UNIVERSIDAD TÉCNICA ESTATAL DE QUEVEDO

FACULTAD DE POSGRADO

MAESTRÍA EN CIENCIAS DE DATOS

Proyecto de investigación previa a la obtención del Grado Académico de Magíster en Ciencias de Datos

TEMA

"MODELO PREDICTIVO DE DESNUTRICIÓN INFANTIL EN EL ECUADOR: DISTRITO ZONA 5"

AUTORA

ING. ANGÉLICA NEOMÍ CARRIÓN GONZÁLEZ

DIRECTOR

ING. ORLANDO RAMIRO ERAZO MORETA, PhD

QUEVEDO – ECUADOR

2024

RESUMEN

En Ecuador, la desnutrición infantil se ha convertido en un desafío en el área de salud pública, marcando el desarrollo de los niños. Esta incidencia no solo afecta en lo físico sino también en lo cognitivo, emocional y social, siendo una desventaja para el ciclo de su vida. La implementación de programas sociales y el disponer de herramientas para combatir el aumento, la evaluación de factores y la predicción de la desnutrición infantil han sido un reto para las autoridades del país.

Ante esta situación, este trabajo persigue disponer de un modelo que ayude en la predicción de la desnutrición infantil, centrándose en una zona concreta de Ecuador, la zona 5 (Santa Elena, Guayas, Los Ríos, Galápagos). Para ello, se empleó una base de datos otorgada por el Ministerio de Salud Pública que incluye registros de pacientes infantiles de 2021 y 2022. La recolección, unificación de datos, limpieza, tratamiento de valores perdidos y normalización de variables están entre las técnicas utilizadas en el estudio. Para disponer del modelo se utilizaron algoritmos de Aprendizaje Automático como regresión logística, *Random Forest*, *K-Nearest Neighbors*, Árbol de clasificación y *XGBoost.* Los resultados indican que *XGBoost* tiene la mayor precisión en la predicción de la desnutrición infantil. Los indicadores clave como edad, peso, talla e índice de masa corporal fueron identificados a través del análisis de los datos; estos son esenciales para evaluar el estado nutricional de los niños. El modelo obtenido ha demostrado ser una herramienta útil para identificar la desnutrición temprana, lo que ayudaría en la implementación de intervenciones preventivas y terapéuticas más efectivas.

Palabras clave: Desnutrición Infantil, Modelo predictivo, Aprendizaje automático, Salud pública

ABSTRACT

In Ecuador, child malnutrition has become a public health challenge, impacting children's development. This incidence affects not only the physical aspect but also the cognitive, emotional, and social dimensions, posing a disadvantage throughout their life cycle. The implementation of social programs and the availability of tools to combat the rise, evaluate factors, and predict child malnutrition has been a challenge for the country's authorities.

In response to this situation, this work aims to provide a model that helps predict child malnutrition, focusing on a specific area of Ecuador, zone 5 (Santa Elena, Guayas, Los Ríos, Galápagos). To achieve this, a database from the Ministry of Public Health, including child patient records from 2021 and 2022, was used. Data collection, unification, cleaning, treatment of missing values, and normalization of variables are among the techniques used in the study. Machine learning algorithms such as logistic regression, Random Forest, K-Nearest Neighbors, classification tree, and *XGBoost* were used to evaluate the model.

The results indicate that *XGBoost* has the highest accuracy in predicting child malnutrition. Key indicators such as age, weight, height, and body mass index were identified through data analysis; these are essential for assessing children's nutritional status. The predictive model has proven to be a useful tool for identifying early malnutrition, which would help in implementing more effective preventive and therapeutic interventions.

Keywords: Child Malnutrition, Predictive Model, Machine Learning, Public Health

CERTIFICACIÓN DE CULMINACÓN DE PROYECTO DE INVESTIGACIÓN

Quevedo, agosto del 2024.

El suscrito, Ing. Orlando Erazo Moreta, PhD., docente de la Universidad Técnica Estatal de Quevedo, certifica que la Ing. Angélica Noemí Carrión González, realizó el Proyecto de Investigación de grado titulado ““**MODELO PREDICTIVO PARA DETERMINAR LA DESNUTRICIÓN INFANTIL EN EL ECUADOR”,** previo a la obtención del título de Magíster en Ciencia de Datos, bajo mi dirección, habiendo cumplido con las disposiciones reglamentarias establecidas para el efecto.

Firmado electrónicamente por:
ORLANDO RAMIRO ERAZO MORETA

Ing. Orlando Erazo Moreta, PhD.

