ISSN: 1692-7257 - Volumen 1 – Número 43 - 2024

RCTA
Revista Colombiana de Tecnologías de Avanzada
UNIPAMPLONA

Digital Object Identifier: 10.24054/rcta.v1i43.2822

Aplicación de machine learning y metodología CRISP-DM para la clasificación precisa de severidad en casos de dengue

Application of machine learning and CRISP-DM methodology for accurate severity classification of dengue

MSc. Carlos Alberto Mejía Rodríguez [1], MSc. Miguel Alberto Rincón Pinzón [1]
MSc. Luís Palmera Quintero [1], Esp. Lina Marcela Arévalo Vergel [1]

[1] ***Universidad Popular del Cesar,*** *Ingeniería de sistemas, Grupo de Investigación GIDEATIC, Aguachica, César, Colombia.*

Correspondencia: *calbertomejia@unicesar.edu.co*

Recibido: *15 octubre 2023.* ***Aceptado:*** *17 diciembre 2023.* ***Publicado:*** *16 marzo 2024.*

Cómo citar: *C. A. Mejia Rodriguez, M. A. Rincon Pinzon, L. M. Palmera Quintero, y L. M. Arevalo Vergel, «Aplicación de machine learning y metodología CRISP-DM para la clasificación precisa de severidad en casos de dengue», RCTA, vol. 1, n.º 43, pp. 78–85, mar. 2024.*
Recuperado de https://ojs.unipamplona.edu.co/index.php/rcta/article/view/2822

Esta obra está bajo una licencia internacional
Creative Commons Atribución-NoComercial 4.0.

Resumen: El proyecto se centra en clasificar con precisión la severidad de los casos de Dengue en Casanare, Colombia, utilizando Machine Learning (ML) y la metodología CRISP-DM. La variable objetivo es “clasificación final”, que categoriza los casos en dengue sin signos de alarma y con signos de alarma. Se probaron varios modelos y técnicas, destacando 'RandomForest' como el más efectivo debido a su alto rendimiento, alcanzando una precisión del 100%. La mejora en la clasificación permitirá una identificación temprana y precisa de la gravedad de los casos, lo que, a su vez, puede mejorar la atención médica y las estrategias de intervención. Se utilizo la base de datos “Casos de Dengue en Casanare por servicio hospitalario, relación tipo de persona, síntomas y estado hospitalario” para respaldar el análisis.

Palabras clave: Ciencia de Datos, CRISP-DM, Dengue, Machine Learning.

Abstract: The project focuses on accurately classifying the severity of Dengue cases in Casanare, Colombia, using Machine Learning (ML) and the CRISP-DM methodology. The target variable is "final classification," which categorizes cases into Dengue without warning signs and Dengue with warning signs. Several models and techniques were tested, with 'RandomForest' standing out as the most effective due to its high performance, achieving an accuracy of 100%. This improvement in classification will enable early and precise identification of case severity, which, in turn, can enhance medical care and intervention strategies. The "Dengue Cases in Casanare by hospital service, person type, symptoms, and hospital status" database was used to support the analysis.

Keywords: CRISP-DM, Data Science, Dengue, Machine Learning.

1. INTRODUCCIÓN

La captura de datos en la actualidad es más masiva que nunca, sin embargo, no se debería limitar su aplicación al simple almacenamiento. Como destaca [1], "En la medida en que nuestras organizaciones reconozcan el gran valor que tienen los datos, seremos testigos de muchas más implementaciones". Este reconocimiento ha dado lugar a un concepto revolucionario conocido como Big Data. Para [2] el término Big Data, a menudo traducido como datos masivos, surgió en el inicio del siglo XXI, especialmente en campos científicos como la astronomía y la genética, impulsado por la explosión en la disponibilidad de datos. Ejemplos notables incluyen el proyecto Sloan Digital Sky Survey, que generó más datos en meses que en toda la historia hasta el momento de la astronomía, y el proyecto del genoma humano, que produjo enormes cantidades de datos genéticos. Sin embargo, en los últimos años, la masificación de datos se ha extendido a todos los ámbitos con el aumento de dispositivos conectados a Internet, el auge de las redes sociales y el Internet de las cosas (IoT). Además, muchos de estos datos son accesibles de manera abierta, lo que permite su explotación global. Pero a pesar de la abundancia de datos, su verdadero valor reside en su análisis e interpretación.

