Jhon Vilcarana Tintaya

Documento.pdf

My Files

My Files

Universidad Peruana Union

Detalles del documento

Identificador de la entrega
trn:oid:::29566:543848416

Fecha de entrega
1 ene 2026, 9:53 a.m. GMT-5

Fecha de descarga
1 ene 2026, 9:55 a.m. GMT-5

Nombre del archivo
Documento.pdf

Tamaño del archivo
3.2 MB

20 páginas

13.713 palabras

76.111 caracteres

3% Similitud general

El total combinado de todas las coincidencias, incluidas las fuentes superpuestas, para ca...

Filtrado desde el informe

- Bibliografía
- Texto citado
- Texto mencionado
- Coincidencias menores (menos de 12 palabras)

Fuentes principales

2% Fuentes de Internet
1% Publicaciones
3% Trabajos entregados (trabajos del estudiante)

Marcas de integridad

N.º de alertas de integridad para revisión

No se han detectado manipulaciones de texto sospechosas.

Los algoritmos de nuestro sistema analizan un documento en profundidad para buscar inconsistencias que permitirían distinguirlo de una entrega normal. Si advertimos algo extraño, lo marcamos como una alerta para que pueda revisarlo.

Una marca de alerta no es necesariamente un indicador de problemas. Sin embargo, recomendamos que preste atención y la revise.

Fuentes principales

2% Fuentes de Internet
1% Publicaciones
3% Trabajos entregados (trabajos del estudiante)

Fuentes principales

Las fuentes con el mayor número de coincidencias dentro de la entrega. Las fuentes superpuestas no se mostrarán.

1 Trabajos entregados
University of Auckland on 2019-07-05 1%

2 Internet
export.arxiv.org <1%

3 Trabajos entregados
Infile on 2025-06-01 <1%

4 Trabajos entregados
Tilburg University on 2025-11-26 <1%

5 Internet
pci.uas.edu.mx <1%

6 Internet
aaltodoc.aalto.fi <1%

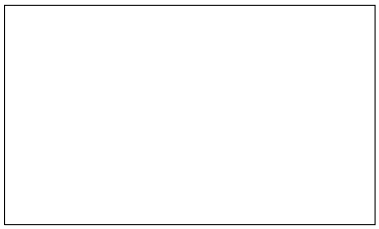
7 Internet
insiderlatam.com <1%

8 Internet
pesquisa.bvsalud.org <1%

Graphical Abstract

Fuzzy Entropy Based Classification Model for the Identification of Factors Contributing to the Prevalence of Anemia in Peruvian Children

Jhon-Vilcarana, Jorge-Quispe, Hubert-Falcon, Soria-Quijaite, Saboya-Rios



Highlights

Fuzzy Entropy Based Classification Model for the Identification of Factors Contributing to the Prevalence of Anemia in Peruvian Children

Jhon-Vilcarana, Jorge-Quispe, Hubert-Falcon, Soria-Quijaite, Saboya-Rios

- Research highlights item 1
- Research highlights item 2
- Research highlights item 3

Fuzzy Entropy Based Classification Model for the Identification of Factors Contributing to the Prevalence of Anemia in Peruvian Children[⋆,⋆⋆]

Jhon-Vilcarana[a,*,1], Jorge-Quispe[a,2], Hubert-Falcon[a,3], Soria-Quijaite[a,4] and Saboya-Rios[a,5]

[a]*Escuela de Ingeniería de Sistemas, Universidad Peruana Unión, KM 19.5 Carretera Central, Ñaña, Lima, 15011, Lurigancho Chosica, Perú*

ARTICLE INFO

Keywords:
Fuzzy Entropy
XGBoost
Classification
Hybrid Model
Public Health
Anemia

ABSTRACT

La anemia infantil constituye uno de los principales problemas de salud pública en países en desarrollo, donde las deficiencias nutricionales afectan gravemente el crecimiento y desarrollo de millones de niños. De acuerdo con la Organización Mundial de la Salud (OMS), cerca del 40% de los niños de 6 a 59 meses padecen esta condición a nivel mundial. La identificación temprana es fundamental para diseñar intervenciones efectivas; no obstante, los métodos tradicionales y los modelos de aprendizaje automático convencionales presentan limitaciones para manejar la complejidad e incertidumbre propias de los datos clínicos. En este estudio se propone un modelo de clasificación que combina lógica difusa, entropía difusa y el clasificador Extreme Gradient Boosting (XGBoost) con ajuste de umbral. El objetivo es capturar tanto la incertidumbre como las relaciones no lineales presentes en los registros médicos. Se evaluaron siete configuraciones, Random Forest y XGBoost en sus versiones básicas y ajustadas, sus respectivas extensiones con entropía difusa, y finalmente el modelo híbrido propuesto. Los resultados indican que el modelo XGBoost con entropía difusa y ajuste de umbral alcanzó el mejor desempeño, logrando un accuracy de 0.909, un F_1-score de 0.870, un ROC-AUC de 0.92 y un PR-AUC de 0.91, superando a todas las configuraciones de referencia. Estos hallazgos demuestran que la integración de representaciones difusas de la incertidumbre con técnicas avanzadas de clasificación ofrece una estrategia robusta y flexible para apoyar la toma de decisiones en salud pública y fortalecer la detección de la prevalencia de anemia infantil.

1. Introducción

La anemia infantil es uno de los mayores problemas de salud pública que afectan a los países en desarrollo, donde las deficiencias nutricionales tienen un impacto profundo en la vida de millones de niños (Nambiema, Robert & Yaya, 2019). Según la Organización Mundial de la Salud (OMS), aproximadamente el 40% de los niños de 6 a 59 meses presentan anemia, lo que corresponde a 269 millones de niños y niñas en todo el mundo. (World Health Organization, 2023). En América Latina, la anemia en niños menores de cinco años sigue siendo un problema de salud importante, y Perú no es una excepción (Rosas-Jiménez, Tercan, Horstick, Igboegwu, Dambach, Louis, Winkler & Deckert, 2022).

En la actualidad, la prevalencia de anemia infantil en menores de tres años en el Perú alcanzó un 43.7%, evidenciando un incremento de 0,6 puntos porcentuales respecto al año precedente según la Encuesta Nacional de Demografía y Salud (ENDES) 2024, afectando con mayor severidad a la población residente en áreas rurales en comparación con las urbanas; asimismo, la desnutrición crónica en menores de cinco años mostró un aumento, pasando del 11.5% en 2023 al 12.1% en 2024, lo que representa un desafío significativo para la salud pública nacional y subraya la necesidad urgente de implementar estrategias efectivas para salvaguardar el desarrollo y bienestar de la infancia peruana (Infobae, 2025). Esta condición no solo pone en riesgo el desarrollo físico y cognitivo de los niños, con consecuencias a largo plazo como bajo rendimiento dificultades académicas en la adolescencia (Hoque, Khan & Chowdhury, 2021).

Este problema es más grave en comunidades de bajos recursos, especialmente en zonas rurales de Perú, donde la falta de acceso a alimentos nutritivos y servicios de salud incrementa su prevalencia, destacando la urgencia de abordarlo de manera integral (kassab Córdova, Mendez-Guerra, Quevedo-Ramirez, Espinoza, Enriquez-Vera & Robles-Valcárcel, 2022). Los métodos tradicionales de análisis epidemiológico, que se basan principalmente en estadísticas descriptivas y análisis de regresión, no logran capturar la complejidad de las interacciones entre los múltiples factores que contribuyen a la prevalencia de anemia, estas herramientas son

[⋆] This document is the results of the research project funded by the National Science Foundation.

[⋆⋆] The second title footnote which is a longer text matter to fill through the whole text width and overflow into another line in the footnotes area of the first page.

This note has no numbers. In this work we demonstrate a_b the formation Y_1 of a new type of polariton on the interface between a cuprous oxide slab and a polystyrene micro-sphere placed on the slab.

[*] Corresponding author

[**] Principal corresponding author

jhonvilcaran@gmail.com (Jhon-Vilcarana); jorgequispe@gmail.com (Jorge-Quispe); hubertfalcon@gmail.com (Hubert-Falcon); soriaquijaite@gmail.com (Soria-Quijaite); saboyarios@gmail.com (Saboya-Rios)

https://upeu.edu.pe (Jhon-Vilcarana); https://upeu.edu.pe (Jorge-Quispe); https://upeu.edu.pe (Hubert-Falcon); https://upeu.edu.pe (Soria-Quijaite); https://upeu.edu.pe (Saboya-Rios)

ORCID(s): 0009-0007-1912-220X (Jhon-Vilcarana)

[1] Jhon Vilcarana fue responsable principal del manuscrito.

[2] Hubert Falcon contribuyó en análisis y revisión.

[3] Jorge Quispe realizó recopilación de datos.

[4] Soria Quijaite colaboró en análisis estadístico.

[5] Saboya Rios supervisó y editó el trabajo.

útiles para identificar tendencias generales, pero tienden a simplificar la realidad, ignorando las interacciones no lineales y las incertidumbres inherentes en los datos (Mutlaq, ALOTAIBI, HUSAYKAN, Alotaibi, Alotaibi, Alotibi, Alotaibi, Abualssayl, MUBARAK, Almutairi, BINNWEJIM & Khabrani, 2024).

A medida que la disponibilidad de grandes volúmenes de datos ha aumentado, gracias a los avances en la recolección de datos y las tecnologías de información, se ha vuelto evidente que los enfoques tradicionales no son suficientes para abordar problemas de salud tan multifactoriales como la anemia (Ma, 2024). En respuesta a las limitaciones de los métodos tradicionales, la comunidad científica ha comenzado a explorar el uso de técnicas más avanzados basadas como los modelos de machine learning (Pullakhandam & McRoy, 2024).

En Perú se han aplicado técnicas de minería de datos y aprendizaje automático para predecir y clasificar la anemia infantil, destacándose el uso de algoritmos como Naive Bayes, regresión logística, KNN, árboles de decisión y Random Forest (Céspedes Panduro, 2022; Marcos Valdez, Navarro Ortiz, Quinteros Peralta, Tirado Julca, Valentín Ricaldi & Calderón-Vilca, 2023; Ramirez & Karina; Mayhua, Huamani, Tovar & Challa, 2022). En este contexto, Marcos Valdez et al. (2023) reportaron una exactitud del 70% en la predicción de anemia en niños menores de cinco años, mientras que Ramirez & Karina aplicaron Random Forest a datos de la ENDES 2022 en niños de 6 a 59 meses, obteniendo una sensibilidad del 65.9% y una especificidad del 63.6%. De forma complementaria, Mayhua et al. (2022) utilizaron árboles de decisión en Arequipa y alcanzaron una precisión del 74.5% en el diagnóstico de anemia. A pesar de estos avances, los modelos empleados han evidenciado porcentajes de precisión relativamente bajos y dificultades para gestionar la incertidumbre inherente a los datos de salud pública (Harshavardhanan & Nene, 2020).

Estudios recientes han destacado la eficacia de la entropía difusa en el análisis de problemas de salud, mostró resultados prometedores en la identificación de factores complejos en condiciones crónicas (Cao & Lin, 2018). Un sistema de inferencia que combina entropía difusa con DEMATEL alcanzó una precisión del 98.69% en la detección temprana de enfermedades cardiovasculares, superando métodos anteriores (Mariadoss & Augustin, 2023). De igual manera, un prototipo de predicción que integra entropía difusa con técnicas de aprendizaje automático logró una precisión del 98% en la clasificación de riesgo de diabetes, resaltando el potencial de esta técnica en la identificación temprana de la enfermedad (Ruby, Chandran, Jain, Chaithanya & Patil, 2023). Asimismo, en el estudio de la variabilidad de la frecuencia cardíaca, la entropía difusa mostró resultados consistentes y confiables, ofreciendo una forma más estable de analizar los datos en comparación con los métodos tradicionales. Esto cobra especial importancia en investigaciones cardiovasculares, donde comprender con precisión los cambios en el ritmo cardíaco puede marcar la diferencia en la detección temprana y el seguimiento de distintas condiciones de salud (Liu, Li, Zhao, Liu, Zheng, Liu & Liu, 2013). Estos hallazgos muestran que la entropía difusa ha sido utilizada con buenos resultados en distintos ámbitos de la salud, especialmente en aquellos donde los datos presentan un alto grado de complejidad. Sin embargo, su aplicación para identificar los factores que influyen en la prevalencia de la anemia infantil sigue siendo un tema poco abordado y con un amplio potencial de investigación (Salas Ccapacca, 2018).

En el contexto peruano, los estudios más recientes sobre la predicción y clasificación de la anemia infantil se han centrado principalmente en variables clínicas básicas y de fácil obtención, como el peso, la talla, la edad y el sexo (Rivera, Cardenas, Castillo-Sequera & Wong, 2024). Aunque estos trabajos han aportado información relevante, aún enfrentan dificultades para capturar la interacción entre múltiples factores y la incertidumbre que caracteriza los datos de salud pública (Incorvaia, Hond & Asgari, 2024). Esto ha limitado la precisión de los modelos y su aplicación práctica en el diseño de estrategias efectivas, especialmente en zonas rurales y en poblaciones más vulnerables del país (Westgard, Orrego-Ferreyros, Calderón & Rogers, 2020).