DIRECTOR

DECLARACIÓN DE AUTORÍA Y CESIÓN DE DERECHOS

Yo, **Angélica Noemí Carrión González**, declaro que la investigación aquí descrita es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

La Universidad Técnica Estatal de Quevedo, puede hacer uso de los derechos correspondientes a este documento, según lo establecido por la Ley de Propiedad Intelectual, por su Reglamento y por la normatividad institucional vigente.

Ing. Angélica Noemí Carrión González

C.C. #0940802788

DEDICATORIA

Dedico este trabajo a mis padres, cuyo amor, comprensión y apoyo constante han sido el pilar fundamental en mi desarrollo personal y académico.

A mis hijos, quienes con su alegría y motivación constante me han impulsado a seguir avanzando en el aprendizaje.

A mis mentores, cuya guía y enseñanza han dejado una marca indeleble en mi formación.

Y a todos aquellos que, con su apoyo y ejemplo, han sido parte esencial en la culminación de este proyecto.

Ing. Angélica Noemí Carrión González

AGRADECIMIENTO

Expreso mi profunda gratitud a Dios, por otorgarme la sabiduría y resiliencia necesarias para alcanzar este logro.

Agradezco a la Universidad Técnica Estatal de Quevedo por brindarme la oportunidad de estudiar y adquirir nuevos conocimientos, permitiéndome avanzar y mejorar en mi desarrollo profesional y académico.

A los mentores que han compartido su conocimiento y experiencia, enriqueciendo mi recorrido académico.

Al director de Tesis, Ing. Orlando Erazo, PhD, por compartir sus conocimientos, su orientación constante y su apoyo incondicional durante el tiempo recorrido en este proceso.

Agradezco a los especialistas en el área de salud por su valiosa participación en este proyecto y por compartir sus conocimientos sobre la desnutrición infantil, lo cual ha sido fundamental para su desarrollo.

Finalmente, extiendo mi gratitud Ministerio de Salud Pública, Distrito Zona 5, por su valiosa colaboración al proporcionar la base de datos necesaria para esta investigación. Su apoyo ha sido esencial para el desarrollo de este proyecto.

Ing. Angélica Noemí Carrión González

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

INTRODUCCIÓN

Según la Organización de las Naciones Unidas para la Alimentación y la Agricultura (FAO), cita en "El Estado de la seguridad alimentaria y la nutrición en el mundo" (2021), en el año 2020, alrededor de 156 millones de menores de cinco años sufrían de desnutrición en todo el mundo. Este problema de salud pública ha sido ampliamente documentado en la literatura científica. Por ejemplo, Black et al. (2013) destacan en su artículo publicado en The Lancet que la desnutrición contribuye a casi la mitad de las muertes en menores de cinco años a nivel global, lo que subraya la magnitud del problema. Además, la FAO (2020) también reporta que, pese a los esfuerzos internacionales, la prevalencia de la desnutrición infantil sigue siendo alarmantemente alta, afectando el desarrollo físico y cognitivo de millones de niños.

En América Latina, la desnutrición infantil se registra como uno de los principales problemas de salud pública y bienestar social; es una de las mayores causas de mortalidad y morbilidad evitable en los niños. Además, Rivera (2019) señala que la causa se encuentra relacionada con los problemas sociales y económicos y con las políticas de salud de la gran mayoría de países de la región.

En Ecuador, la desnutrición infantil es un problema de salud pública que perjudica a niños y adolescentes, a pesar de los intentos de diferentes instituciones públicas o privadas. Según Figueroa y Ruiz (2023), los datos oficiales de las Naciones Unidas publicados en el año 2022 indican que el "23,1 % de niños menores de cinco años del territorio ecuatoriano tiene desnutrición crónica infantil o retraso en su crecimiento". La desnutrición crónica en Ecuador afecta al 27,2 % de los niños, según los datos de la Encuesta Nacional de Desnutrición Infantil (ENDI) realizada por el Instituto

Nacional de Estadística y Censos (2023) en 2022-2023. Este dato destaca la magnitud del problema y la necesidad de implementar políticas más efectivas para reducir la desnutrición infantil en el país. Con estas cifras, el país se sitúa como la segunda nación en América Latina y el Caribe con la mayor prevalencia de desnutrición crónica en niños, superado solo por Guatemala como lo describe la FAO (2021).

Con base en la situación problemática expresada, este trabajo se enfoca en la búsqueda de un modelo que ayude a estimar la desnutrición infantil en Ecuador con miras a abordarla con cifras proyectadas basadas en la realidad. Este modelo permitirá conocer los riesgos de manera generalizada de desnutrición temprana, lo que facilitaría intervenciones preventivas y terapéuticas rápidas, para abordar esta limitación.