Dentro del ámbito del Big Data, es fundamental comprender los conceptos de dato, información y conocimiento. Según [3], el dato es la unidad más elemental y cruda que se puede analizar. La información es el resultado de operaciones realizadas en los datos, perdiendo ciertos detalles, pero ganando generalidad. Finalmente, el conocimiento es una abstracción que guía la transformación de datos en información, considerando el contexto y equilibrando aplicabilidad y comprensión de matices para su interpretación precisa. Una manera general de entender la relación entre estos elementos es reconocer su adaptabilidad a distintos usuarios, quienes, debido a sus perspectivas individuales, pueden situarlos en niveles de abstracción diferentes. Para un usuario, lo que constituye información puede ser considerado como datos por otro. Entonces dependiendo del enfoque y del nivel de agregación, la entidad que se considera dato, información o conocimiento varía

Para extraer conocimiento a partir de grandes cantidades de información, es necesario llevar a cabo un proceso de minería de datos. De acuerdo con [4] la minería de datos se centra en aprovechar grandes cantidades de información. No obstante, para trascender la mera etapa de análisis de datos, se han diseñado metodologías específicas con el objetivo de mejorar y optimizar el proceso de extracción de conocimiento. Estas metodologías ofrecen un marco estructurado para obtener, depurar y aplicar el conocimiento adquirido de manera efectiva en entornos organizacionales. El modelo KDD (Knowledge Discovery in Databases) y la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) son enfoques fundamentales en el campo de la minería de datos. KDD establece etapas clave para proyectos exitosos, incluyendo selección, preparación, búsqueda de patrones, evaluación y refinamiento de modelos. CRISP-DM, por otro lado, se basa en KDD, pero se adapta específicamente a las necesidades industriales, organizando el proceso en fases, procesos y actividades, desde la comprensión del negocio y la preparación de datos hasta el modelado, la evaluación y la implementación de resultados.

En las etapas de modelado, los métodos de Machine Learning (ML) entran en función y, según [5] estos algoritmos pueden aprender de los datos de entrada para mejorar el rendimiento en tareas específicas a través de procesos de entrenamiento. Este campo interdisciplinario combina estadísticas e informática y se divide en aprendizaje supervisado (con ejemplos de entrada y objetivos conocidos) y no supervisado (sin información adicional). Problemas como la clasificación y la regresión son supervisados, mientras que la extracción de características y la agrupación son ejemplos de aprendizaje no supervisado.

Ahora bien, [6] destacan que, en el sector de la salud, tanto el diagnóstico como la toma de decisiones médicas conllevan un razonamiento bajo incertidumbre. Los médicos evalúan la información proporcionada por el paciente, su historial clínico y su experiencia para determinar las probabilidades de que el paciente pueda tener una determinada afección. Por lo tanto, es crucial realizar una estimación precisa de los riesgos asociados para tomar decisiones médicas adecuadas.

En cuanto a la aplicación del aprendizaje automático en investigaciones epidemiológicas, [7] enfatizan la importancia crítica que ha tenido el control de epidemias a lo largo de la historia. Como respuesta a este desafío, proponen la utilización de técnicas de ML como una herramienta novedosa para desarrollar estrategias de control óptimas que aborden múltiples tipos de intervenciones. Este

enfoque histórico y la aplicación de tecnología avanzada resaltan la urgente necesidad de abordar de manera efectiva el control de epidemias en la sociedad contemporánea.

El dengue como enfermedad viral puede suscitar desafíos de control y prevención. Según [8] el dengue representa un reto crítico en Colombia, un país hiperendémico para esta enfermedad y comprender las tendencias epidemiológicas es esencial para políticas de salud efectivas. Según [9], el dengue abarca una amplia variedad de síntomas, desde leves hasta graves. La detección temprana y el manejo adecuado son vitales para reducir la mortalidad. Para lograrlo, se pueden emplear modelos de ML como la Regresión Logística y los Árboles de Decisión, entre otros. Estos modelos se entrenan con datos de casos de dengue, que incluyen síntomas y resultados de pruebas, además se combinan o complementa las fortalezas de diferentes modelos, mejorando así la precisión de las predicciones.