A nivel internacional, también se han identificado factores relevantes para la anemia infantil y condiciones relacionadas (Ajmer, 2020). En India, los principales determinantes incluyen el estado anémico de la madre, el nivel socioeconómico, el nivel educativo materno y la edad del niño (Meitei, Saini, Mohapatra et al., 2022). En Etiopía, se reportaron como factores clave el tiempo de acceso al agua potable, el bajo peso materno, el tamaño pequeño al nacer y la edad del niño mayor de 30 meses (Bitew, Sparks & Nyarko, 2022). Finalmente, en China, investigaciones en adultos mayores encontraron que la exposición a metales como Cu, As y Te contribuyó significativamente a la prevalencia de anemia (Qiao, Shen, Duan, Wang, He, Zhang, Dai, Li, Chen, Li, Zhao, Liu, Yang, Zhang, Guan & Sun, 2024). En conjunto, las investigaciones realizadas tanto en Perú como en otros países han identificado múltiples factores relacionados con la anemia infantil y han señalado las limitaciones de los métodos actuales para analizar este problema (Marcos Valdez et al., 2023). Esto resalta la necesidad de un enfoque más integral que permita manejar la complejidad y la incertidumbre asociadas a la enfermedad (Tesfaye, Tessema & Jarso, 2020). Dicho enfoque debe ir más allá del uso exclusivo de variables tradicionales como peso o talla, incorporando también factores sociodemográficos, socioeconómicos y socioculturales que permitan comprender de forma más completa las causas y condiciones que favorecen a la prevalencia de la anemia infantil (Westgard et al., 2020).

Frente a esta necesidad, el presente estudio propone un modelo basado en entropía difusa orientado a analizar e identificar los factores asociados a la prevalencia de la anemia en niños y niñas. Esta metodología combina los principios de la lógica difusa y la teoría de la información, lo que facilita manejar la incertidumbre y las relaciones no lineales que suelen presentarse en los datos de salud. A partir de información proveniente del contexto peruano, el modelo

busca ofrecer una clasificación más precisa de los factores de riesgo vinculados a la anemia, constituyéndose en una herramienta adaptable y aplicable a otros entornos epidemiológicos similares. Este enfoque puede servir de apoyo para diseñar estrategias de intervención en salud pública más efectivas y basadas en evidencia.

El estudio se distingue por aplicar un modelo de clasificación innovador que integra la entropía difusa, técnica que combina la lógica difusa con el concepto de entropía para manejar la variabilidad y la incertidumbre propias de los datos clínicos y sociales (Sun, Wong, Wang & Tong, 2017). A nivel internacional, se han empleado diversos métodos de aprendizaje automático, como *Elastic Net*, *Random Forest* y regresiones logísticas, para estudiar los factores relacionados con la anemia infantil. Sin embargo, en el contexto peruano, la mayoría de los trabajos se ha centrado únicamente en la predicción de la enfermedad a partir de variables clínicas (Davalos Carrera & Cortez Paredes, 2025). Hasta el momento, no se había desarrollado un modelo enfocado en analizar la prevalencia ni en identificar de manera detallada los factores que contribuyen a su persistencia (Gurudatha & Majhi, 2024).

Los enfoques actuales han permitido detectar ciertos patrones, su precisión suele ser limitada, alcanzando niveles de desempeño inferiores al 74%. Esto restringe su utilidad para ofrecer análisis predictivos más detallados que orienten decisiones efectivas en salud pública (Marcos Valdez et al., 2023). En este escenario, la entropía difusa representa un avance metodológico importante, al permitir interpretar mejor los datos y mejorar la clasificación de factores de riesgo, superando las limitaciones de los métodos convencionales (Sukanesh & Harikumar, 2006).

2. Contribuciones clave del trabajo propuesto

Con el objetivo de lograr una alta relevancia clínica y precisión en el contexto peruano, este trabajo introdujo un modelo híbrido novedoso que combinó ingeniería de características basada en entropía difusa, clasificación XGBoost y optimización del umbral de decisión. A través de este enfoque, se identificó que la Entropía Difusa era la clave para manejar la ambigüedad inherente en los datos médicos, facilitando un análisis profundo de los factores que contribuían a la prevalencia de anemia en niños. Más allá de la teoría, esta investigación ofreció un recurso práctico para diseñar políticas públicas focalizadas donde más se necesitaban. Su valor radicó en su capacidad de adaptación; fue una solución que no solo resolvió el problema inmediato, sino que podía replicarse para abordar otros retos multifactoriales en la salud pública.

1. **Integración de características basadas en Entropía Difusa:** A diferencia de los estudios tradicionales que dependían únicamente de variables básicas clínicas sin procesar, esta investigación incorporó la Entropía Difusa como técnica de ingeniería de características. Este enfoque no solo procesó los datos, sino que cuantificó la incertidumbre inherente a las mediciones de salud. Al hacerlo, se dotó al modelo de la sensibilidad necesaria para distinguir casos de anemia difíciles de clasificar, reflejando así la verdadera complejidad de las condiciones médicas.

2. **Marco XGBoost Optimizado por Umbral:** El estudio introdujo una estrategia de optimización especializada para la arquitectura XGBoost. En lugar de depender de umbrales de decisión predeterminados (0.5), este estudio implementó una técnica de maximización del F1-Score para ajustar dinámicamente el umbral de clasificación. Este proceso mitigó eficazmente el sesgo hacia la clase mayoritaria que se encontraba a menudo en conjuntos de datos médicos desbalanceados. Como resultado, el modelo propuesto logró una exactitud de prueba superior del 90.9%, destacando una mejora significativa sobre Random Forest estándar y enfoques no optimizados.

3. **Gestión robusta de la incertidumbre clínica:** Al crear una sinergia entre los principios de la lógica difusa y los algoritmos de gradient boosting, el estudio demostró que se podía lograr una alta sensibilidad diagnóstica (Recall: 88.3%) incluso con datos demográficos complejos y desbalanceados. Este enfoque no solo mejoró el rendimiento predictivo para los casos positivos subrepresentados como los niños con prevalencia de anemia, sino que también ofreció una herramienta computacional escalable y no invasiva para el tamizaje temprano en entornos de salud pública con recursos limitados en Perú.

3. Metodología

El propósito del modelo es identificar con mayor precisión los factores que contribuyen a la prevalencia de la anemia en niños peruanos, incluso en aquellos que ya han recibido tratamiento. Para ello, se propone el modelo de clasificación basado en entropia difusa para la identificacion de factores que contribuyen en la prevalencia de anemia en niños peruanos, esta incorpora una capa de entropía difusa con el objetivo de representar y gestionar de manera más efectiva la incertidumbre presente en los datos clínicos.

Este enfoque parte del reconocimiento de una realidad común en los entornos de salud como en la Red de Salud de Huarochirí en Perú, información incompleta, datos de distintas fuentes o formatos, y valores que no siempre son consistentes. En ese contexto, la entropía difusa permite traducir dicha ambigüedad en una medida numérica que el modelo puede interpretar y aprovechar, aportando una representación más clara de la coondición de cada paciente.

Más allá de la predicción convencional, el modelo busca identificar patrones tempranos que puedan anticipar la continuidad de la anemia en el tiempo. De esta manera, no solo contribuye al diagnóstico, sino también a la comprensión de los factores que dificultan la recuperación, ofreciendo información valiosa para orientar intervenciones más oportunas y efectivas en la salud infantil. Así, se prioriza el uso

de herramientas que no solo alcancen una mayor precisión, sino que también ayuden a tomar decisiones clínicas más informadas y justificadas, especialmente en poblaciones vulnerables.

El modelo de clasificación no solo introduce una técnica novedosa en el contexto peruano, sino que también aporta un marco replicable para analizar otros problemas complejos de salud pública, donde los datos no siempre son claros y las decisiones requieren considerar múltiples factores con diferentes grados de certeza.

3.1. Descripción del negocio

El ciclo CRISP-DM como se demuestra en la Figure 1 ordena el trabajo en pasos claros. Se empezó por entender el problema, se buscó determinar con datos de clínicos qué niños presentaban una mayor probabilidad de seguir con anemia pasado el tratamiento estándar de seis meses (Ministerio de Salud (Perú), 2024). Con este objetivo, se revisó la evidencia disponible, se definió el objetivo de clasificación y se acordó desde el inicio que la incertidumbre de los datos debía medirse y visualizarse de forma explícita utilizando el concepto de entropía difusa.

Se trabajó con historias clínicas de la Red de Salud de Huarochirí entre 2020 y 2024. Se unió todas las consultas de cada niño para formar un solo registro por paciente y se exploró la calidad de la información: distribución de variables, consistencia y huecos en los registros.

Se agrupó y limpió las consultas, se eliminó duplicados y se corrigieron valores imposibles. Los faltantes se completaron con un valor aleatorio dentro de un rango sencillo alrededor de la media de cada variable, usando una desviación estándar por debajo y por encima como límites. Se definió la etiqueta del estudio de forma operativa: hay prevalencia de anemia cuando los controles por anemia se mantienen durante seis meses o más desde la primera consulta; en caso contrario, no hay prevalencia. Para facilitar el aprendizaje del modelo, se codificó las variables categóricas, se estandarizó las numéricas y se promedió las mediciones repetidas por paciente, como el promedio de peso, talla y hemoglobina.

Para cada variable se estableció rangos de trabajo usando solo los datos de entrenamiento y se definieron funciones de pertenencia triangulares que convierten cada valor en tres niveles fáciles de leer, bajo, medio y alto. A partir de esos niveles se calculó la entropía difusa por variable y también un resumen global por paciente. Cuando la entropía es baja, la señal es clara, cuando es alta, la información es ambigua y conviene decidir con más cautela.

Con todo esto se formó el conjunto final de entrada y se pasó al modelado. Se usó un clasificador XGBoost que recibe las variables originales junto con los grados difusos y la entropía. Se ajustó el modelo con validación cruzada estratificada y una búsqueda ordenada de parámetros para evitar sobreajuste y quedarnos con una configuración estable. También se comparó con líneas base sin capa difusa para mostrar el aporte de las nuevas variables.

La salida del modelo es una probabilidad de prevalencia que se convirtió en una etiqueta con un umbral elegido durante el entrenamiento para equilibrar precisión y sensibilidad, tomando como guía el F_1-score de la clase con prevalencia. La evaluación se hace una sola vez sobre el conjunto de prueba separado desde el inicio, y reporta exactitud, precisión, sensibilidad, F_1-score, áreas bajo las curvas ROC y Precision–Recall, además de la matriz de confusión y las curvas con el punto operativo marcado. El resultado es un flujo claro, reproducible y entendible que muestra qué señales sostienen la decisión y con qué grado de certeza.

3.2. Comprensión de los datos

La primera fase del estudio consistió en explorar y analizar las características del conjunto de datos, con el fin de comprender la estructura de la información, identificar variables clave y evaluar la calidad de los datos. El dataset utilizado fue proporcionado por la municipalidad de Huarochirí y contiene registros clínicos y demográficos de pacientes pediátricos atendidos entre los años 2020 y 2024. Este conjunto de datos abarca una variedad de variables clínicas, resultados de laboratorio y factores relacionados con el sistema de salud, lo que permite una visión general de los factores que podrían influir en la prevalencia de la anemia en niños peruanos. Para asegurar la calidad del análisis, se realizó una inspección inicial del dataset con el objetivo de detectar posibles inconsistencias, valores faltantes o redundantes. Se identificaron tanto variables clínicas clave (peso, talla, hemoglobina) como demográficas (edad, género, etnia, tipo de financiador). Una de las principales características del dataset es su tamaño y la complejidad en datos registros provenientes de toda la Red de Salu fueron procesados mediante un riguroso proceso de limpieza. Asimismo, se identificaron variables que no aportaban valor al análisis, los cuales fueron excluidos del modelo para evitar sesgo en el dataset. Finalmente, se filtraron los registros de pacientes con edades comprendidas entre 0 y 5 años, así como aquellos que habían recibido atención por anemia. Dado que los archivos estaban separados por meses, se procedió a la consolidación de los datos mediante el número de documento del paciente, agrupando todas las consultas correspondientes a cada individuo a lo largo del período de estudio. Este proceso tuvo como objetivo estructurar un subconjunto de datos que sirviera como base para el modelo de clasificación basado en entropía difusa; estas variables se seleccionaron con ayuda de la nutricionista del centro de salud.

3.3. Preparación de los datos

La preparación de los datos se realizó tras seleccionar las variables relevantes para la investigación, con el objetivo de identificar pacientes con y sin prevalencia de anemia, considerando la duración estándar del tratamiento de seis meses. Inicialmente, se clasificaron los pacientes en función de sus consultas por anemia en años diferentes, marcando como prevalentes a aquellos con recurrencia del diagnóstico y como no prevalentes a quienes lograron recuperarse tras el tratamiento. Para garantizar un análisis completo, se utilizó el número de documento de identidad como identificador

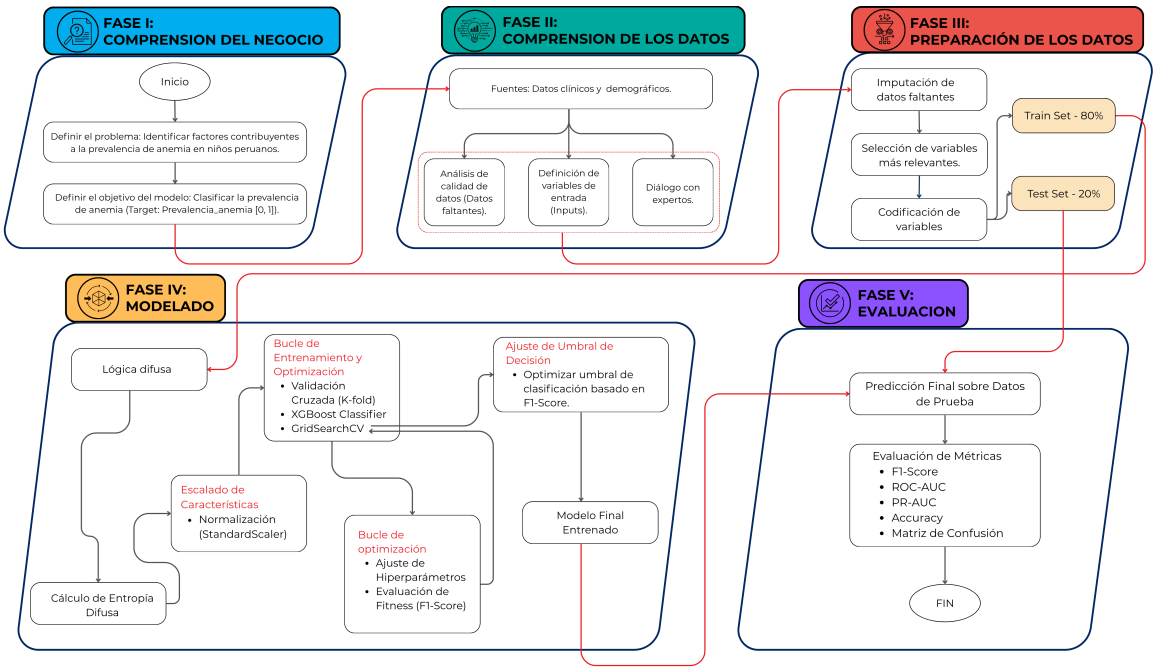


Figure 1: El siguiente diagrama ilustra el proceso metodológico completo, destacando cada etapa desde la comprensión inicial del problema hasta la evaluación final de los resultados.