El resto del documento se encuentra estructurado en capítulos. El Capítulo I, Contextualización de la problemática, aborda la ubicación y contextualización del problema, la situación actual, la delimitación del problema, los objetivos y la justificación. El Capítulo II, Marco Teórico, comprende la fundamentación conceptual, la fundamentación teórica y la fundamentación legal. El Capítulo III, Marco Metodológico, incluye el tipo de investigación, los métodos utilizados en la investigación, la construcción metodológica del objeto de investigación, la elaboración del marco teórico, la recolección de la información, y el procesamiento y análisis de los datos. El Capítulo IV, Resultados y Discusión, se enfoca en la presentación de los resultados obtenidos y su discusión. Por último, en el Capítulo V, Conclusiones y Recomendaciones, se presentan las conclusiones derivadas de la investigación y las recomendaciones pertinentes.

CAPÍTULO I

MARCO CONTEXTUAL DE LA INVESTIGACIÓN

1.1. UBICACIÓN Y CONTEXTUALIZACIÓN DE LA PROBLEMÁTICA

La desnutrición infantil es un desafío para cualquier nación debido a sus efectos negativos en la salud, la educación, el trabajo, y la economía familiar y social como lo describen Moncayo et al. (2021). La desnutrición infantil es un problema grave en América Latina, y se relaciona con la enfermedad y muerte evitable de niños. Esto se debe a factores sociales, económicos y políticos deficientes en los países de la región. Sin embargo, existen países como Costa Rica, Brasil y Chile que tienen políticas de salud efectivas y que han logrado reducir la desnutrición infantil.

En Ecuador, la desnutrición infantil sigue siendo un problema constante que afecta la salud y desarrollo de los menores. Según la Organización de las Naciones Unidas para la Alimentación y la Agricultura (2021) aproximadamente uno de cada tres niños menores de cinco años presenta alguna forma de malnutrición, como retraso en el crecimiento, emaciación o sobrepeso. Esta situación persiste como un problema de salud pública importante, a pesar de los esfuerzos gubernamentales y no gubernamentales.

Debido a las limitadas asignaciones presupuestarias para la atención de la desnutrición infantil, como mencionan Locks et al. (2015), el problema es igualmente preocupante a nivel de distritos. Los planes para abordarla carecen de protocolos adecuados y se manejan en conjunto con los gobiernos municipales, lo que dificulta su implementación efectiva.

Por otro lado, según las estadísticas de la Encuesta Nacional de Desnutrición Infantil (ENDI), realizada en 2022-2023 por el Instituto Nacional de Estadística y Censos (INEC) (2023), la desnutrición crónica en Ecuador afecta al 27,2 % de los niños.

Este dato subraya la necesidad urgente de dirigir políticas efectivas para abordar y reducir la incidencia de la desnutrición infantil en estas áreas. La ENDI proporciona una base sólida y actualizada de datos, destacando la magnitud del problema y la importancia de implementar estrategias de intervención más focalizadas y eficientes. Esta situación pone de manifiesto la necesidad de contar con más opciones para enfrentar la desnutrición infantil. Se requiere de herramientas tecnológicas que posean características que permitan evaluarla e idealmente hasta predecirla, facilitando la toma de decisiones y la implementación de intervenciones adecuadas, con acciones eficaces en el corto tiempo. De forma más concreta, sería de utilidad disponer de un modelo que permita hacer estimaciones para identificar a los niños en riesgo de desnutrición. Es por esto que, con el presente estudio, se pretende usar técnicas de Ciencia de Datos tratando de contribuir en al ámbito de la salud y por el bienestar de los niños con desnutrición en Ecuador.

1.2. SITUACIÓN ACTUAL DE LA PROBLEMÁTICA

La situación actual de la problemática se caracteriza por la persistente prevalencia de la desnutrición infantil en Ecuador, a pesar de los esfuerzos realizados para abordar esta situación. La desnutrición infantil sigue siendo un desafío de salud pública significativo y un obstáculo para el desarrollo integral de los niños ecuatorianos.

De acuerdo con Moncayo et al. (2021), la desnutrición infantil afecta negativamente el crecimiento físico y cognitivo de los niños. Puede provocar un mayor riesgo de enfermedades, un bajo rendimiento escolar y un desarrollo socioemocional comprometido con efectos a largo plazo. Esto mantiene un ciclo de desventajas que daña la calidad de vida de los niños y su futuro.

Las medidas públicas son un medio apropiado para combatir la desnutrición. De hecho, Ecuador está atravesando cambios nutricionales y epidemiológicos que están llevando a una reducción de la desnutrición. Sin embargo, la realidad es que nuestro país está insuficientemente nutrido en comparación con su nivel de desarrollo.