2. METODOLOGÍA

La investigación se lleva a cabo siguiendo la metodología CRISP-DM, que según [10], establece las etapas necesarias para la realización de proyectos de minería de datos. Es una guía de acceso gratuito basada en el proceso KDD. La metodología CRISP-DM se compone de seis fases principales: comprensión del negocio, comprensión de los datos, preparación de datos, modelado, evaluación e implementación. Un esquema del flujo de las fases. Ver fig. 1.

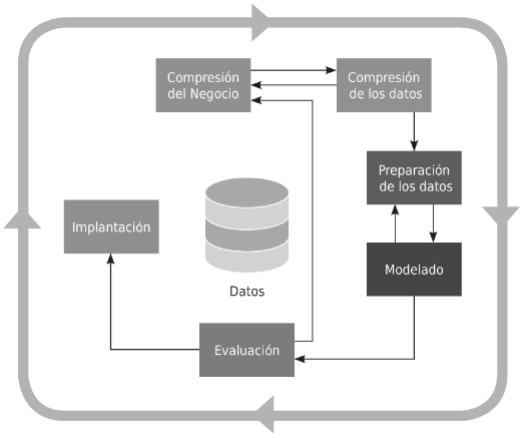

Fig. 1. *Flujo CRISP-DM*
Fuente: *Castillo Romero, J. A. (2019). Big data. IFCT128PO. IC Editorial.*

El estudio se basa en el conjunto de datos "Casos de Dengue en Casanare por servicio hospitalario," descargado del portal Datos Abiertos de Colombia. Este dataset consta de 54 columnas y 2010 filas, cada una representando un caso de dengue atendido en centros hospitalarios. Los datos fueron recopilados entre octubre de 2022 y febrero de 2023. La variable objetivo, denominada "clasfina" o "Clasificación Final," se utiliza para determinar la gravedad del dengue en pacientes y se divide en dos categorías: "Dengue sin signos de alarma" y "Dengue con signos de alarma". Esta variable es esencial, ya que es la que se pretende que los modelos sean capaces de aprender a predecir. El objetivo es desarrollar el mejor modelo de Machine Learning para clasificar la gravedad de la enfermedad en nuevos pacientes.

3. RESULTADOS

El trabajo se organiza siguiendo las fases y actividades de la metodología CRISP-DM. A continuación, se describen los resultados de cada fase.

3.1. Comprensión del negocio

3.1.1. Determinar los objetivos de la Organización

Según [11] el dengue es una infección viral transmitida por mosquitos, común en áreas tropicales y subtropicales. En muchos casos, las personas infectadas no presentan síntomas, pero cuando estos surgen, incluyen fiebre alta, dolores de cabeza, náuseas y erupciones cutáneas. La mayoría se recupera en una o dos semanas, pero en ocasiones, la enfermedad se agrava y requiere hospitalización, incluso puede ser mortal. Aunque existen medicamentos para aliviar los síntomas del dengue, actualmente no hay un tratamiento específico.

La información sobre la entidad suministradora de los datos se presenta en la tabla 1.

Tabla 1: *Información de la Entidad*

Área/dependencia	Secretaría de Salud y Ambiente Municipal
Nombre/Entidad	E.S.E. Salud Yopal
Departamento	Casanare
Municipio	Yopal
Orden	Territorial
Sector	Salud y Protección Social

3.1.2. Determinación de los objetivos del proyecto

Elaborar un modelo de predicción del diagnóstico final de la gravedad de dengue en pacientes

atendidos en hospitales en base a los datos históricos de atención de estos. Se establece que la efectividad mínima del modelo de predicción debe ser del 99%.