único, permitiendo rastrear y consolidar todos los registros de un mismo paciente en el dataset. Adicionalmente, se recopilaron datos sobre consultas realizadas en otras especialidades médicas, excluyendo aquellas relacionadas con nutrición, ya que estas se enfocaban principalmente en la suplementación o tratamiento específico de la anemia, lo que podría sesgar los resultados.

Una vez seleccionadas las variables relevantes, se agruparon las consultas de los pacientes registrados entre 2020 y 2024, identificando redundancias en el dataset. Se observó que múltiples registros pertenecían a los mismos pacientes y que las variables categóricas, como género, etnia y financiador, permanecían constantes a lo largo de sus consultas, por lo que se consolidaron en un único registro por paciente; además, se calcularon los promedios por paciente de peso, talla y hemoglobina, obteniendo las variables Peso promedio, Talla promedio y Hemoglobina promedio como representaciones centrales de todas las citas registradas durante el seguimiento. Sobre este registro consolidado se definió la variable objetivo "Prevalencia de anemia", se etiquetó como 1 como prevalencia cuando el control por anemia se extendió seis meses o más desde la primera consulta, y 0 como no prevalencia cuando hubo una sola consulta o todas las consultas por anemia ocurrieron dentro de los primeros seis meses. Adicionalmente, se detectaron inconsistencias significativas en las variables clínicas como peso, talla y hemoglobina, con valores atípicos extremos como pacientes con pesos irreales de 400 kg y una cantidad considerable de datos faltantes. Para abordar las limitaciones asociadas a los datos faltantes como se observa en la Figure 2, se aplicó una imputación mediante muestreo aleatorio uniforme (Bhushan, Kumar & Pokhrel, 2024). Primero, se calcularon la media (μ) y la desviación estándar (σ) a partir de los valores observados de cada variable. Luego, se definió un intervalo simétrico alrededor de la media, comprendido entre $\mu - \sigma$ y $\mu + \sigma$. Cada valor faltante se reemplazó por un valor imputado x^* como se muestra en la Figure 3.

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i, \qquad \sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2},$$
$$L_{\text{inf}} = \mu - \sigma, \qquad L_{\text{sup}} = \mu + \sigma,$$
$$x^* \sim \mathcal{U}\left(\mu - \sigma,\ \mu + \sigma\right) \tag{1}$$

Después de limpiar y consolidar el conjunto de datos, se redujó significativamente la cantidad de registros, se pasó de 32 807 datos a 6 339 datos, manteniendo únicamente información de pacientes únicos. Esta depuración fue clave, ya que permitió organizar mejor los datos, agrupando en columnas representativas información relevante como el número de diagnósticos por paciente, los distintos niveles de anemia detectados, y las mediciones clínicas (peso, talla y hemoglobina) tanto en la primera como en la última consulta registrada. Durante la transformación de los datos, se observó que algunas variables numéricas presentaban una gran dispersión, lo cual podía generar ruido o interpretaciones erróneas al momento de modelar. Se aplicó distintos tipos de codificación según la naturaleza de cada variable. En el caso de las variables categóricas, como el género, la etnia y el tipo de seguro, se utilizó codificación numérica (Label Encoding), asignando un valor específico a cada categoría.

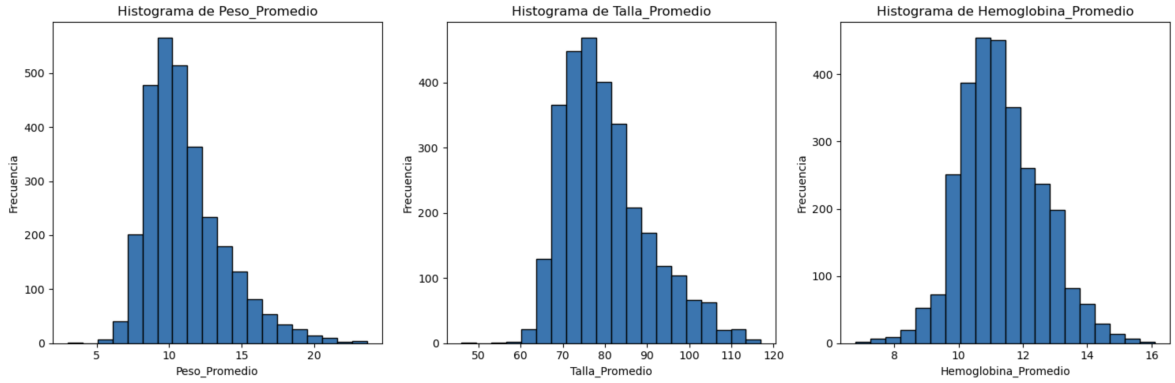


Figure 2: Distribución original de la variable con presencia de datos faltantes. Se observan interrupciones en la densidad debido a la ausencia de registros en ciertas regiones del dominio.

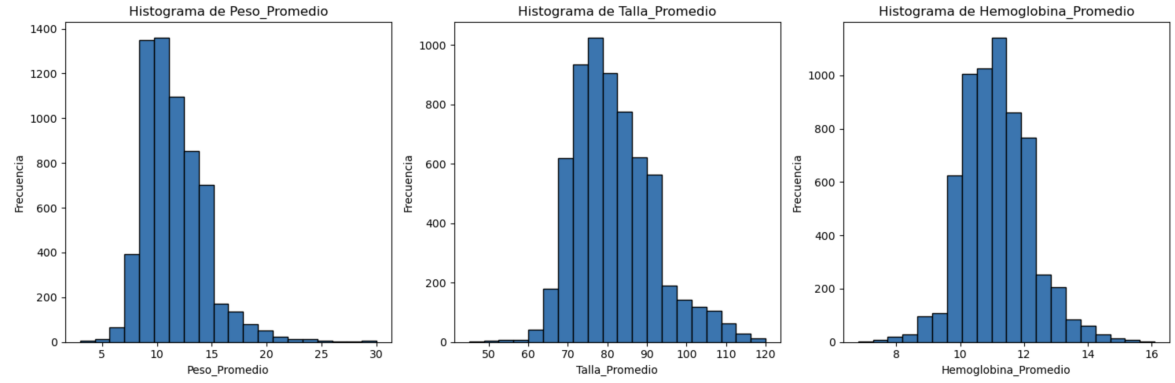


Figure 3: Distribución de la variable después de la imputación. Los valores faltantes fueron reemplazados por muestreo aleatorio uniforme dentro del intervalo $[\mu - \sigma,\ \mu + \sigma]$, manteniendo la dispersión natural de los datos sin sesgar hacia la media.

Esto permitió conservar la información cualitativa de forma manejable para el modelo.

Con las variables numéricas más dispersas, como peso, talla y hemoglobina, se decidió agrupar sus valores en intervalos definidos por sus rangos observados. Así, cada valor quedó clasificado como "bajo", "medio" o "alto", lo que facilitó no solo su interpretación clínica, sino también su incorporación al sistema difuso.

Este proceso fue esencial para dejar el conjunto de datos listo para el modelado, buscando siempre un balance entre simplificar sin perder detalle, y preparar la información para que el modelo de entropía difusa pudiera aprovechar al máximo la riqueza clínica contenida en los registros.

Se usó SelectKBest para identificar las variables que realmente aporten al modelo (Table 3) y descartar lo que añade ruido, esta técnica ordena cada variable predictora según su aporte para distinguir mejor los factores que contribuyen en la prevalencia de anemia (Saeed & Hama, 2023). El cálculo se hizo con todas las variables para determinar cuáles eran más relevantes, mediante el puntaje obtenido, este control evita redundancias, reduce el tamaño del conjunto y hace más estable el entrenamiento. Se determinó que no todo se decide por el número, en salud existen factores que aunque no destaquen en el ranking, son relevantes por su peso clínico o social. Por ese motivo se conservó algunas variables de bajo puntaje cuando aportaban contexto que el modelo podía aprovechar. Esta estrategia trae beneficios concretos, reduce el riesgo de sobreajuste. Así se desminuyó el costo de cómputo y se aceleró el procesamiento, al trabajar con menos columnas y señales más limpias. SelectKBest dió una guía objetiva para priorizar predictores, la revisión de redundancias y el criterio clínico impidió perder información útil, para lograr un conjunto de entrada claro, compacto e interpretable, listo para alimentar la siguiente etapa del estudio del modelado, donde se incorporó lógica difusa y entropía difusa para tratar explícitamente la incertidumbre (Ayyanar, Jeganathan, Parthasarathy, Jayaraman & Lakshminarayanan, 2022).

3.4. Modelado

El modelo propuesto combina lógica difusa y entropía difusa con el objetivo de manejar la incertidumbre y la variabilidad inherente a los datos clínicos (Jin & Garg, 2023). La lógica difusa permite representar variables en términos lingüísticos —como bajo, medio y alto— mediante funciones de pertenencia que capturan transiciones graduales entre valores (Kaczmarek-Majer, Rutkowska & Hryniewicz, 2022). Este enfoque resulta especialmente valioso para variables como hemoglobina, peso, talla o años en consulta, donde los límites estrictos entre categorías no reflejan

Table 1
Variables categóricas y numéricas transformadas mediante codificación, utilizadas en el modelo de clasificación.

Columna	Descripción	Valores codificados
Genero	Género del paciente	0 = Femenino, 1 = Masculino
Etnia	Etnia del paciente	1 = Mestizo, 2 = Blanco, 3 = Negro, 4 = Otros
Financiador	Tipo de financiamiento	1 = S.I.S, 2 = Usuario, 3 = Otros
Citas_Anemia	N.º de citas relacionadas con anemia	1 = 0–10, 2 = 11–20, 3 = 21–40, 4 = >40
Otras_Citas	N.º de otras citas	1 = 0–50, 2 = 51–100, 3 = 101–150, 4 = 151–200, 5 = >200
Diagnostico_Repetitivo	Nivel de diagnóstico repetitivo	1 = 0–10, 2 = 11–30, 3 = 31–50, 4 = >50
Citas_Primer_Anio	Citas en el primer año del paciente	1 = 1–5, 2 = 6–10, 3 = >10
Citas_Ultimo_Anio	Citas en el último año del paciente	1 = 1–5, 2 = 6–10, 3 = >10
Citas_2020	Citas en el año 2020	1 = 0–5, 2 = 6–10, 3 = >10
Citas_2021	Citas en el año 2021	1 = 0–5, 2 = 6–10, 3 = >10
Citas_2022	Citas en el año 2022	1 = 0–5, 2 = 6–10, 3 = >10
Citas_2023	Citas en el año 2023	1 = 0–5, 2 = 6–10, 3 = >10
Citas_2024	Citas en el año 2024	1 = 0–5, 2 = 6–10, 3 = >10
A_Leve	Grado de anemia leve	1 = 0–10, 2 = 11–20, 3 = 21–40, 4 = >40
A_Moderado	Grado de anemia moderada	1 = 0–5, 2 = 6–10, 3 = >10
A_Severo	Grado de anemia severa	1 = 0–2, 2 = 3–5, 3 = >5
1Cita_Peso	Peso en la primera cita (kg)	1 = ≤10, 2 = 11–20, 3 = >20
Ucita_Peso	Peso en la última cita (kg)	1 = ≤10, 2 = 11–20, 3 = >20
Peso_Combinado	Último o primer valor existente	1 = ≤10, 2 = 11–20, 3 = >20
1Cita_Talla	Talla en la primera cita (cm)	1 = ≤50, 2 = 51–100, 3 = >100
Ucita_Talla	Talla en la última cita (cm)	1 = ≤50, 2 = 51–100, 3 = >100
Talla_Combinada	Último o primer valor existente	1 = ≤50, 2 = 51–100, 3 = >100
1Cita_Hemoglobina	Hemoglobina en la primera cita (g/dL)	1 = ≤8, 2 = 9–12, 3 = >12
Ucita_Hemoglobina	Hemoglobina en la última cita (g/dL)	1 = ≤8, 2 = 9–12, 3 = >12
Hemoglobina_Combinada	Último o primer valor existente	1 = ≤8, 2 = 9–12, 3 = >12
Peso_Promedio	Peso promedio durante todas las consultas	1 = ≤10, 2 = 11–20, 3 = >20

Table 2
Variables que no requirieron codificación, ya que presentaban rangos acotados y eran directamente utilizables en el modelo.