En el informe "El Estado Mundial de la Infancia 2019" de UNICEF (2019), se destaca que aproximadamente 500,000 niños menores de cinco años sufren desnutrición y los avances en los últimos 30 años han sido limitados. Aunque se han implementado políticas y programas para combatir la desnutrición infantil en el Ecuador, en algunas regiones y comunidades del país se observa la alta prevalencia de este problema tal como indican Huiracocha et al. (2012).

En Ecuador, el incremento en el precio de los productos ocasionó que el presupuesto general del Estado creciera a lo largo de la década anterior. Esto permitió la implementación de políticas sociales dirigidas a combatir la malnutrición infantil. Sin embargo, los resultados no fueron los esperados. A pesar del aumento en las políticas y programas a favor de la lucha contra la desnutrición infantil, muchos niños en su primera infancia siguen experimentando síntomas de desnutrición, como retraso en el crecimiento, bajo peso para su edad y deficiencias nutricionales(Rivera, 2019).

La toma de decisiones para abordar la desnutrición infantil es un proceso complejo que requiere herramientas que apoyen la toma de decisiones. La identificación temprana de niños en riesgo es esencial para brindar intervenciones preventivas y de tratamiento de manera oportuna, lo que enfatiza la necesidad urgente de desarrollar un modelo predictivo basado en datos. En concreto, la situación actual de

la desnutrición infantil en Ecuador resalta la importancia de la búsqueda de estas herramientas para mejorar la toma de decisiones y la calidad de vida de la población infantil en el país.

1.3. PROBLEMA DE INVESTIGACIÓN

1.3.1. Problema general

¿Cómo disponer de un modelo de predicción que ayude en la prevención de la desnutrición infantil en niños ecuatorianos del distrito zona 5 de hasta cinco años?

1.3.2. Problemas derivados

¿Cuáles son los indicadores más relevantes para evaluar y ayudar a prevenir la desnutrición infantil?

¿Cómo estaría estructurado un modelo que permita hacer predicciones de la desnutrición infantil en el distrito zona 5 del Ecuador?

¿Cómo comprobar la efectividad del modelo de predicción de datos escogido para prevenir la desnutrición infantil en el distrito zona 5 del Ecuador?

1.4. DELIMITACIÓN DEL PROBLEMA

- **CAMPO:** Tecnología de información y la comunicación
- **ÁREA:** Tecnología de información y la comunicación
- **LÍNEA:** Ciencias de datos
- **LUGAR:** Ciudad Quevedo
- **TIEMPO:** De septiembre 2023 a marzo del 2024 de investigación

1.5. OBJETIVOS

1.5.1. Objetivo General

Proponer un modelo predictivo utilizando técnicas de análisis de datos y aprendizaje automático con el fin de ayudar en la prevención de la desnutrición infantil de niños ecuatorianos distrito zona 5.

1.5.2. Objetivos Específicos

1. Seleccionar los indicadores relevantes con base en los parámetros médicos establecidos y aplicando técnicas estadísticas que ayuden a evaluar la desnutrición infantil para la construcción de un modelo predictivo.
2. Examinar los modelos y algoritmos de aprendizaje automático pertinentes para analizar datos relacionados con la desnutrición infantil del distrito zona 5 en Ecuador, estimando el grado de precisión en las predicciones a través del entrenamiento y la validación de los resultados.
3. Determinar el modelo con un rendimiento favorable, con mayor grado de precisión en las predicciones de la desnutrición infantil, mediante los resultados de un conjunto de datos de validación.

1.6. JUSTIFICACIÓN

El fin del presente trabajo es ayudar a obtener información para la toma de decisiones para la prevención de la desnutrición infantil del distrito zona 5 del Ministerio de Salud Pública del Ecuador. La falta de nutrientes necesarios para crecer y desarrollarse adecuadamente afecta negativamente la calidad de vida actual y futura de los niños. Esto representa una amenaza constante para su desarrollo físico y cognitivo. Al predecir áreas o grupos de población con mayor riesgo de desnutrición, se pueden asignar recursos de manera más eficiente, enfocándolos en programas de nutrición específicos y asegurando que lleguen a quienes más lo necesitan.

La generación de un modelo predictivo utilizando técnicas de análisis de datos y aprendizaje automático es el objetivo general de este estudio, con la finalidad de anticipar y advertir la desnutrición infantil, identificando factores de riesgo en la población infantil, antes de que se desarrollen enfermedades graves. Esto puede facilitar una intervención temprana, efectiva y contribuir al bienestar general y al potencial futuro de los niños.