3.1.3. Planificar tareas.

En esta fase, lo recomendado es adoptar enfoques ágiles de desarrollo, se recomienda especialmente Scrum, pues el objetivo principal es desarrollar modelos de ML, que están relacionados con el desarrollo de algoritmos de computadora. Y como indica [12] Scrum es un marco de trabajo ampliamente reconocido en la industria del desarrollo de software. Esto implica formar equipos, organizar el trabajo en tareas generales significativas y desglosarlas en actividades detalladas.

3.2. Comprensión de los datos

El conjunto de datos corresponde a la relación de casos de Dengue ocurridos en el municipio de Yopal, departamento de Casanare en Colombia, desagregado por tipo de persona, síntomas y servicios hospitalarios, cada fila es un caso atendido. El detalle del dataset se presenta en la tabla 2.

Tabla 2: *Información de Datos*

Idioma	Español
Cobertura Geográfica	Municipal – Yopal
Fecha Emisión	2022-10-24
Última actualización	2023-02-17
Filas	2.010
Columnas	54
URL Normativa	Ir.

3.2.1. Exploración de datos

Lo ideal en este momento sería presentar una descripción detallada de cada variable o columna del conjunto de datos. Sin embargo, por razones de practicidad en el informe, solo se proporciona el detalle de la variable objetivo en la Tabla 3.

Tabla 3: *Detalle variable objetivo*

Nombre Atributo	Clasfinal
Descripción	Clasificación de pacientes según síntomas y protocolo de seguimiento
Tipo de datos	Categórica/ String
Distribución de valores	Destinitos: 3 Nulos: 0 Valores: Dengue con síntomas de alarma: 1020 (50.7%). Dengue sin síntomas de alarma: 988 (49.3%)

En este apartado es relevante un resumen estadístico de las variables numéricas. Un ejemplo concreto de este tipo de resumen se encuentra en la Tabla 4, donde se detallan estadísticas descriptivas de la variable "edad_"

Tabla 4: *Descripción estadística de la variable "edad"*

count	2010
mean	19.736816
std	18.016333
min	1
25%	7
50%	13
75%	27
max	101

Los resultados de la Tabla 4 indican presencia de valores atípicos, esto también se puede apreciar en la distribución de datos ver fig.2.

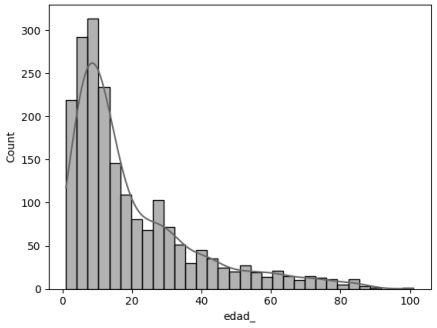

Fig. 2. *Distribución de la edad de los pacientes.*

El resumen de la variable indica que los casos se concentran significativa en pacientes menores de 13 años, lo cual sugiere una marcada tendencia de la enfermedad en los niños. Estos resultados apoyan la afirmación realizada por [13], la cual sostiene que el dengue hemorrágico (DHF) y el síndrome de choque por dengue (DSS), dos variantes graves de la enfermedad, tienen una mayor incidencia en niños menores de 15 años.

Este análisis demográfico, que destaca la vulnerabilidad de los niños al dengue grave, tiene aplicaciones más allá de la investigación médica. En el contexto del aprendizaje automático, podría utilizarse como un indicador para establecer clústeres en datos no etiquetados.

3.3. Preparación de los datos

Según [4] "Los datos, en el mundo real, suelen estar incompletos y tener inconsistencias. Por esto es necesario prepararlos antes de ejecutar modelos de analítica".

Una problemática frecuente en proyectos Big data son los valores nulos y atípicos, sobre estos últimos [14] enfatizan en que la presencia de valores atípicos (outliers) en los datos puede tener un impacto significativo en la evaluación de modelos de ML, influenciando la percepción del rendimiento del modelo y la interpretación de las características importantes.