Columna	Descripción	Rango / Valores
Edad_inicio_consulta	Edad del paciente al inicio de la consulta	0 – 5
Total_Anios_En_consulta	Duración total en años en consulta	0 – 5
Edad_fin_consulta	Edad del paciente al final de la consulta	0 – 5
Diagnostico_Definitivo	Nivel del diagnóstico definitivo	0 – 7
Diagnostico_Presuntivo	Nivel del diagnóstico presuntivo	0 – 5
P_Recuperado	Estado de recuperación	0 – 4
Prevalencia_anemia	Indicador de prevalencia de anemia	0 = Sin anemia, 1 = Con anemia

adecuadamente la realidad clínica (Shimizu, Hayakawa, Takada, Okada & Kiyama, 2021).

En este contexto, la lógica difusa y la entropía difusa cumplen roles complementarios (Arya & Kumar, 2020). La primera constituye el marco de representación de la información, asignando a cada variable un grado de pertenencia continuo a categorías lingüísticas, evitando así las limitaciones de los umbrales rígidos (Ali, Mekhilef, Nukman & Razak, 2022). La segunda actúa como medida derivada que cuantifica la incertidumbre de dichas asignaciones, valores bajos de entropía reflejan certeza, mientras que valores elevados indican máxima dispersión e incertidumbre (Sun, Wang, Ding, Qian & Xu, 2021). De esta manera, la lógica difusa describe cómo se estructura la información clínica, mientras que la entropía difusa proporciona una estimación del nivel de incertidumbre en esa representación (de Sá Ferreira, 2023).

Para cada variable se definió un universo de discurso basado en los valores extremos observados tras la limpieza de datos (Nasfi, de Tré & Bronselaer, 2025). Dentro de este intervalo se parametrizaron funciones de pertenencia triangulares, seleccionadas por su simplicidad computacional, facilidad de interpretación y uso extendido en aplicaciones médicas (Chotikunnan, Chotikunnan, Pititheeraphab, Puttasakul, Wongkamhang & Thongpance, 2025). Cada registro del conjunto de datos se transformó en un vector de grados de pertenencia que cuantifica el nivel de asociación del valor con las tres categorías establecidas (Huang, Lu & Yang, 2024). Posteriormente, se calculó la entropía difusa de cada variable a partir de estos grados de pertenencia,

Table 3
Importancia de variables seleccionadas según el algoritmo SelectKBest, ordenadas por relevancia en relación con la variable objetivo (Prevalencia de anemia). Las últimas dos variables, aunque con menor puntuación, fueron incluidas por su relevancia teórica como variables categóricas de interés en el análisis.

Variable	Puntuación
Total_Anios_En_consulta	0.187452
Diagnostico_Definitvo	0.090763
Otras_Citas_Encoded	0.088389
Citas_Anemia_Encoded	0.077629
Etnia_Encoded	0.066477
Diagnostico_Repetitivo_Encoded	0.054762
Talla_Combinada_Encoded	0.054280
Hemoglobina_Combinada_Encoded	0.042371
P_Recuperado	0.039080
Edad_inicio_consulta	0.033796
Genero_Encoded	0.016611
Financiador_Encoded	0.012044

mediante una adaptación de la entropía de Shannon normalizada en el rango [0,1] (Zhang, Jiao, Zhang, Wang & Cui, 2024). Finalmente, se definió una entropía global por paciente como el promedio de las entropías individuales de las variables clínicas seleccionadas (Muthalaly, Kwong, John, van der Geest, Tao, Schaeffer, Tanigawa, Nakamura, Kaneko, Tedrow, Stevenson, Epstein, Kapur, Zei & Koplan, 2019).

Las salidas generadas por el modelo se integraron en un esquema constituido por dos componentes principales (Zhang, Sun & Peng, 2023). En primer lugar, un sistema de inferencia difusa diseñado para generar una puntuación de riesgo de prevalencia de anemia infantil a partir de un conjunto de reglas clínicas del tipo si–entonces, definidas con apoyo de expertos en nutrición pediátrica (Jalaludin, Kiau, Hasim, Lee, Low, Kazim, Hoi & Taher, 2025). Estas reglas modelan relaciones no lineales entre variables y permiten obtener una salida continua, que posteriormente se defuzzifica mediante el método del centroide (Meng, Li, Yang & Gao, 2025). En segundo lugar, un algoritmo de clasificación supervisada basado en Extreme Gradient Boosting (XGBoost), incorporando tanto los grados de pertenencia como las medidas de entropía como variables adicionales (Tian, Tsai, Zhang, Cai, Xiao, Yu, Zhao & Chen, 2024). Este clasificador fue seleccionado por su capacidad para modelar interacciones complejas, manejar datos heterogéneos y ofrecer un alto rendimiento en entornos clínicos (Ma, Meng, Yan, Yan, Chai & Song, 2020).

Este enfoque no solo incrementa la precisión predictiva del modelo, sino que también mejora su capacidad explicativa (Sarstedt & Danks, 2021). La entropía difusa aporta una medida clara del nivel de incertidumbre en los datos, mientras que las reglas clínicas aseguran trazabilidad en la toma de decisiones. De esta manera, el sistema logra un balance entre rendimiento estadístico y transparencia, condiciones esenciales en aplicaciones médicas (Bhardwaj & Sharma, 2021).

El modelo propuesto integra la flexibilidad de la lógica difusa, la capacidad de cuantificar incertidumbre mediante entropía y la potencia predictiva de XGBoost, resultando en una herramienta robusta que combina interpretabilidad clínica con alto desempeño predictivo (Nimmala, Mahendar, Manasa, Lakshmi, Kiran & Raghavendra, 2024).

3.4.1. Lógica difusa

Los conjuntos difusos constituyen la base de la lógica difusa y permiten modelar la incertidumbre en dominios donde las fronteras entre categorías no son nítidas (Bargues, Cruz, Álvarez & Prieto-Gómez, 2017). A diferencia de los conjuntos clásicos, en los que la pertenencia de un elemento se define de manera binaria ($x \in A$ o $x \notin A$), en un conjunto difuso cada elemento x posee un grado de pertenencia $\mu_A(x)$ expresado en el intervalo continuo $[0, 1]$ (García, Sandoval, Vega & Herrera, 2021). Un conjunto difuso A definido sobre un universo de discurso X se describe como:

$$A = \{(x, \mu_A(x)) \mid x \in X,\ \mu_A(x) \in [0, 1]\}, \tag{2}$$

donde la función de membresía $\mu_A(x)$ asigna a cada valor de x un grado que refleja la intensidad con la que cumple la propiedad característica del conjunto (Jun & Xin, 2019).

En este estudio, las variables clínicas y sociales asociadas a la anemia infantil fueron transformadas en términos lingüísticos como bajo, medio y alto. Cada uno de estos términos se modeló mediante funciones de pertenencia triangulares (of Computer Science & Engineering, 2020).

$$\mu_{\text{bajo}}(x) = \begin{cases} 1, & x \le a, \\ \dfrac{b - x}{b - a}, & a < x < b, \\ 0, & x \ge b \end{cases} \tag{3}$$

$$\mu_{\text{medio}}(x) = \begin{cases} 0, & x \le a \ \text{ o } \ x \ge c, \\ \dfrac{x - a}{b - a}, & a < x < b, \\ \dfrac{c - x}{c - b}, & b \le x < c \end{cases} \tag{4}$$

$$\mu_{\text{alto}}(x) = \begin{cases} 0, & x \le a, \\ \dfrac{x - a}{b - a}, & a < x < b, \\ 1, & x \ge b \end{cases} \tag{5}$$

donde a, b, c corresponden a parámetros definidos a partir del rango observado en los datos.

Este enfoque permite que un valor pueda pertenecer simultáneamente, con diferentes grados, a más de una categoría. Como un nivel de hemoglobina cercano al umbral entre bajo y medio puede tener $\mu_{\text{bajo}}(x) = 0.4$ y $\mu_{\text{medio}}(x) =$

0.6. Esto refleja la gradualidad propia de la práctica clínica y evita la pérdida de información en las fronteras rígidas.

De esta manera, los conjuntos difusos proveen la representación inicial sobre la cual se construyen cálculos de incertidumbre (entropía difusa) y reglas de inferencia, integrándose al modelo de clasificación para fortalecer su capacidad predictiva.

3.4.2. *Inferencia difusa*

Una vez definidas las funciones de pertenencia, el siguiente paso en el modelo consiste en la inferencia difusa, la cual permite razonar a partir de reglas que se expresan de la siguiente forma:

$$R_j : \quad \text{Si } x_1 \text{ es } A_1^j \wedge x_2 \text{ es } A_2^j \wedge \cdots \wedge x_n \text{ es } A_n^j \Rightarrow y \text{ es } B^j, \quad (6)$$

donde cada x_i es una variable de entrada, A_i^j representa un conjunto difuso asociado por *bajo*, *medio*, *alto*) y B^j corresponde a un conjunto difuso de salida (Camacho, Iglesias, Herrera & Aboukheir, 2021).

La activación de una regla se determina aplicando operadores difusos sobre los grados de pertenencia de las entradas (van Lambalgen, s.f.).

$$\begin{aligned} \mu_{A\wedge B}(x) &= \min\{\mu_A(x), \mu_B(x)\}, \\ \mu_{A\vee B}(x) &= \max\{\mu_A(x), \mu_B(x)\}, \\ \mu_{\neg A}(x) &= 1 - \mu_A(x). \end{aligned} \quad (7)$$

De este modo, un mismo valor de entrada puede activar simultáneamente varias reglas con distintos grados de intensidad. El motor de inferencia combina estas contribuciones para generar una salida difusa que refleja la incertidumbre y la gradualidad propias del dominio clínico (Chen, Shang, Su, Keravnou-Papailiou, Zhao, Antoniou & Shen, 2021).

Por un lado, hace que el modelo sea más fácil de interpretar, ya que traduce su funcionamiento interno a reglas que podemos comprender. Por otro lado, enriquece los datos originales al generar patrones de activación más detallados. Estos nuevos patrones son los que se utilizó para calcular la entropía difusa y realizar la clasificación supervisada. (Zhao, Wu & Deng, 2023).

Este método permitió detectar relaciones complejas y matices clínicos que se perderían con una categorización demasiado rígida, fortaleciendo así tanto la capacidad predictiva del modelo como su transparencia. (Cho, Zhan & Ghosh, 2022).

3.4.3. *Entropía difusa*

La entropía es una medida fundamental para cuantificar la incertidumbre en cualquier sistema (Shannon, 1948). En el campo de la lógica difusa, se aplicó este concepto usando la entropía difusa, sirvió para evaluar el nivel de imprecisión de los conjuntos difusos (Singh & Sharma, 2021). La clave de esta métrica es que permitió capturar cómo se distribuyen los grados de pertenencia. Esto aporta una visión mas clara y robusta de la incertidumbre que es inherente a datos complejos, como los clínicos. (Wijesuriya, Moreno-Betancur, Carlin, Silva & Lee, 2022).

$$H(p) = -\sum_{i=1}^{K} p_i \log(p_i), \quad (8)$$

En esta fórmula p_i representa el grado de pertenencia normalizado de un elemento, calculado para cada una de las K categorías (De Luca & Termini, 1972).

En la práctica, un valor de entropía cercano a 0 significa que un registro encaja claramente en una categoría, lo que indica que tiene baja incertidumbre. Por el contrario, un valor que se acerca a 1 indica que la asignación de ese registro es mucho más ambigua (Ramos, Franco-Pedroso, Lozano-Diez & González-Rodríguez, 2018).

De esta manera, la entropía difusa permite clasificar los registros en función de la información que aportan (Pandey, Mishra, Rani, Ali & Chakrabortty, 2023). Entropía baja, observaciones con categorías bien definidas, útiles para reforzar patrones existentes (Buscemi, Schindler & Šafránek, 2022). Entropía alta, observaciones ambiguas que contienen información crítica para la detección de casos inciertos, relevantes en el ámbito clínico (vSafr'anek & Thingna, 2020).

Esta característica permite mejorar el proceso de clasificación al proporcionar una medida explícita de la incertidumbre asociada a cada variable (Yin, Liu, Sun & Xia, 2025).

3.4.4. *Sistema de clasificación supervisado*

Una vez transformadas las variables originales mediante lógica difusa e inferencia, se procede a la etapa de clasificación supervisada (Lin & Zhang, 2024). El objetivo de esta fase es aprender un modelo predictivo capaz de distinguir entre los casos positivos y negativos de la variable de interés (prevalencia de anemia) a partir de los atributos enriquecidos con grados de pertenencia y entropía difusa (Boadh, Chaudhary, Dahiya, Dogra, Rathee, Kumar & Rajoria, 2022).

Para este fin se utilizó Extreme Gradient Boosting (XGBoost), un algoritmo basado en el ensamble de múltiples árboles de decisión entrenados de manera secuencial y optimizados mediante gradiente descendente (Chen & Guestrin, 2016). El modelo combina clasificadores débiles $h_t(x)$ en una predicción final (Dhanka & Maini, 2024).

$$\hat{y}_i = \sum_{t=1}^{T} f_t(x_i), \quad f_t \in \mathcal{F}, \quad (9)$$

donde $\mathcal{F}$ es el espacio de árboles de decisión y T el número de iteraciones (Tian et al., 2024).

$$\mathcal{L}(\phi) = \sum_i l(y_i, \hat{y}_i) + \sum_t \Omega(f_t) \quad (10)$$

siendo $l(y_i, \hat{y}_i)$ la función de pérdida logística y $\Omega(f_t)$ un término de regularización que penaliza la complejidad

de los árboles para evitar sobreajuste (Hastie, Tibshirani & Friedman, 2009).

Una ventaja fundamental de este clasificador es su capacidad para manejar datos desbalanceados mediante el ajuste del parámetro `scale_pos_weight`, que compensa la desproporción entre clases (Sutou & Wang, 2024). De esta manera, el modelo mantiene un equilibrio entre sensibilidad (detección de casos positivos) y especificidad (evitar falsos positivos), lo cual es crítico en aplicaciones médicas (Hong, Moon, Choi, Kang, Ye, Kim, Kim, Cho, Choi, eun Lee, Choi, Kim, Heo, Han & Hong, 2025).