El primer paso importante es conocer e identificar factores relevantes mediante métodos estadísticos que permitan evaluar y predecir la desnutrición infantil. Este paso ayuda a comprender la dinámica de la desnutrición y seleccionar las variables más influyentes en el proceso. Además, se propone para el desarrollo del modelo utilizar una variedad de modelos y algoritmos como regresión logística, SVM (Máquinas de vectores soporte), *Random Forest*, Árbol de clasificación y *XGBoost* para realizar las predicciones de la desnutrición infantil en el Ecuador. Esto implica

la aplicación de técnicas avanzadas de aprendizaje automático y el análisis de datos para obtener un modelo adaptable.

Prevenir la desnutrición infantil no solo tiene un impacto en la salud, sino que también mejora la calidad de vida de los niños y sus familias. Por ello, este proyecto se alinea con los Objetivos de Desarrollo Sostenible, en particular con el primero (relacionado con el fin del hambre) y el tercero (que insta a la salud y el bienestar de la población). El objetivo es evaluar la efectividad del modelo ya desarrollado utilizando un conjunto de datos de referencia, contribuyendo así a la identificación temprana y prevención de la desnutrición infantil. Esto permitirá valorar los resultados de las predicciones y, en consecuencia, su utilidad como herramienta para la toma de decisiones efectivas para la prevención y el abordaje temprano de la desnutrición. Además, representa un paso significativo hacia la promoción de políticas y programas de salud más eficaces y focalizados en las necesidades de los niños en riesgo de desnutrición.

CAPÍTULO II

MARCO TEÓRICO

2.1 FUNDAMENTACIÓN CONCEPTUAL

2.1.1 Infancia

El Fondo de las Naciones Unidas, citado en Jiménez (2012), describe a la infancia como un sujeto reconocido con derechos sociales y civiles en igualdad de condiciones al adulto. Infancia, para UNICEF (2016), es el período de mayor y más rápido desarrollo en la vida de una persona. Durante esta etapa se construyen las bases del futuro de cada niño, de su salud, bienestar y educación.

Jiménez (2012) cita que la infancia está basada en la idea de que el niño es un ser incapaz, irresponsable e incompleto, en proceso de formación, no autónomo y dependiente. Esta visión justifica que los adultos mantengan una especie de derecho de propiedad sobre los niños. En este contexto, se establece que los niños no son sujetos de derechos, sino objetos sobre los cuales otros sujetos (como el Estado, la sociedad, la escuela y los padres) ejercen sus derechos.

2.1.2 Desnutrición infantil

Ortega (2019) hace referencia a la desnutrición como un conjunto de manifestaciones clínicas, alteraciones bioquímicas y antropométricas producidas por la ingesta deficiente y/o aprovechamiento biológico de macronutrientes ocasionando la insatisfacción de requerimientos nutricionales.

Suryawan et al. (2022) menciona la importancia de la intervención temprana se subraya en varios estudios científicos. Por ejemplo, un estudio de revisión sistemática publicado en Cambridge Core, Suryawan et al. (2022) indica que la desnutrición en los primeros 1,000 días de vida puede tener efectos duraderos en el desarrollo neurocognitivo, afectando negativamente el rendimiento escolar y la capacidad de aprendizaje. Además, Mark et al. (2020) destacan que la desnutrición

es una “pandemia silenciosa” que afecta a millones de niños globalmente, con implicaciones significativas para su salud y supervivencia.

2.1.3. Tipos de desnutrición.

2.1.3.1. Desnutrición crónica

Este tipo de desnutrición se caracteriza por una carencia de los nutrientes necesarios durante un tiempo prolongado por lo que aumenta el riesgo de que contraiga enfermedades afectando al desarrollo físico e intelectual del niño. La desnutrición crónica se manifiesta principalmente por un retraso en el crecimiento, medido según la relación altura-edad. Este trastorno indica una deficiencia nutricional a largo plazo en la dieta, provocando deterioro físico y mental. En algunos casos, si no se trata a tiempo, la causa se vuelve irreversible (Durán-Pincay et al., 2022).

UNICEF (2021) describe que “un niño con desnutrición crónica puede tener problemas de aprendizaje en la edad escolar, sobrepeso, obesidad y enfermedades no transmisibles, como hipertensión o diabetes en la vida adulta, y dificultades para insertarse en el mercado laboral”. Además de afectar a las personas que la padecen, la desnutrición crónica tiene un fuerte impacto en el desarrollo económico y social de los países.

En Ecuador, los gastos asociados a la malnutrición como salud, educación y pérdida de productividad representan el 4,3 % del producto interno bruto (Cueva Moncayo et al., 2021).También, Moncayo et al. (2021) mencionan que la desnutrición crónica infantil afecta a aproximadamente 156 millones de niños en todo el mundo. Por ello, se constituye en una problemática multicausal que retrasa el crecimiento de los menores con relación a su edad, e impacta negativamente y de manera definitiva en su desarrollo.