Para [15] es crucial comprender que, en entornos con un alto número de dimensiones, algunos algoritmos no funcionan de manera eficaz. La reducción de dimensionalidad aborda este desafío al convertir un conjunto de datos con múltiples dimensiones (variables o atributos) en datos con dimensiones menores, garantizando al mismo tiempo que se conserve información relevante de manera concisa. Esta técnica desempeña un papel fundamental en simplificar conjuntos de datos complejos y es ampliamente aplicada en el ámbito del aprendizaje automático. Hay que tener en cuenta que existen distinciones entre la selección de características y la reducción de dimensionalidad, aunque ambos enfoques tienen como objetivo primordial mejorar tanto la eficiencia como la precisión en el análisis de datos.

Preparar los datos implica asegurar su calidad, y para lograrlo, es esencial llevar a cabo la limpieza de datos, que según [16] "En el proceso de limpieza de datos (en inglés, data cleansing o data scrubbing) se llevan a cabo actividades de detección, eliminación o corrección de instancias corrompidas o inapropiadas en los juegos de datos".

Para llevar a cabo estas actividades, es fundamental revisar y establecer las reglas de negocio en colaboración con los diversos actores involucrados, como proveedores de servicios y usuarios. A través de esta colaboración, se pueden definir rangos o formatos de datos válidos. En algunos casos, la documentación puede ser el único medio necesario para establecer estas reglas, o bien, puede complementar el proceso.

En esta fase de preparación de datos, se toman medidas para mejorar la calidad y utilidad de los mismos. Se identifican y eliminan las variables que no aportan información significativa, como aquellas en las que más del 95% de sus valores son idénticos. Los valores faltantes se reemplazan utilizando la moda para variables categóricas y la media para variables numéricas. Además, se transforma la fecha de ingreso en el número de mes para considerar posibles patrones estacionales. Se calculan los días con síntomas de un paciente en atención restando la fecha de inicio de síntomas de la fecha de ingreso.

Siguiendo con la reducción de dimensiones, se utiliza un algoritmo de Eliminación Hacia Atrás, también conocida como Backward Elimination, técnica que, según [17], tiene como propósito principal la construcción de un modelo de regresión múltiple de alta calidad que tenga la menor cantidad de atributos posible, pero sin sacrificar la capacidad predictiva del modelo.

Al concluir esta fase, el conjunto de datos queda definido de la siguiente manera: 1994 filas y 12 columnas. La Tabla 3 presenta las variables finales que serán empleadas en el modelado.

Tabla 5: *Variables clasificadas para el modelado*

Columna	Descripción	Tipo
Semana	Semana epidemiológica según el calendario vigente	Número
sexo_	Sexo del paciente	Texto
ciclo_de_vida	Etapa de vida del paciente	Texto
Dolrretroo	Hallazgos semiológicos Dolor Retro ocular	Texto
Malgias	Hallazgos semiológicos mialgias	Texto
Artralgia	Hallazgos semiológicos Artralgias	Texto
dolor_abdo	Hallazgos semiológicos Dolor abdominal	Texto
Vomito	Hallazgos semiológicos vomito	Texto
pac_hos_	Se encuentra el paciente Hospitalizado	Texto
Conducta	Conducta del paciente	Texto
dias_sint	Días con síntomas del paciente	Número
Clasfinal	Clasificación final	Texto

3.4. Modelado

Según [18], el modelado en Machine Learning utiliza datos históricos etiquetados para entrenar modelos que pueden hacer predicciones precisas en nuevos datos sin etiquetas. Comienza con la recopilación de datos, seguida de la división en conjuntos de entrenamiento y prueba. El conjunto de entrenamiento se utiliza para enseñar al modelo a identificar patrones y relaciones entre las características y los resultados. La evaluación se realiza con el conjunto de prueba para medir la precisión del modelo. Si es necesario, se ajustan parámetros o se prueban diferentes algoritmos. Una vez que se considera efectivo, el modelo se utiliza para hacer predicciones en nuevos datos.

Siguiendo esta dirección, se empleen bibliotecas implementadas en Python que proporcionan diversas técnicas de aprendizaje automático para procesar datos y generar modelos. El código correspondiente a estas actividades se encuentra compilado en cuadernos de Jupyter disponibles en el repositorio de GitHub. La experimentación incluirá la realización de pruebas con diversos algoritmos y la optimización de hiperparámetros. Los resultados de estas evaluaciones se presentarán de manera resumida en las secciones siguientes.