En este estudio, el clasificador fue integrado dentro de un `pipeline` junto con el preprocesamiento y la transformación difusa, lo que asegura que las evaluaciones se realicen sin fuga de información (Park, Oh & Pedrycz, 2022). Los resultados del sistema de clasificación se evaluaron con métricas robustas como F_1-score, Accuracy, área bajo la curva ROC (ROC-AUC) y área bajo la curva Precisión-Recall (PR-AUC), que permiten medir su desempeño en contextos de desbalance y su utilidad clínica (de la Cruz Huayanay, Bazán & Russo, 2024).

3.4.5. Validación cruzada e hiperparámetros

Para garantizar una evaluación robusta del modelo y evitar el sobreajuste, se empleó la técnica de validación cruzada estratificada (Szeghalmy & Fazekas, 2023). En esta estrategia, el conjunto de datos D se divide en k subconjuntos disjuntos o folds, de manera que en cada iteración uno de ellos se utiliza como conjunto de validación mientras que los restantes $k-1$ se emplean para entrenamiento (Li, 2023).

$$\mathrm{CV}(k) = \frac{1}{k}\sum_{i=1}^{k} \mathcal{L}\left(f^{(-i)}, D_i\right) \tag{11}$$

D_i representa el fold i utilizado para validación, $f^{(-i)}$ es el modelo entrenado sin dicho subconjunto, y $\mathcal{L}(\cdot)$ corresponde a la función de pérdida o métrica de evaluación seleccionada como F_1-score (Barros, Costa & Araújo, 2021) . Este promedio sobre k iteraciones proporciona una estimación más estable del desempeño general del clasificador.

Una vez establecida la validación cruzada, se procedió a la búsqueda de hiperparámetros Table 4 mediante Randomized Search, técnica que consiste en explorar de manera aleatoria un subconjunto del espacio de posibles configuraciones (Bergstra & Bengio, 2012). Cada configuración evaluada se entrena y valida siguiendo el esquema de k folds, seleccionándose finalmente la que optimiza la métrica objetivo (Soper, 2021).

$$\theta^* = \arg\min_{\theta\in\Theta} \frac{1}{k}\sum_{i=1}^{k} \mathcal{L}\left(f_\theta^{(-i)}, D_i\right) \tag{12}$$

donde θ denota un conjunto de hiperparámetros como n-estimadores, tasa de aprendizaje y profundidad máxima Θ es el espacio de búsqueda definido y θ^* representa la configuración óptima (Karl, Pielok, Moosbauer, Pfisterer, Coors, Binder, Schneider, Thomas, Richter, Lang, Garrido-Merch'an, Branke & Bischl, 2022).

Table 4
Espacio de búsqueda de hiperparámetros y configuración óptima seleccionada para el modelo.

Hiperparámetro	Espacio de Búsqueda	Valor Óptimo
n_estimators	{300, 500, 700, 900}	**700**
learning_rate	{0.01, 0.05, 0.1, 0.2}	**0.01**
max_depth	{3, 4, 5, 6}	**6**
min_child_weight	{1, 3, 5}	**1**
subsample	{0.7, 0.8, 0.9, 1.0}	**0.7**
colsample_bytree	{0.7, 0.8, 0.9, 1.0}	**0.9**
gamma	{0, 0.5, 1, 2}	**0**
reg_lambda (L2)	{0.0, 0.5, 1.0, 2.0}	**1.0**
reg_alpha (L1)	{0.0, 0.25, 0.5, 1.0}	**0.5**

Este enfoque, al combinar validación cruzada con búsqueda aleatoria, permite encontrar un balance entre eficiencia computacional y capacidad de generalización, evitando que el modelo XGBoost se sobreajuste a una sola partición de los datos y garantizando un rendimiento más confiable en el conjunto de prueba (Malviya, Dangi, Jadhav & Kishore, 2025).

Finalmente, se implementó un ajuste del umbral de decisión con el objetivo de optimizar la detección de la clase positiva (Esposito, Landrum, Schneider, Stiefl & Riniker, 2021). En los problemas de clasificación, el umbral estándar suele ser 0.5. Sin embargo, en un escenario clínico y con clases desbalanceadas, ese valor no es el más adecuado, ya que puede disparar el número de falsos negativos o positivos (Rajaraman, Ganesan & Antani, 2021). Para enfrentar este reto, se evaluó múltiples umbrales y se eligió aquel que dió el mejor balance entre precisión y exhaustividad. El criterio principal fue seleccionar el valor que maximizara el F_1-score (Abed & Akbaş, 2024). Este procedimiento permitió encontrar un umbral óptimo, que resultó ser ligeramente superior al convencional. Se mejoró la capacidad del modelo para discriminar los casos de prevalencia de anemia clínicamente relevantes (Kattou, Montandrau, Rekik, Delentdecker, Brini, Zannis & Beaussier, 2022). De esta manera el sistema mostró un equilibrio más adecuado entre detectar correctamente a los pacientes afectados y reducir los errores de clasificación, lo que aumenta su utilidad práctica en contextos de salud (Ueda & Morishita, 2023).

3.5. Evaluación

El modelo se evaluó utilizando métricas ampliamente reconocidas en el aprendizaje supervisado (Rainio, Teuho & Klén, 2024), siguiendo la flujo detallado en la Figure 4. La elección de estas métricas es clave, ya que permiten valorar no solo la exactitud global, sino también la capacidad del modelo para identificar correctamente los casos positivos y negativos de prevalencia de anemia (Gómez, Urueta, Salas-Álvarez, Riaño & Ramírez-González, 2025).

En primer lugar, se consideró la exactitud (accuracy), definida como la proporción de predicciones correctas sobre el total de instancias (202, 2024).

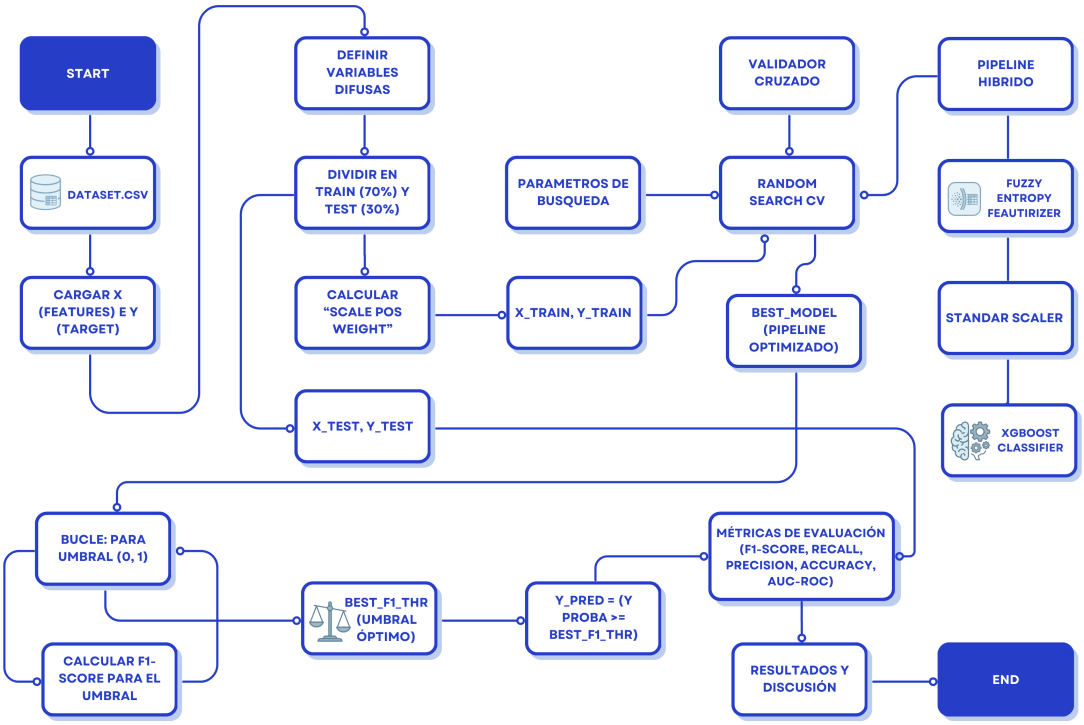


Figure 4: Diagrama de flujo de la metodología propuesta. Este diagrama ilustra las etapas clave, incluyendo la definición de variables difusas, el preprocesamiento de los datos, el proceso de optimización del pipeline híbrido mediante búsqueda aleatoria con validación cruzada, y el bucle de ajuste del umbral para maximizar el F_1-score.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN},$$

donde TP y TN representan los verdaderos positivos y negativos, mientras que FP y FN corresponden a falsos positivos y falsos negativos respectivamente (Desai, 2023).

Dado que la exactitud puede ser poco informativa en contextos de clases desbalanceadas, se empleó también el F_1-score, métrica que combina precisión y exhaustividad en un solo valor, calculado como una media que da mayor peso al valor más bajo entre ambas (Hemmatian, Hajizadeh & Nazari, 2025). Esto permite reflejar de forma más equilibrada el desempeño del modelo en escenarios donde una de las métricas puede ser desproporcionadamente alta frente a la otra (Tanaka & Yon, 2024).

$$\text{Precision} = \frac{TP}{TP + FP}, \qquad \text{Recall} = \frac{TP}{TP + FN},$$

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

El F_1-score es especialmente relevante en problemas clínicos donde los falsos negativos representan un riesgo significativo al no detectar a un niño con prevalencia de anemia (Saputra, Sunat & Ratnaningsih, 2023).

Adicionalmente, se emplearon métricas basadas en curvas de evaluación (Li, 2024). El Área bajo la curva ROC (ROC-AUC) mide la capacidad discriminativa del modelo entre clases, mientras que el Área bajo la curva Precisión-Recall (PR-AUC) resulta más informativa en escenarios con desbalance de clases, al centrarse en la proporción de positivos detectados correctamente (Imani, Beikmohammadi & Arabnia, 2025).

Finalmente, se utilizó la matriz de confusión como herramienta visual para interpretar el desempeño del clasificador, mostró explícitamente las cantidades de TP, FP, FN y TN. Esta matriz permite comprender qué tipo de errores comete el modelo y sirve como complemento interpretativo a las métricas cuantitativas (Chicco & Jurman, 2022).

La combinación de estas métricas ofrece una evaluación integral del modelo, balanceando la visión global de la exactitud con medidas sensibles al desbalance y al costo clínico de los errores (Garza, Williams, Ounpraseuth, Hu, Lee, Snowden, Walden, Simon, Devlin, Young & Zozus, 2024).

Table 5
Comparación de rendimiento entre diferentes configuraciones de modelos de referencia y el modelo propuesto con entropía difusa.

Modelo	Accuracy	Recall	Precisión	F_1-score	ROC-AUC	PR-AUC
RF (Sin tunear)	0.837	0.739	0.776	0.757	0.945	0.907
XGBoost (Sin tunear)	0.829	0.743	0.755	0.749	0.961	0.933
RF (Tuneado)	0.806	0.619	0.771	0.687	0.916	0.862
XGBoost (Tuneado)	0.822	0.642	0.800	0.712	0.941	0.904
RF + Entropía Difusa	0.897	0.840	**0.870**	0.850	0.942	0.900
XGBoost + Entropía Difusa	0.902	**0.884**	0.841	0.862	0.959	0.931
Modelo propuesto (XGBoost + Entropía Difusa + ajuste de umbral)	**0.909**	0.883	0.858	**0.870**	**0.962**	**0.937**

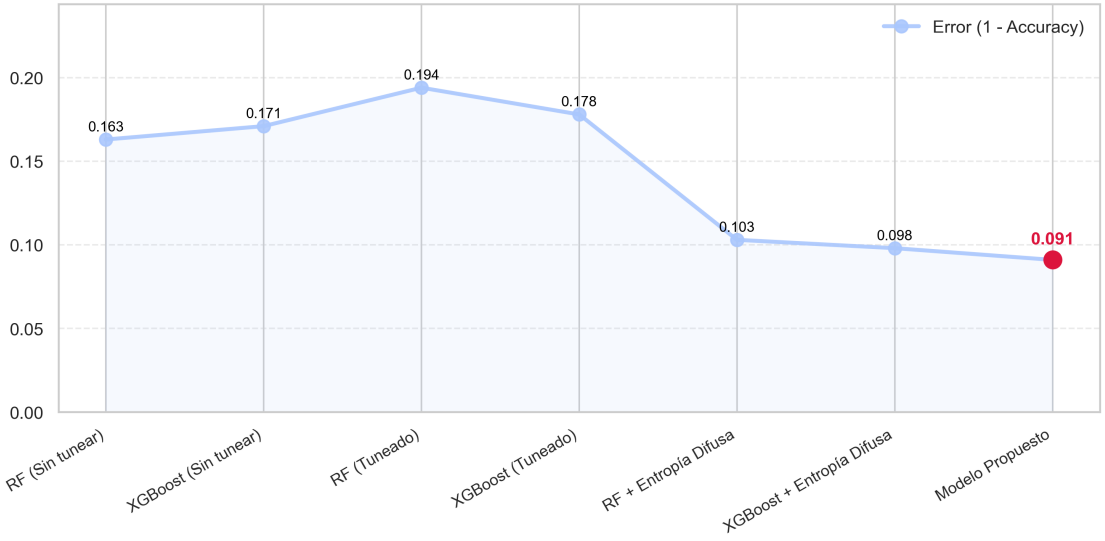


Figure 5: El Modelo Propuesto que integra XGBoost, entropía difusa y un ajuste óptimo del umbral, consigue el menor error, lo que subraya la contribución de la entropía difusa para la gestión de la incertidumbre y la mejora del rendimiento predictivo.

