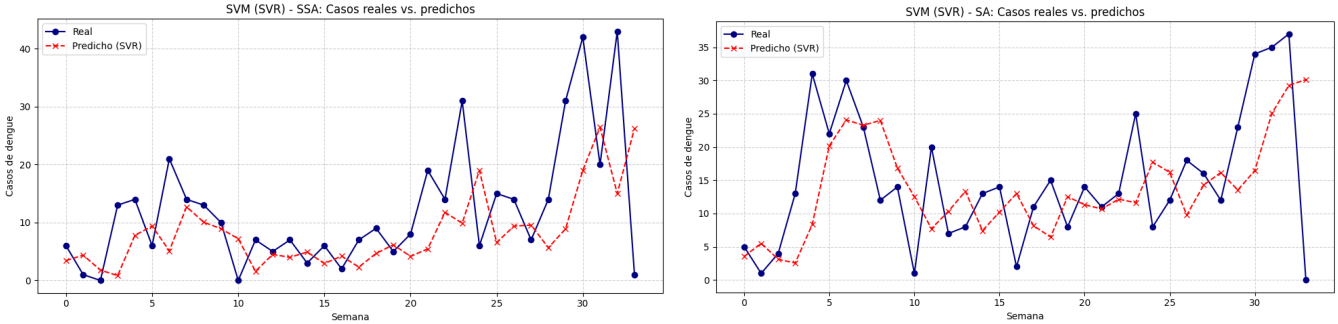
Anexo 1. Prototipo para la utilización del modelo

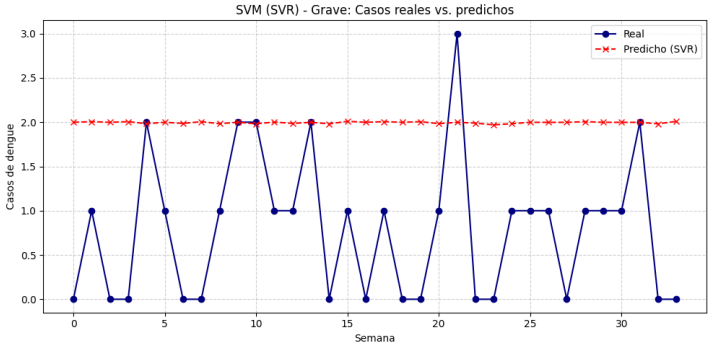
El prototipo del modelo predictivo se encuentra disponible en este link: https://incidencia-dengue-app-fylb5xhwq8zmjzrguqxg2n.streamlit.app/

Anexo 2. Graficas comparativas entre el valor predicho vs el valor final. Modelo Random forest optimizado



Anexo 3. Graficas comparativas entre el valor predicho vs el valor final. Modelo SVM optimizado





Anexo 4. Permiso de la institución para realizar el estudio

Hospital Infantil Napoleón Franco Pareja, Dirección Científica Cartagena de Indias, D. T. y C. 16 de enero de 2026

A QUIEN CORRESPONDA:

En mi calidad de **Director Científico** del Hospital Infantil Napoleón Franco Pareja (Casa del Niño), por medio de la presente hago constar que la institución otorga el aval y la autorización formal al estudiante **JOEL DORIA ATENCIA**, identificado con cédula de ciudadanía No. **1.193.034.755**, para el uso y tratamiento de los datos históricos de incidencia de Dengue correspondientes al periodo comprendido entre los años **2019 y 2022**.

Esta autorización se concede con el fin exclusivo de desarrollar su trabajo de investigación para optar al título de **Magíster en Ciencia de Datos**, titulado:

"MODELO DE APRENDIZAJE AUTOMÁTICO PARA LA PREDICCIÓN DEL COMPORTAMIENTO EPIDEMIOLÓGICO DEL DENGUE EN UN HOSPITAL PEDIÁTRICO DE CARTAGENA DE INDIAS"

El uso de dicha base de datos queda sujeto a las siguientes consideraciones:

- **Confidencialidad:** El investigador se compromete a garantizar el anonimato de los pacientes, asegurando que no se divulgará información que permita su identificación, cumpliendo con la Ley 1581 de 2012 (Habeas Data).
- **Uso Académico:** Los datos serán utilizados estrictamente para fines de investigación, modelado predictivo y análisis estadístico dentro del marco de la tesis mencionada.
- **Ética:** El proyecto deberá sigue los lineamientos del Comité de Ética de la institución el manejo responsable de la información clínica.

Reconocemos la importancia de este estudio para fortalecer la capacidad de respuesta de nuestra institución y de la ciudad ante los brotes epidemiológicos de Dengue, mediante el uso de tecnologías avanzadas de *Machine Learning*.

Atentamente,

Hernando Pinzón Redondo

Director Científico

Hospital Infantil Napoleón Franco Pareja Cartagena de Indias, Colombia