Los algoritmos utilizados son LogisticRegression, KNeighborsClassifier y RandomForestClassifier.

La Regresión Logística, para [19], es un valioso enfoque analítico para resolver problemas de clasificación, como determinar si una nueva muestra se relaciona mejor con una categoría específica. El modelo de regresión logística más común aborda un resultado binario; algo que puede tomar dos valores.

Este enfoque se puede aplicar al contexto de la clasificación de la gravedad del dengue en pacientes atendidos en centros de salud, permitiendo diferenciar entre casos de Dengue con síntomas de alarma y casos de Dengue sin síntomas de alarma.

KNeighborsClassifier, según [20] es un clasificador que se basa en la evidencia proporcionada por las instancias de aprendizaje cercanas al patrón que se está clasificando. Los parámetros del método se ajustan de manera que se maximice la verosimilitud evidencial, la cual es una extensión de la función de verosimilitud diseñada para manejar datos inciertos. Este clasificador ha demostrado superar a otros métodos en situaciones de aprendizaje parcialmente supervisado.

Se utilizo el conjunto de entrenamiento para entrenar el modelo KNeighborsClassifier. Este modelo se ajusta a los datos de entrenamiento y aprende a identificar patrones que relacionan las características de los pacientes con la gravedad del Dengue.
RandomForestClassifier, Según [21] se utiliza para intentar mejorar la precisión en comparación con la regresión lineal, ya que el bosque aleatorio puede aproximar mejor la relación entre los objetivos y las características.

A continuación, se presentan los resultados sobresalientes de cada técnica junto con sus respectivas combinaciones de parámetros.

Un primer paso se propone utilizar "Label Encoding" para convertir las variables categóricas en formato numérico, facilitando así su procesamiento por los algoritmos de machine learning. La variable objetivo se codifica como 0 para Dengue con síntomas de alarma y 1 para Dengue sin síntomas de alarma.

3.4.1. LogisticRegression

Parametros: C=1.0, class_weight=None, dual=False, fit_intercept=True, intercept_scaling=1, l1_ratio=None, max_iter=1000, multi_class='auto', n_jobs=None, penalty='l2', random_state=None, solver='lbfgs', tol=0.0001, verbose=0, warm_start=False.
Rendimiento obtenido: 0.9038076152304609

En la fig. 3 se presenta la matriz de confusión que resume los resultados del modelo usando la técnica de LogisticRegression en términos de aciertos y desaciertos en la clasificación.

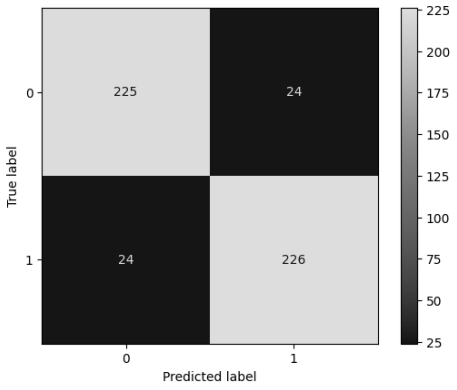

Fig. 3. *Matriz de confusión del modelo LogisticRegression*

3.4.2. KNeighborsClassifier

Parametros: algorithm='kd_tree', leaf_size=30, metric='minkowski', metric_params=None, n_jobs=None, n_neighbors=10, p=2, weights='uniform
Rendimiento obtenido: 0.7414829659318637

En la figura 4 se muestra la matriz de confusión que resume el rendimiento del modelo utilizando la técnica de KNeighborsClassifier.