4. Resultados y discusión

El modelo propuesto fue evaluado en comparación con siete configuraciones de referencia, cuatro modelos base (Random Forest y XGBoost, con y sin ajuste de hiperparámetros), dos modelos extendidos con entropía difusa y finalmente la versión final con ajuste de umbral. Los resultados se presentan en la Table 5, mientras que la Figure 5 ofrece una comparación visual directa del error 1 - Accuracy de cada configuración, complementados con las curvas ROC y PR que ilustran el desempeño de cada configuración.

Los modelos base mostraron un desempeño moderado, Random Forest sin ajuste alcanzó un accuracy de 0.837 y un F_1-score de 0.757, mientras que XGBoost sin ajuste obtuvo 0.829 de accuracy y 0.749 de F_1-score. Con tuning básico los resultados fueron más bajos, con un F_1-score de 0.687 en Random Forest y 0.712 en XGBoost. Estos valores confirman que ajustes superficiales de hiperparámetros no bastan para explotar el potencial de los algoritmos en datos clínicos.

La incorporación de entropía difusa mejoró de forma consistente las métricas: Random Forest + Entropía Difusa alcanzó un accuracy de 0.897 y un F_1-score de 0.850, mientras que XGBoost + Entropía Difusa obtuvo un accuracy de 0.902 y un F_1-score de 0.862. Finalmente, el modelo propuesto (XGBoost + Entropía Difusa + ajuste de umbral) logró los mejores resultados globales accuracy de 0.909, recall de 0.883, precisión de 0.858, F_1-score de 0.870, ROC-AUC de 0.962 y PR-AUC de 0.937. Estos valores confirman la robustez del modelo para discriminar entre casos positivos y negativos incluso en escenarios de desbalance.

En compración con investigaciones previas realizadas en el Perú, los resultados representan una mejora sustancial. Marcos Valdez et al. (2023) reportó una exactitud del 70% en la predicción de anemia en menores de cinco años; Ramirez & Karina empleó Random Forest sobre datos ENDES en niños de 6 a 59 meses y alcanzó una sensibilidad de 65.9% y especificidad de 63.6%; mientras que Mayhua et al. (2022) obtuvo una precisión del 74.5% con árboles de decisión en Arequipa. De igual forma, Céspedes Panduro (2022) concluyó que los algoritmos tradicionales no superaban niveles aceptables de desempeño. Frente a estos antecedentes, el modelo propuesto supera en más de 15 puntos porcentuales

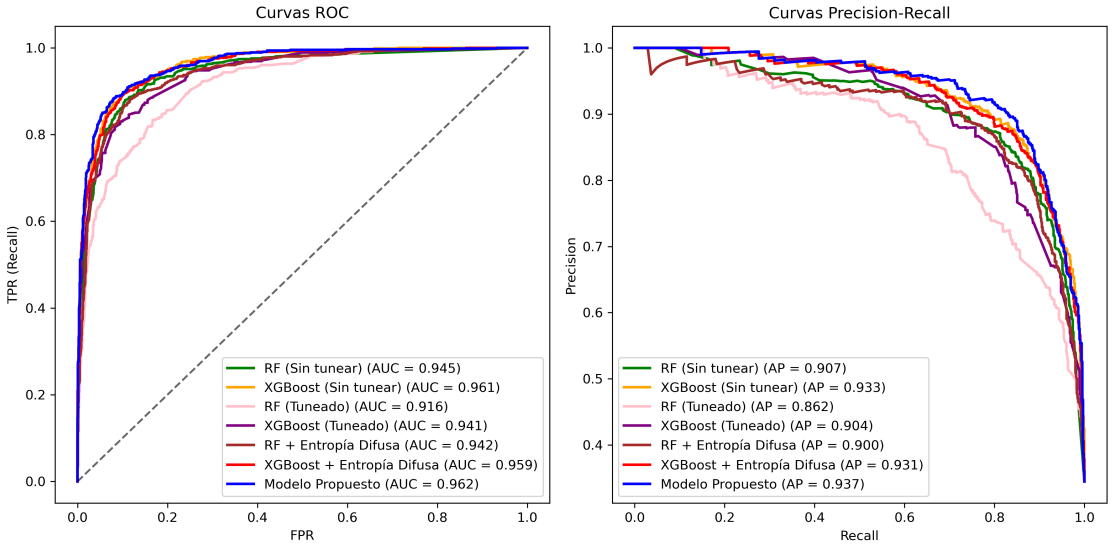


Figure 6: Curvas ROC y PR de los modelos evaluados. El modelo propuesto (XGBoost + entropía difusa + ajuste de umbral) presenta las mayores áreas bajo la curva en ambas métricas, lo que confirma su superior capacidad discriminativa.

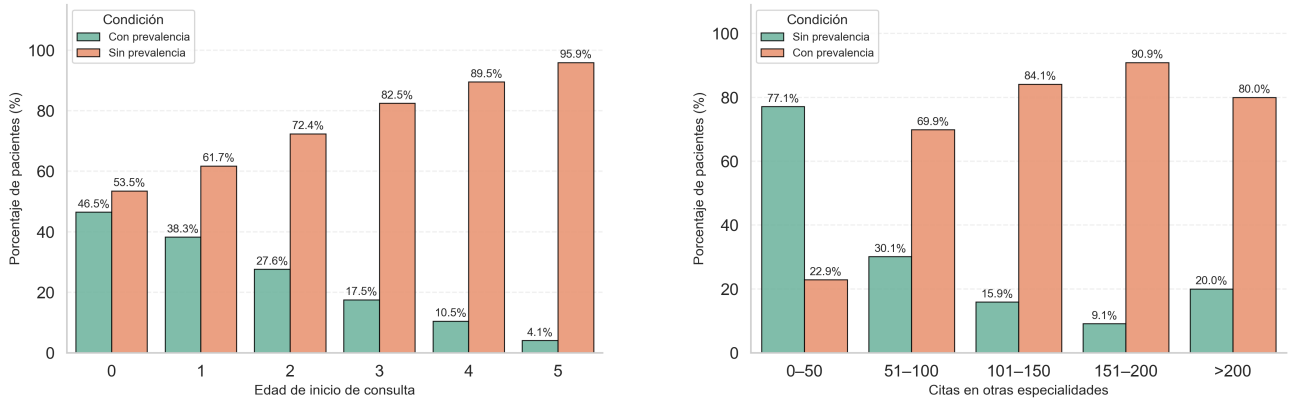


Figure 7: Distribución de pacientes según edad al inicio de consulta y número de otras citas, segmentada por condición (sin prevalencia vs con prevalencia). Ambas gráficas muestran porcentajes de pacientes, permitiendo comparar la distribución de prevalencia de anemia en diferentes variables.

la exactitud reportada, al tiempo que optimiza el balance entre sensibilidad y precisión, aspecto crítico para el tamizaje de anemia infantil.

En comparación con investigaciones previas realizadas en el Perú, los resultados del modelo representan una mejora sustancial. Estudios anteriores, como los de Marcos Valdez et al. (2023), reportaron una exactitud del 70% en la predicción de anemia en menores de cinco años; Ramirez & Karina empleó Random Forest sobre datos ENDES en niños de 6 a 59 meses alcanzando una sensibilidad de 65.9% y especificidad de 63.6%; mientras que Mayhua et al. (2022) obtuvo una precisión del 74.5% con árboles de decisión en Arequipa. Céspedes Panduro (2022) concluyó que los algoritmos tradicionales no superaban niveles aceptables de desempeño. Frente a estos antecedentes, el modelo supera en más de 15 puntos porcentuales la exactitud reportada, mejorando el balance entre sensibilidad y precisión, lo que es crítico para el tamizaje de anemia infantil.

Una de las contribuciones clave de la investigación es la incorporación de un conjunto más amplio de variables, más allá de los indicadores clínicos tradicionales como hemoglobina, peso y talla. Como se muestra en la Table 3, estas variables incluyen características demográficas, historial de consultas y factores de seguimiento médico. Esta integración permite evaluar la prevalencia de anemia desde una perspectiva más holística y refleja la complejidad real de la atención infantil.

En la Figure 7 se encontró hallazgos relevantes dentro del análisis, los pacientes que inician su atención a edades más tempranas presentan una mayor probabilidad de desarrollar anemia y los pacientes que desarrollan prevalencia de anemia tienden a tener una gran cantidad de citas en otras especialidades lo cual evidencia que pueden tener otras

Algorithm 1 Fuzzy Entropy–XGBoost Classification Model

Input: Dataset D, list of fuzzy variables V_{fuzzy}
Output: Optimized model $\mathcal{M}_{opt}$ and evaluation metrics

function FUZZYENTROPYFEATURIZER(X, V_{fuzzy})
  for each variable v in V_{fuzzy} **do**
    Define universe $U_v = [0, v_{max}]$ with step Δ
    Compute membership functions:
      $\mu_{low}, \mu_{medium}, \mu_{high} \leftarrow \text{Triangular}(U_v)$
  end for
  for each record $x_i \in X$ **do**
    for each variable v in V_{fuzzy} **do**
      Compute fuzzy memberships μ_b, μ_m, μ_a
      Normalize: $P = (\mu + \varepsilon)/(\sum \mu + 3\varepsilon)$
      Compute entropy $E_v = -\frac{\sum P \log P}{\log 3}$
    end for
    x_i[entropy] $\leftarrow$ mean(E_v)
  end for
  return X with added fuzzy-entropy feature
end function

Split D into training and testing sets: $(X_{train}, X_{test}, y_{train}, y_{test})$
Compute class weight $w = \frac{neg}{pos}$
Build pipeline:
  $[FuzzyEntropyFeaturizer, StandardScaler, XGBoost(w)]$
Perform grid search on hyperparameters with 5-fold cross-validation
Select best model $\mathcal{M}_{opt}$ based on F1-score
Predict probabilities on X_{test}: y_{proba}
for each threshold $t \in [0, 1]$ **do**
  Compute $F1(t)$, $Precision(t)$, $Recall(t)$
end for
Select t^* maximizing $F1(t)$
Compute final predictions $y_{pred} = (y_{proba} \geq t^*)$
Evaluate metrics: Accuracy, ROC-AUC, PR-AUC, Confusion Matrix
return $\mathcal{M}_{opt}$ and performance metrics

enfermedades las cuales permitan que esos pacientes continuen con anemia a pesar de los tratamientos, esto resalta la necesidad de fortaleces las estrategias de deteccion e intervencion mas tempranas. Estos resultados evidencian que los enfoques actuales solo se basan unicamente en indicadores clinicos basicos, es por eos que se hace necesario incorporar factores adicionales que permitan una comprensión más integral de las causas y condiciones asociadas a la prevalencia de la anemia infantil. Finalmente, los resultados no solo confirman la utilidad de los modelos de aprendizaje automático en contextos clínicos, sino que también destacan la originalidad del estudio al integrar factores adicionales, ofreciendo un enfoque más completo que las investigaciones previas y aportando herramientas más robustas para la identificación factores que contribuyen a la prevalencia de anemia infantil.

4.1. Limitaciones

El modelo se desarrolló utilizando registros clínicos que abarcan variables sociodemográficas básicas, características de las consultas médicas y mediciones clínicas como peso, talla y hemoglobina. Esto permitió un análisis sólido de los factores asociados a la anemia infantil, aunque todavía no se incluyeron aspectos como el nivel educativo, la dieta o las condiciones del hogar, que también influyen en la salud infantil. Más que una limitación, esto representa una oportunidad de crecimiento. El modelo puede ampliarse fácilmente para integrar este tipo de factores en el futuro, lo que haría sus predicciones más completas y útiles en distintos contextos. Asimismo, el enfoque propuesto puede aplicarse a bases de datos más grandes y de diferentes regiones, lo que ayudaría a confirmar su utilidad en poblaciones diversas. Esta capacidad de adaptarse y crecer muestra que el modelo no solo es útil en el contexto actual, sino que también tiene potencial para convertirse en una herramienta práctica y escalable dentro de la salud pública.

5. Conclusiones

El objetivo principal de este estudio fue desarrollar un modelo de clasificación basado en entropía difusa para la identificación de los factores que contribuyen a la prevalencia de anemia en niños peruanos. El pipeline híbrido integra tres etapas clave, un preprocesamiento avanzado mediante entropía difusa para capturar la incertidumbre inherente a los datos, un clasificador robusto XGBoost optimizado, y una etapa crucial de post-procesamiento que es un bucle de optimización de umbral explícitamente diseñado para poder maximizar el F_1-score en lugar del accuracy. Los resultados confirman la eficacia del enfoque propuesto, el modelo final, que integra XGBoost, entropía difusa y ajuste de umbral, alcanzó un accuracy del 90.9%, un recall de 88.3%, una precisión de 85.8% y un F_1-score de 0.870. Además, los valores del área bajo la curva ROC-AUC de 0.962 y PR-AUC de 0.937 muestran un desempeño estadísticamente significativo, "la entropía difusa por sí sola produce un salto drástico en el rendimiento, y el modelo propuesto final supera a todas las configuraciones previas al alcanzar el error más bajo de 0.091", lo que valida la robustez del modelo para detectar los casos de prevalencia de anemia infantil. Estos hallazgos demuestran que incorporar la entropía difusa es clave para gestionar la incertidumbre y mejorar la capacidad predictiva frente a los modelos tradicionales. Esto, representa un avance metodológico en el análisis de datos clínicos. Además, el modelo destaca por su potencial de escalabilidad y replicabilidad. Su diseño permite adaptarlo a otras bases de datos clínicas, ampliarlo con factores sociodemográficos y socioeconómicos, e implementarlo en distintos contextos epidemiológicos. Esto abre la puerta al diseño de políticas públicas más focalizadas y a estrategias de intervención en salud infantil con mayor sustento científico. Como trabajo futuro, se recomienda validar el modelo en muestras más amplias y diversas. También será crucial incorporar variables adicionales que ayuden a explicar mejor los determinantes

sociales y culturales de la anemia. Al hacerlo, el enfoque no solo contribuirá al conocimiento científico, sino que se consolidará como una herramienta práctica para fortalecer la salud pública y mejorar los indicadores de nutrición infantil en el Perú.

turnitin Página 20 de 23 - Engrega de integridad Identificador de la entrega trn:oid:::29566:543848416

CRediT authorship contribution statement

Jhon-Vilcarana: . **Jorge-Quispe:** Recolección de datos. **Hubert-Falcon:** Análisis, Revisión. **Soria-Quijaite:** Análisis estadístico. **Saboya-Rios:** Supervisión, Edición.