2.1.3.2. Desnutrición aguda

Manosalvas (2019) cita que La insuficiencia de peso respecto de la talla se denomina emaciación o desnutrición aguda. Esta desnutrición se mide mediante el perímetro del brazo comparándolo con estándares internacionales. Una vez comprobado que un niño sufre de desnutrición aguda moderada requiere de un tratamiento inmediato. Por otro lado, Fernández et al. (2022) mencionan que la desnutrición aguda implica un peso infantil inferior al estándar para su estatura. Esta desnutrición exige atención urgente porque afecta procesos vitales de niños, embarazadas y madres lactantes. La desnutrición aguda en niños se identifica por bajo peso para su estatura, posible emaciación, y debe tratarse rápidamente para evitar complicaciones y mortalidad.

2.1.3.3. Desnutrición aguda grave o severa

Fernández et al. (2022) mencionan que es la forma de desnutrición más grave. El niño tiene un peso muy por debajo del estándar de referencia para su altura; se consideran tres desviaciones estándar. Se mide también por el perímetro del brazo. Altera todos los procesos vitales del niño y conlleva un alto riesgo de mortalidad. El riesgo de muerte para un niño con desnutrición aguda grave es nueve veces superior que para un niño en condiciones normales; requiere de atención médica urgente.

2.1.4. Ciencia de Datos

Silva y Martínez (2021) mencionan que la Ciencia de Datos se enfoca en analizar datos para obtener información relevante para empresas. Combina diferentes disciplinas para analizar datos de gran volumen. Este análisis permite a los científicos de datos abordar y responder preguntas sobre eventos pasados, sus causas, eventos futuros y posibles acciones basadas en los resultados.

Por otro lado, Córdova et al. (2018) citan que la importancia de la Ciencia de Datos

radica en su capacidad para generar significado a partir de los datos, mediante herramientas, métodos y tecnología. Las organizaciones modernas están saturadas de datos; hay una abundancia de dispositivos que pueden recolectar y guardar información automáticamente. En el comercio electrónico, la medicina, las finanzas y otros aspectos, los sistemas en línea y los portales de pago recopilan más información. Así, en la actualidad existe mucha información textual, auditiva, visual y gráfica disponible.

2.1.5. Áreas que componen la Ciencia de Datos

2.1.5.1. Estadística

Argota et al. (2022) describen que la estadística se ocupa del tratamiento de datos y la inferencia poblacional mediante la observación de hechos; es decir, define los resultados esperados, la población, unidades de observación, variables, métodos, plan de muestreo, tamaño de la muestra y factores, métodos estadísticos, diseño de experimentos, entre otros aspectos.

2.1.5.2. Inteligencia artificial

Silva et al. (2021) hacen referencia que la Inteligencia Artificial (IA) se basa en algoritmos absolutos del todo o nada. Vale decir que su resultado se evalúa como correcto o incorrecto según resuelva exactamente una tarea o no, sin admitir errores; si funciona bien en todos los casos se implementa (Prol Castelo et al ., 2022). Las IA emplean algoritmos y modelos matemáticos para procesar datos y tomar decisiones basadas en aprendizaje automático. La IA mejora su precisión y eficiencia con el tiempo.

2.1.5.3. Visualización de datos

Cairo (2017) describe que la visualización de datos es una técnica fundamental en la Ciencia de Datos que va más allá de la simple representación gráfica de información. Una visualización de datos no es sólo para ser vista, como si fuese un dibujo, sino para ser leída e interpretada con atención, interrogándose a uno mismo no sólo sobre lo que el gráfico revela, sino sobre lo que puede no estar mostrando. Para una visualización de datos apropiada, las palabras y los gráficos deben estar unidos para formar argumentos y como sustento de discusiones y debates, reforzándose mutuamente.

2.1.6. Técnicas de ciencias de datos

2.1.6.1.Análisis exploratorio de datos (AED)

Se trata de un método que prioriza el análisis de datos, existiendo numerosos criterios para ello. El objetivo del AED es examinar los datos antes de usar cualquier técnica estadística para obtener una comprensión básica de los datos y las relaciones entre las variables analizadas. Cualquier cálculo (promedios, desviaciones, correlaciones, etc.) debe ser precedido por un análisis visual de los datos (Browne et al., 2021). El primer objetivo del AED es encontrar patrones o modelos en las distribuciones univariadas de datos y encontrar anomalías y errores, utilizando una variedad de técnicas gráficas y buscando estimadores no paramétricos, libres de distribución o simplemente estimadores robustos (Browne et al., 2021).