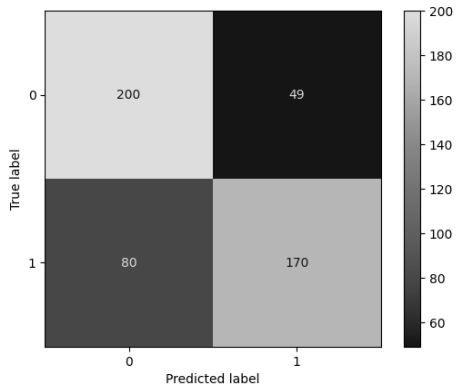

Fig. 4. *Matriz de confusión del modelo KNeighborsClassifier*

3.4.3. RandomForestClassifier

Parametros: bootstrap=True, ccp_alpha=0.0, class_weight=None, criterion='gini', max_depth=None, max_features='sqrt', max_leaf_nodes=None, max_samples=None, min_impurity_decrease=0.0, min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0, n_estimators=100, n_jobs=None, oob_score=False, random_state=42, verbose=0, warm_start=False
Rendimiento obtenido: 1.0

En la fig. 5 se muestra la matriz de confusión que resume el rendimiento del modelo utilizando la técnica de KNeighborsClassifier.

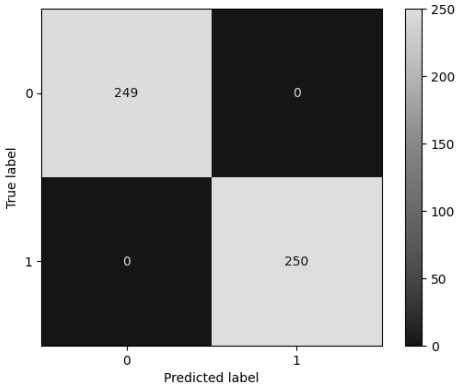

Fig. 5. *Matriz de confusión del modelo RandomForestClassifier*

Durante el proceso de prueba de modelos, se emplearon varios modelos y se aplicaron configuraciones específicas para mejorar su rendimiento. Se ajustaron hiperparámetros clave para cada modelo con el objetivo de encontrar combinaciones óptimas.

3.5. Evaluación

Tras experimentar con los diferentes modelos y probar diferentes ajustes en los hiperparámetros, es la técnica RandomForestClassifier la que se adapta mejor a las características de los datos y con el que se obtiene una eficiencia del 100% en la clasificación de la complejidad del dengue.

3.6. Implementación

Sería ideal evaluar el rendimiento del modelo con nuevos datos de la entidad suministradora, pero estos no se actualicen frecuentemente en el portal donde se disponen abiertamente. Otra alternativa de implementación es integrar el modelo con la plataforma de registro de los datos para generar automáticamente clasificaciones sobre la gravedad del dengue. Esto proporcionaría al profesional la opción de utilizar estas predicciones como una herramienta de apoyo al emitir el diagnóstico final, ya sea dengue con síntomas de alarma o sin síntomas de alarma.

4. CONCLUSIONES

Se aplicó la metodología CRISP-DM para lograr una eficiente clasificación de la gravedad del dengue en pacientes de Casanare, Colombia, utilizando técnicas de aprendizaje automático como LogisticRegression, KNeighborsClassifier y RandomForestClassifier.

Los resultados evidencian que el modelo RandomForestClassifier logró una tasa de clasificación correcta del 100% en la gravedad del dengue de los pacientes, lo que podría ser fundamental para el tratamiento temprano de la enfermedad y tener un impacto significativo en la salud pública de la región.

Para investigaciones futuras, se sugiere la integración de estos modelos en la plataforma de registro de datos para generar predicciones automáticas, brindando a los profesionales de la salud una herramienta de apoyo en los diagnósticos.

Se logra comprobar la efectividad de la minería de datos y el aprendizaje automático como herramientas predictivas para el manejo del dengue, con importantes implicaciones para la atención médica y la prevención de la enfermedad en la región.

REFERENCIAS

[1] Medina L., E. H. Big Data: Los Datos como Generadores de Valor. Universidad Peruana de Ciencias Aplicadas. 2023.

[2] Casas R., J., Nin G., J., & Julbe L., F. (2019). Big data: análisis de datos en entornos masivos. Editorial UOC.

[3] López M., J. J. y Zarza, G. (2017). La ingeniería del big data: cómo trabajar con datos. Editorial UOC. Barcelona, España.

[4] Maldonado, S. (2022). Analytics y Big Data: ciencia de los Datos aplicada al mundo de los negocios. RIL editores.