References

, 2024. Proceedings of the 2024 workshop on human-in-the-loop data analytics. Proceedings of the 2024 Workshop on Human-In-the-Loop Data Analytics doi:10.1145/3665939.

Abed, M.S.A., Akbaş, A., 2024. An approach in melanoma skin cancer segmentation with bat optimization algorithm. International Journal of Imaging Systems and Technology 34. doi:10.1002/ima.23119.

Ajmer, S., 2020. Childhood anemia in bihar, india: Patterns and determinants , 01–09.

Ali, M., Mekhilef, S., Nukman, Y., Razak, B.A., 2022. Laser simulator logic: A novel inference system for highly overlapping of linguistic variable in membership functions. J. King Saud Univ. Comput. Inf. Sci. 34, 8019–8040. doi:10.1016/j.jksuci.2022.07.017.

Arya, V., Kumar, S., 2020. Knowledge measure and entropy: a complementary concept in fuzzy theory. Granular Computing 6, 631 – 643. doi:10.1007/s41066-020-00221-7.

Ayyanar, M., Jeganathan, S., Parthasarathy, S., Jayaraman, V., Lakshminarayanan, A., 2022. Predicting the cardiac diseases using selectkbest method equipped light gradient boosting machine. 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI) , 117–122doi:10.1109/ICOEI53556.2022.9777224.

Bargues, J.L.F., Cruz, M., Álvarez, L., Prieto-Gómez, R.E., 2017. Una metodología basada en números difusos para la evaluación de criterios evaluables mediante un juicio de valor =a methodology based on fuzzy numbers for the evaluation of evaluable criteria through a value judgment 3, 49–57. doi:10.20868/ADE.2017.3571.

Barros, R., Costa, E., Araújo, J., 2021. Evaluation of classifiers for non-technical loss identification in electric power systems. International Journal of Electrical Power & Energy Systems 132, 107173. doi:10.1016/J.IJEPES.2021.107173.

Bergstra, J., Bengio, Y., 2012. Random search for hyper-parameter optimization. J. Mach. Learn. Res. 13, 281–305. doi:10.5555/2503308.2188395.

Bhardwaj, N., Sharma, P., 2021. An advanced uncertainty measure using fuzzy soft sets: Application to decision-making problems. Big Data Min. Anal. 4, 94–103. doi:10.26599/BDMA.2020.9020020.

Bhushan, S., Kumar, A., Pokhrel, R., 2024. Synthetic imputation methods for domain mean under simple random sampling. Franklin Open doi:10.1016/j.fraope.2024.100101.

Bitew, F.H., Sparks, C.S., Nyarko, S.H., 2022. Machine learning algorithms for predicting undernutrition among under-five children in ethiopia. Public Health Nutrition 25, 269–280. doi:10.1017/S1368980021004262.

Boadh, R., Chaudhary, K., Dahiya, M., Dogra, N., Rathee, S., Kumar, A., Rajoria, Y.K., 2022. Analysis and investigation of fuzzy expert system for predicting the child anaemia. Materials Today: Proceedings doi:10.1016/j.matpr.2022.01.094.

Buscemi, F., Schindler, J., Šafránek, D., 2022. Observational entropy, coarse-grained states, and the petz recovery map: information-theoretic properties and bounds. New Journal of Physics 25. doi:10.1088/1367-2630/accd11.

Camacho, O., Iglesias, E., Herrera, M., Aboukheir, H., 2021. Fuzzy logic-based control: From fundamentals to applications control .

Cao, Z., Lin, C.T., 2018. Inherent fuzzy entropy for the improvement of eeg complexity evaluation. IEEE Transactions on Fuzzy Systems 26, 1032–1035. doi:10.1109/TFUZZ.2017.2666789.

Céspedes Panduro, B., 2022. Aplicación del algoritmo" random forest" para un modelo de clasificación sobre la tenencia de anemia de niños del perú .

Chen, T., Guestrin, C., 2016. Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, New York, NY, USA. p. 785–794. URL: https://doi.org/10.1145/2939672.2939785, doi:10.1145/2939672.2939785.

Chen, T., Shang, C., Su, P., Keravnou-Papailiou, E., Zhao, Y., Antoniou, G., Shen, Q., 2021. A decision tree-initialised neuro-fuzzy approach for clinical decision support. Artificial intelligence in medicine 111, 101986. doi:10.1016/j.artmed.2020.101986.

Chicco, D., Jurman, G., 2022. An invitation to greater use of matthews correlation coefficient in robotics and artificial intelligence. Frontiers in Robotics and AI 9. doi:10.3389/frobt.2022.876814.

Cho, Y., Zhan, X., Ghosh, D., 2022. Nonlinear predictive directions in clinical trials. Comput. Stat. Data Anal. 174, 107476. doi:10.1016/j.csda.2022.107476.

Chotikunnan, P., Chotikunnan, R., Pititheeraphab, Y., Puttasakul, T., Wongkamhang, A., Thongpance, N., 2025. Comparative analysis of fuzzy membership functions for step and smooth input tracking in a 3-axis robotic manipulator. Journal of Fuzzy Systems and Control doi:10.59247/jfsc.v3i1.278.

of Computer Science, I.J., Engineering, 2020. Fuzzy logic based model using triangular membership functions. IJRDO Journal of Computer Science and Engineering 6, 1–6. URL: https://www.ijrdo.org/index.php/cse/article/download/659/609. accedido: 23-Sep-2025.

de la Cruz Huayanay, A., Bazán, J.L., Russo, C., 2024. Performance of evaluation metrics for classification in imbalanced data. Comput. Stat. 40, 1447–1473. doi:10.1007/s00180-024-01539-5.

kassab Córdova, A.A., Mendez-Guerra, C., Quevedo-Ramirez, A., Espinoza, R., Enriquez-Vera, D., Robles-Valcárcel, P., 2022. Rural and urban disparities in anemia among peruvian children aged 6-59 months: a multivariate decomposition and spatial analysis. Rural and remote health 22 2, 6936. doi:10.22605/RRH6936.

Davalos Carrera, J.I., Cortez Paredes, K.M., 2025. Modelos de machine learning para la detección temprana de deficiencias nutricionales en la infancia. B.S. thesis.

De Luca, A., Termini, S., 1972. A definition of a nonprobabilistic entropy in the setting of fuzzy sets theory. Information and Control 20, 301–312. doi:10.1016/S0019-9958(72)90199-4.

Desai, D., 2023. Dev's formulae for diagnostic test accuracy meta-analysis doi:10.1101/2023.03.27.23287186.

Dhanka, S., Maini, S., 2024. A hybridization of xgboost machine learning model by optuna hyperparameter tuning suite for cardiovascular disease classification with significant effect of outliers and heterogeneous training datasets. International journal of cardiology 420, 132757. doi:10.1016/j.ijcard.2024.132757.

Esposito, C., Landrum, G., Schneider, N., Stiefl, N., Riniker, S., 2021. Ghost: Adjusting the decision threshold to handle imbalanced data in machine learning. Journal of chemical information and modeling doi:10.1021/acs.jcim.1c00160.

García, J.E., Sandoval, A.V., Vega, J.E.S., Herrera, B.E.G., 2021. Comparación del nivel de desempeño de una competencia usando tres instrumentos, dos basados en rúbrica y otro basado en lógica difusa: A comparison of the level of competency using three instruments; two rubric based instruments and a fussy logic-based instrument 2, 123–145. doi:10.46990/RELEP.2020.2.4.245.

Garza, M.Y., Williams, T., Ounpraseuth, S., Hu, Z., Lee, J., Snowden, J.N., Walden, A.C., Simon, A.E., Devlin, L.A., Young, L.W., Zozus, M.N., 2024. Error rates of data processing methods in clinical research: A systematic review and meta-analysis of manuscripts identified through pubmed. International journal of medical informatics 195, 105749. doi:10.1016/j.ijmedinf.2024.105749.

Gurudatha, S., Majhi, R., 2024. Robust machine learning methods for prediction of childhood anemia – a case of the empowered action group states of india. 2024 2nd International Conference on Recent Advances in Information Technology for Sustainable Development (ICRAIS) , 188–193doi:10.1109/ICRAIS62903.2024.10811745.

Gómez, J., Urueta, C.P., Salas-Álvarez, D., Riaño, V.H., Ramírez-González, G.A., 2025. Anemia classification system using machine learning. Informatics 12, 19. doi:10.3390/informatics12010019.

Harshavardhanan, S., Nene, M., 2020. Capturing and modeling uncertainty in prognostics and health management using machine learning. 2020 5th

International Conference on Communication and Electronics Systems (ICCES) , 1235–1241doi:10.1109/ICCES48766.2020.9137918.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed., Springer. doi:10.1007/978-0-387-84858-7.

Hemmatian, J., Hajizadeh, R., Nazari, F., 2025. Addressing imbalanced data classification with cluster-based reduced noise smote. PLOS ONE 20. doi:10.1371/journal.pone.0317396.

Hong, C., Moon, Y., Choi, H.W., Kang, E.K., Ye, S.H., Kim, J.K., Kim, Y.H., Cho, H., Choi, H., eun Lee, D., Choi, Y., Kim, T.M., Heo, S.G., Han, N., Hong, K.M., 2025. Abstract 5057: Enhancing clinical applications by evaluation of sensitivity and specificity in whole exome sequencing. Cancer Research doi:10.1158/1538-7445.am2025-5057.

Hoque, M., Khan, M., Chowdhury, 2021. Prevalence of anemia among primary school children in mymensingh: A school-based cross-sectional study .

Huang, J., Lu, X., Yang, J., 2024. Sparse membership affinity lasso for fuzzy clustering. IEEE Transactions on Fuzzy Systems 32, 1553–1563. doi:10.1109/TFUZZ.2023.3327688.

Imani, M., Beikmohammadi, A., Arabnia, H., 2025. Comprehensive analysis of random forest and xgboost performance with smote, adasyn, and gnus under varying imbalance levels. Technologies doi:10.3390/technologies13030088.

Incorvaia, G., Hond, D., Asgari, H., 2024. Uncertainty quantification of machine learning model performance via anomaly-based dataset dissimilarity measures. Electronics doi:10.3390/electronics13050939.

Infobae, 2025. Anemia infantil en menores de 3 años subió a 43,7% y la desnutrición crónica en menores de 5 años llegó al 12,1%, según la endes 2024. https://www.infobae.com/peru/2025/04/02/anemia-infantil-en-menores-de-3-anos-subio-a-437-y-la-desnutricion-cronica-en-menores-de-5-anos-llego-al-121-segun-la-endes-2024/. Consultado: 2 de mayo de 2025.

Jalaludin, M.Y., Kiau, H.B., Hasim, S., Lee, W.K., Low, A., Kazim, N.H.N., Hoi, J.T., Taher, S.W., 2025. A noninvasive approach to assess the prevalence of and factors associated with anemia risk in malaysian children under three years of age: Cross-sectional study. JMIR Pediatrics and Parenting 8. doi:10.2196/58586.

Jin, J., Garg, H., 2023. Intuitionistic fuzzy three-way ranking-based topsis approach with a novel entropy measure and its application to medical treatment selection. Adv. Eng. Softw. 180, 103459. doi:10.1016/j.advengsoft.2023.103459.

Jun, Y.B., Xin, X.L., 2019. Complex fuzzy sets with application in bck/bci-algebras. Bulletin of the Section of Logic 48, 173–188. doi:10.18778/0138-0680.48.3.02.

Kaczmarek-Majer, K., Rutkowska, A., Hryniewicz, O., 2022. Evolving membership functions in fuzzy linguistic summarization .

Karl, F., Pielok, T., Moosbauer, J., Pfisterer, F., Coors, S., Binder, M., Schneider, L., Thomas, J., Richter, J., Lang, M., Garrido-Merch'an, E.C., Branke, J., Bischl, B., 2022. Multi-objective hyperparameter optimization in machine learning—an overview. ACM Transactions on Evolutionary Learning and Optimization 3, 1 – 50. doi:10.1145/3610536.

Kattou, F., Montandrau, O., Rekik, M., Delentdecker, P., Brini, K., Zannis, K., Beaussier, M., 2022. Critical preoperative hemoglobin value to predict anemia-related complications after cardiac surgery. Journal of cardiothoracic and vascular anesthesia doi:10.1053/j.jvca.2022.01.013.

van Lambalgen, M., s.f. Script fuzzy logic. https://www.blutner.de/uncert/Fuzzy.pdf. Material de referencia en teoría de conjuntos y lógica difusa.

Li, J., 2023. Asymptotics of k-fold cross validation. J. Artif. Intell. Res. 78, 491–526. doi:10.1613/jair.1.13974.

Li, J., 2024. Area under the roc curve has the most consistent evaluation for binary classification. PLOS ONE 19. doi:10.1371/journal.pone.0316019.

Lin, G., Zhang, Y., 2024. Fuzzy neural logic reasoning for robust classification. ACM Transactions on Knowledge Discovery from Data 19, 1 – 29. doi:10.1145/3704728.