2.1.6.2.Clasificación

Antunes (2019) define que la clasificación consiste en ordenar los datos en grupos o categorías específicas. Las computadoras están entrenadas para identificar y ordenar datos. Los conjuntos de datos conocidos se utilizan para crear algoritmos de

decisión en una computadora que procesa y categoriza rápidamente los datos.

2.1.6.3. Regresión

Hasan et al. (2020) describen a la regresión como el método para encontrar una relación entre dos puntos de datos que aparentemente no se relacionan. La conexión se suele modelar en torno a una fórmula matemática y se representa en forma de gráfico o curvas. Cuando se conoce el valor de un punto de datos, se utiliza la regresión para predecir el otro punto de datos.

2.1.6.4. Regresión Logística

Fiuza Pérez & Rodríguez Pérez (2000) definen que la regresión logística es aquella que halla, para cada individuo, según los valores de una serie de variables (Xi), la probabilidad (p) de que presente el efecto estudiado. Una transformación logarítmica de dicha ecuación, a la que se le llama logit, consiste en convertir la probabilidad (p) en odds (comparación de probabilidad de evento). De aquí surge la ecuación de la regresión logística, que es parecida a la ecuación de la regresión lineal múltiple.

2.1.6.5. Máquina de vectores de soporte (SVM)

SVM es un método eficaz para crear un clasificador. Su objetivo es crear un límite de decisión entre dos clases que permitan la predicción de etiquetas a partir de uno o más vectores de características, Además, es muy eficaz para reconocer patrones sutiles en conjuntos de datos complejos Huang et al. (2018).

2.1.6.6. Árbol de Clasificación (AC)

Los árboles de clasificación se emplean para asignar unidades experimentales a las clases de una variable dependiente a partir de sus mediciones en uno o más predictores. Los AC se emplean para establecer unidades a cada uno de los dos grupos

definidos por la variable respuesta (Beltrán & Barbona, 2019). Otro estudio, realizado por Fenta et al. (2021), resalta cómo las estrategias de optimización local pueden mejorar la estructura de los árboles de clasificación generados por métodos ávidos, minimizando la pérdida de clasificación en cada nodo del árbol y promoviendo soluciones más esparsas (soluciones con pocos elementos significativo o nulos) y generalizables.

2.1.6.7. Bosque aleatorio (Random Forest)

De acuerdo con Espinosa (2020), el algoritmo Bosque aleatorio (*Random Forest*) es una técnica de aprendizaje supervisado que genera múltiples árboles de decisión sobre un conjunto de datos de entrenamiento; los resultados obtenidos se combinan a fin de obtener un modelo único más robusto en comparación con los resultados de cada árbol por separado. Cada árbol se obtiene mediante un proceso de dos etapas:

1. Se genera un número considerable de árboles de decisión con el conjunto de datos. Cada árbol contiene un subconjunto aleatorio de variables m (predictores) de forma que m < M (donde M = total de predictores).
2. Cada árbol crece hasta su máxima extensión**.**

Espinosa-Zúñiga (2020) describe que cada árbol generado por el algoritmo *Random Forest* contiene un grupo de observaciones aleatorias. Las observaciones no estimadas en los árboles, también conocidas como "*out of the bag*" (OOB), se utilizan para validar el modelo. Las salidas de todos los árboles se combinan en una salida final Y (conocida como ensamblado), la cual se obtiene mediante alguna regla. Generalmente, el promedio se utiliza cuando las salidas de los árboles del ensamblado son numéricas, y el conteo de votos se emplea cuando las salidas son categóricas.

2.1.6.8. KNN K-Vecinos más cercanos)

IBM (2023) define que el algoritmo KNN es un método no paramétrico, o basado en instancias, que ha sido considerado como uno de los métodos más simples en minería de datos y aprendizaje automático. El principio del algoritmo KNN es que las muestras más similares que pertenecen a la misma clase tienen una alta probabilidad.

Por lo general, el algoritmo KNN primero encuentra k vecinos más cercanos de una consulta en el conjunto de datos de entrenamiento y, a continuación, predice la consulta con la clase principal en los k vecinos más cercanos.

2.1.6.9. Modelo XGBoost (Aumento extremo de gradiente)

Giraldo y Vargas (2011) menciona que el algoritmo de aumento de gradiente *XGBoost* es conocido por su excelente desempeño en una amplia gama de aplicaciones, como la regresión, la clasificación. *XGBoost* captura relaciones complejas y logra un rendimiento de última generación optimizando de forma iterativa una función de pérdida utilizando un conjunto de árboles de decisión. Si bien otros algoritmos tienen sus propias fortalezas, *XGBoost* proporciona una combinación atractiva de versatilidad y precisión.