[5] Suarez L, A. A., Vazquez S., C. R., & Huffel, S. Van. (2018). Machine learning approaches for ambulatory electrocardiography signal processing.

[6] Rios Insua, D., & Gomez-Ullate Oteiza, D. (2019). Big data: conceptos, tecnologias y aplicaciones. Editorial CSIC Consejo Superior de Investigaciones Cientificas.

[7] Arnst, M., Louppe, G., Van Hulle, R., Gillet, L., Bureau, F., & Denoël, V. (2022). A hybrid stochastic model and its Bayesian identification for infectious disease screening in a university campus with application to massive COVID-19 screening at the University of Liège. Mathematical Biosciences, 347. https://doi.org/10.1016/j.mbs.2022.108805

[8] Gutierrez-Barbosa, H., Medina-Moreno, S., Zapata, J. C., & Chua, J. V. (2020). Dengue Infections in Colombia: Epidemiological Trends of a Hyperendemic Country. Tropical Medicine and Infectious Disease, 5(4).

[9] Gangula, R., Thirupathi, L., Parupati, R., Sreeveda, K., & Gattoju, S. (2023). Ensemble machine learning based prediction of dengue disease with performance and accuracy elevation patterns. Materials Today: Proceedings, 80, 3458–3463. https://doi.org/https://doi.org/10.1016/j.matpr.2021.07.270

[10] Castillo Romero, J. A. (2019). Big data. IFCT128PO. IC Editorial.

[11] Organización Mundial de La Salud. (2023). Dengue y dengue grave. WHO.

[12] Kadenic, M. D., Koumaditis, K., & Junker-Jensen, L. (2023). Mastering scrum with a focus on team maturity and key components of scrum. Information and Software Technology, 153, 107079. https://doi.org/https://doi.org/10.1016/j.infsof.2022.107079

[13] Treatments for dengue: a Global Dengue Alliance to address unmet needs. (2023). The Lancet Global Health. https://doi.org/https://doi.org/10.1016/S2214-109X(23)00362-5

[14] Nariya, M. K., Mills, C. E., Sorger, P. K., & Sokolov, A. (2023). Paired evaluation of machine-learning models characterizes effects of confounders and outliers. Patterns, 4(8), 100791. https://doi.org/https://doi.org/10.1016/j.patter.2023.100791-

[15] Menoyo R., D., Garcia L., E., & Garcia C., A. (2021). Fundamentos de la ciencia de datos. Editorial Universidad de Alcala.

[16] Minguillon, J., Casas, J., & Minguillon, J. (2017). Mineria de datos: modelos y algoritmos. Editorial UOC.

[17] Kotu, V., & Deshpande, B. (2019). Chapter 14 - Feature Selection. In V. Kotu & B. Deshpande (Eds.), Data Science (Second Edition) (pp. 467–490). Morgan Kaufmann. https://doi.org/https://doi.org/10.1016/B978-0-12-814761-0.00014-9

[18] Caballero, R., & Martin, E. (2022). Las bases de big data y de la inteligencia artificial. Los libros de la Catarata.

[19] Edgar, T. W., & Manz, D. O. (2017). Chapter 4 - Exploratory Study. In T. W. Edgar & D. O. Manz (Eds.), Research Methods for Cyber Security (pp. 95–130). Syngress. https://doi.org/https://doi.org/10.1016/B978-0-12-805349-2.00004-2

[20] Denœux, T., Kanjanatarakul, O., & Sriboonchitta, S. (2019). A new evidential K-nearest neighbor rule based on contextual discounting with partially supervised learning. International Journal of Approximate Reasoning, 113, 287–302. https://doi.org/https://doi.org/10.1016/j.ijar.2019.07.009

[21] Malik, A., Javeri, Y. T., Shah, M., & Mangrulkar, R. (2022). Chapter 11 - Impact analysis of COVID-19 news headlines on global economy. In R. C. Poonia, B. Agarwal, S. Kumar, M. S. Khan, G. Marques, & J. Nayak (Eds.), Cyber-Physical Systems (pp. 189–206). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-824557-6.00001-7