Liu, C., Li, K., Zhao, L., Liu, F., Zheng, D., Liu, C., Liu, S., 2013. Analysis of heart rate variability using fuzzy measure entropy. Computers in biology and medicine 43 2, 100–8. doi:10.1016/j.compbiomed.2012.11.005.

Ma, B., Meng, F., Yan, G., Yan, H., Chai, B., Song, F., 2020. Diagnostic classification of cancers using extreme gradient boosting algorithm and multi-omics data. Computers in biology and medicine 121, 103761. doi:10.1016/j.compbiomed.2020.103761.

Ma, Q., 2024. Research on the application of health big data based on semi-supervised learning algorithm, in: 2024 3rd International Conference for Innovation in Technology (INOCON), pp. 1–4. doi:10.1109/INOCON60754.2024.10512045.

Malviya, L., Dangi, R., Jadhav, A., Kishore, J., 2025. Optimization-based hyperparameter tuning using extra trees to classify type-2-diabetes mellitus. 2025 International Conference on Computational, Communication and Information Technology (ICCCIT) , 465–470doi:10.1109/ICCCIT62592.2025.10928114.

Marcos Valdez, A.J., Navarro Ortiz, E.G., Quinteros Peralta, R.E., Tirado Julca, J.J., Valentín Ricaldi, D.F., Calderón-Vilca, H.D., 2023. Aprendizaje automático para predicción de anemia en niños menores de 5 años mediante el análisis de su estado de nutrición usando minería de datos. Computación y Sistemas 27, 749–768.

Mariadoss, S., Augustin, F., 2023. Fuzzy entropy dematel inference system for accurate and efficient cardiovascular disease diagnosis. Computer methods in biomechanics and biomedical engineering , 1–32doi:10.1080/10255842.2023.2245518.

Mayhua, I.A., Huamani, A.C., Tovar, A.V., Challa, M.R., 2022. Aplicación de los árboles de decisión en el diagnóstico de anemia en niños de la ciudad de arequipa. Innovación y Software 3, 26–39.

Meitei, A., Saini, A., Mohapatra, B., et al., 2022. Predicting child anaemia in the north-eastern states of india: a machine learning approach. International Journal of System Assurance Engineering and Management 13, 2949–2962. URL: https://doi.org/10.1007/s13198-022-01765-4, doi:10.1007/s13198-022-01765-4.

Meng, B., Li, F., Yang, F., Gao, Q., 2025. Centroid-free k-means with balanced clustering. IEEE Signal Processing Letters 32, 1191–1195. doi:10.1109/LSP.2025.3547665.

Ministerio de Salud (Perú), 2024. Minsa presentó los servicios para la prevención y control de la anemia en niños menores de 3 años, mujeres adolescentes y gestantes. https://www.gob.pe/institucion/minsa/noticias/934442-minsa-presento-los-servicios-para-la-prevencion-y-control-de-la-anemia-en- Accedido el 21 de octubre de 2025.

Muthalaly, R., Kwong, R., John, R., van der Geest, R.J., Tao, Q., Schaeffer, B., Tanigawa, S.I., Nakamura, T., Kaneko, K., Tedrow, U., Stevenson, W., Epstein, L., Kapur, S., Zei, P.C., Koplan, B., 2019. Left ventricular entropy is a novel predictor of arrhythmic events in patients with dilated cardiomyopathy receiving defibrillators for primary prevention. JACC. Cardiovascular imaging doi:10.1016/j.jcmg.2018.07.003.

Mutlaq, H.S., ALOTAIBI, K.A.T., HUSAYKAN, I.M., Alotaibi, S.S.F., Alotaibi, F.H.H., Alotibi, F.A.A., Alotaibi, E.S.E., Abualssayl, A.A., MUBARAK, A.M., Almutairi, M.H., BINNWEJIM, M.S., Khabrani, M.A.A., 2024. Anemia: An updated review for healthcare professionals. Journal of Ecohumanism 3, 13225–. URL: https://doi.org/10.62754/joe.v3i8.6232, doi:10.62754/joe.v3i8.6232.

Nambiema, A., Robert, A., Yaya, I., 2019. Prevalence and risk factors of anemia in children aged from 6 to 59 months in togo: analysis from togo demographic and health survey data, 2013–2014. BMC public health 19, 1–9.

Nasfi, R., de Tré, G., Bronselaer, A., 2025. Improving data cleaning by learning from unstructured textual data. IEEE Access 13, 36470–36491. doi:10.1109/ACCESS.2025.3543953.

Nimmala, S., Mahendar, M., Manasa, P., Lakshmi, H.N., Kiran, M.A., Raghavendra, C., 2024. Fuzzy-enhanced xgboost model for classifying kidney disease severity. 2024 4th International Conference on Soft Computing for Security Applications (ICSCSA) , 21–26doi:10.1109/ICSCSA64454.2024.00011.

Pandey, K., Mishra, A., Rani, P., Ali, J., Chakrabortty, R., 2023. Selecting features by utilizing intuitionistic fuzzy entropy method. Decision Making: Applications in Management and Engineering doi:10.31181/dmame07012023p.

Park, S.B., Oh, S.K., Pedrycz, W., 2022. Feature data-driven-reinforced fuzzy radial basis function neural network classifier with the aid of preprocessing techniques and particle swarm optimization. Soft Computing 27, 15443–15462. doi:10.1007/s00500-023-09124-6.

Pullakhandam, S., McRoy, S., 2024. Classification and explanation of iron deficiency anemia from complete blood count data using machine learning. BioMedInformatics 4, 661–672.

Qiao, G., Shen, Z., Duan, S., Wang, R., He, P., Zhang, Z., Dai, Y., Li, M., Chen, Y., Li, X., Zhao, Y., Liu, Z., Yang, H., Zhang, R., Guan, S., Sun, J., 2024. Associations of urinary metal concentrations with anemia: A cross-sectional study of chinese community-dwelling elderly. Ecotoxicology and Environmental Safety 270, 115828. URL: https://doi.org/10.1016/j.ecoenv.2023.115828, doi:10.1016/j.ecoenv.2023.115828.

Rainio, O., Teuho, J., Klén, R., 2024. Evaluation metrics and statistical tests for machine learning. Scientific Reports 14. doi:10.1038/s41598-024-56706-x.

Rajaraman, S., Ganesan, P., Antani, S., 2021. Deep learning model calibration for improving performance in class-imbalanced medical image classification tasks. PLoS ONE 17. doi:10.1371/journal.pone.0262838.

Ramirez, A., Karina, Q., . Random forest como método alternativo para determinar factores asociados a la anemia en niños y niñas entre 6 a 59 meses en el perú, según la encuesta demográfica y salud familiar–endes 2022 .

Ramos, D., Franco-Pedroso, J., Lozano-Diez, A., González-Rodríguez, J., 2018. Deconstructing cross-entropy for probabilistic binary classifiers. Entropy 20. doi:10.3390/e20030208.

Rivera, J., Cardenas, D., Castillo-Sequera, J., Wong, L., 2024. Early prediction model for anemia in infants using clinical data from perú applying supervised machine learning algorithms. 2024 10th International Conference on Optimization and Applications (ICOA) , 1–7doi:10.1109/ICOA62581.2024.10754254.

Rosas-Jiménez, C., Tercan, E., Horstick, O., Igboegwu, E., Dambach, P., Louis, V., Winkler, V., Deckert, A., 2022. Prevalence of anemia among indigenous children in latin america: a systematic review. Revista de Saúde Pública 56. doi:10.11606/s1518-8787.2022056004360.

Ruby, A., Chandran, G.C., Jain, S., Chaithanya, B., Patil, R., 2023. Rffe – random forest fuzzy entropy for the classification of diabetes mellitus. AIMS Public Health 10, 422 – 442. doi:10.3934/publichealth.2023030.

Saeed, M.H., Hama, J., 2023. Cardiac disease prediction using ai algorithms with selectkbest. Medical & Biological Engineering & Computing 61, 3397–3408. doi:10.1007/s11517-023-02918-8.

Salas Ccapacca, M., 2018. Prevención de la desnutrición infantil en San Martín de Porres (Lima, Perú). Tesis de maestría. Pontificia Universidad Católica del Perú. Lima, Perú. URL: http://hdl.handle.net/20.500.12404/15982. acceso abierto, repositorio PUCP-Institucional.

Saputra, D.C.E., Sunat, K., Ratnaningsih, T., 2023. A new artificial intelligence approach using extreme learning machine as the potentially effective model to predict and analyze the diagnosis of anemia. Healthcare 11. doi:10.3390/healthcare11050697.

Sarstedt, M., Danks, N., 2021. Prediction in hrm research–a gap between rhetoric and reality. Human Resource Management Journal doi:10.1111/1748-8583.12400.

Shannon, C.E., 1948. A mathematical theory of communication. Bell System Technical Journal 27, 379–423. doi:10.1002/j.1538-7305.1948.tb01338.x.

Shimizu, Y., Hayakawa, H., Takada, M., Okada, T., Kiyama, M., 2021. Hemoglobin and adult height loss among japanese workers: A retrospective study. PLoS ONE 16. doi:10.1371/journal.pone.0256281.

Singh, S., Sharma, S., 2021. On a generalized entropy and dissimilarity measure in intuitionistic fuzzy environment with applications. Soft Computing 25, 7493 – 7514. doi:10.1007/s00500-021-05709-1.

Soper, D.S., 2021. Greed is good: Rapid hyperparameter optimization and model selection using greedy k-fold cross validation. Electronics doi:10.3390/electronics10161973.

Sukanesh, R., Harikumar, R., 2006. Minimum relative entropy (mre) method for fuzzy based classification of epilepsy risk levels from eeg signals. 2006 International Conference on Biomedical and Pharmaceutical Engineering , 93–98.

Sun, L., Wang, L., Ding, W., Qian, Y., Xu, J., 2021. Feature selection using fuzzy neighborhood entropy-based uncertainty measures for fuzzy neighborhood multigranulation rough sets. IEEE Transactions on Fuzzy Systems 29, 19–33. doi:10.1109/TFUZZ.2020.2989098.

Sun, R., Wong, W., Wang, J., Tong, R., 2017. Changes in electroencephalography complexity using a brain computer interface-motor observation training in chronic stroke patients: A fuzzy approximate entropy analysis. Frontiers in Human Neuroscience 11. doi:10.3389/fnhum.2017.00444.

Sutou, A., Wang, J., 2024. Influence-balanced xgboost: Improving xgboost for imbalanced data using influence functions. IEEE Access 12, 193473–193486. doi:10.1109/ACCESS.2024.3520159.

Szeghalmy, S., Fazekas, A., 2023. A comparative study of the use of stratified cross-validation and distribution-balanced stratified cross-validation in imbalanced learning. Sensors (Basel, Switzerland) 23. doi:10.3390/s23042333.

de Sá Ferreira, F., 2023. Advancements in medical diagnostics through fuzzy logic: A comprehensive review. Revista ft doi:10.69849/revistaft/ar10202310011040.

Tanaka, K., Yon, G.G.V., 2024. Imaginary network motifs: Structural patterns of false positives and negatives in social networks. Soc. Networks 78, 65–80. doi:10.1016/j.socnet.2023.11.005.

Tesfaye, T.S., Tessema, F., Jarso, H., 2020. Prevalence of anemia and associated factors among “apparently healthy” urban and rural residents in ethiopia: A comparative cross-sectional study. Journal of Blood Medicine 11, 89 – 96. doi:10.2147/JBM.S239988.

Tian, J., Tsai, P.W., Zhang, K., Cai, X., Xiao, H., Yu, K., Zhao, W., Chen, J., 2024. Synergetic focal loss for imbalanced classification in federated xgboost. IEEE Transactions on Artificial Intelligence 5, 647–660. doi:10.1109/TAI.2023.3254519.

Ueda, Y., Morishita, J., 2023. Patient identification based on deep metric learning for preventing human errors in follow-up x-ray examinations. Journal of Digital Imaging 36, 1941 – 1953. doi:10.1007/s10278-023-00850-9.

vSafr'anek, D., Thingna, J., 2020. Quantifying information extraction using generalized quantum measurements. Physical Review A doi:10.1103/PhysRevA.108.032413.

Westgard, C., Orrego-Ferreyros, L.A., Calderón, L., Rogers, A.M., 2020. Nutrition uptake and disease of children with anemia in peru, a cross-sectional analysis doi:10.21203/rs.3.rs-40610/v1.

Wijesuriya, R., Moreno-Betancur, M., Carlin, J., Silva, A.D.D., Lee, K.J., 2022. Multiple imputation approaches for handling incomplete three-level data with time-varying cluster-memberships. Statistics in Medicine 41, 4385 – 4402. doi:10.1002/sim.9515.

World Health Organization, 2023. Anemia Consultado: 1 de mayo de 2023.

Yin, T., Liu, N., Sun, H., Xia, S., 2025. Boosting active learning via re-aligned feature space. Knowl. Based Syst. 311, 113085. doi:10.1016/j.knosys.2025.113085.

Zhang, H., Sun, B., Peng, W., 2023. A novel hybrid deep fuzzy model based on gradient descent algorithm with application to time series forecasting. Expert Syst. Appl. 238, 121988. doi:10.1016/j.eswa.2023.121988.

Zhang, X., Jiao, Y., Zhang, D., Wang, X., Cui, Y., 2024. Pixel-based small-window parametric ultrasound imaging for liver tumor characterization. Biomedical engineering letters 14 5, 1011–1021. doi:10.1007/s13534-024-00385-0.

Zhao, H., Wu, Y., Deng, W., 2023. An interpretable dynamic inference system based on fuzzy broad learning. IEEE Transactions on Instrumentation and Measurement 72, 1–12. doi:10.1109/TIM.2023.3316213.