2.1.7. Métricas de evaluación

A los modelos se les debe evaluar la calidad de la clasificación. Según Nieto et al. (2019) puede hacerse por cuatro métricas diferentes: exactitud, precisión (especificidad), *recall* (sensibilidad) y medición puntaje F1. Estos valores se determinan a partir de la matriz de confusión (Tabla 1)

Tabla 1*Matriz de confusión*

		Predicción	
		Predicción	Predicción
A ct ...	Positivo	Verdaderos	Falsos negativos
		Falsos positivos	Verdaderos

Elaborado por: Nieto et al. (2019)

A continuación, se explica cada métrica y su relevancia para el análisis y comparación de modelos predictivos.

Precisión: Mide la proporción de predicciones correctas entre todas las predicciones realizadas por un modelo.

Sensibilidad: Mide la proporción de casos positivos reales que el modelo predijo correctamente.

Especificidad: Mide la proporción de casos negativos reales que el modelo predijo correctamente.

Recall: Mide la proporción de casos positivos reales que el modelo predijo correctamente. Es equivalente a la sensibilidad.

Donde:

- VP: Verdaderos positivos
- VN: Verdaderos negativos
- FP: Falsos positivos
- FN: Falsos negativos

Curva ROC

Armesto (2011) hace referencia a que la curva ROC se utiliza este tipo de análisis estadístico siempre que las pruebas de diagnóstico brinden resultados medidos en escala continua, por intervalos u ordinal. Este tipo de análisis permite evaluar la

capacidad de una prueba de diagnóstico para distinguir entre estados de salud alternativos mutuamente excluyentes (sano/enfermo, positivo/negativo, etc.)

Fenta et al. (2021) hacen referencia a que las curvas ROC son gráficos que muestran la sensibilidad en función de los falsos positivos (1-especificidad) de las pruebas diagnósticas. En cada punto de la curva, se representa un par de sensibilidad/ (1-especificidad) que corresponde a un nivel de decisión específico.

2.1.8. Tecnologías utilizadas en Ciencia de Datos

2.1.8.1. Tableu

Munawar et al. (2022) describen Tableau como una solución patentada de inteligencia empresarial muy preferida por su capacidad de ofrecer visualizaciones interactivas de manera rápida gracias a su naturaleza de arrastrar y soltar. Tableau ofrece una amplia variedad de opciones de gráficos, incluyendo pastel y gráficos de barras. Los usuarios pueden crear paneles informativos instantáneamente a partir de diversos conjuntos de datos, realizar agregaciones, y resaltar o profundizar en gráficos con facilidad. Incluso los usuarios novatos pueden crear visualizaciones para iluminar hechos en grandes conjuntos de datos. Tableau se conecta sin esfuerzo a datos almacenados en Excel, CSV y archivos de texto, reconociendo campos y formatos automáticamente.

2.1.8.2.R Studio

AstraEd (2022) describe a R como un lenguaje de programación que se mantiene en un ambiente para cómputo estadístico y gráfico. Permite al usuario (o programador) escribir una serie de instrucciones u órdenes de manera organizada, concentrándose en el manejo, análisis, procesamiento y visualización de datos (Hernández & Usuga, 2024).

2.1.8.3.Python

Basu (2023) describe a Python es un lenguaje de programación potente y elegante, fácil de leer y de entender. Demuestra la mayoría de estas características comunes a muchos otros idiomas y es útil para aplicaciones del mundo real. También es software libre, tiene un estándar de implementación, una comunidad grande y amigable de hackers a su alrededor. Python persigue eliminar la ambigüedad. Al igual que al familiarizarse con cualquier idioma extranjero, una vez que se lee el código lo suficiente, cualquier fragmento de código que encontrado empezará a tener sentido (Basu, 2023).

2.2. FUNDAMENTACIÓN TEÓRICA

Un estudio realizado en dos centros infantiles del Cantón Francisco Orellana Mero & Dario. (2021), Ecuador, investigó el impacto de los factores socioeconómicos y educativos en la desnutrición infantil. La investigación utilizó un diseño descriptivo y correlacional, abarcando una muestra de 73 infantes de 1 a 3 años. Los resultados revelaron una alta prevalencia de desnutrición leve, seguida de desnutrición aguda y crónica. Se encontró una correlación positiva muy alta entre los factores socioeconómicos y educativos y los niveles de desnutrición, con un valor de correlación de Pearson de 0,8, lo que subraya la importancia de estos factores en la nutrición infantil.

Otro trabajo que utiliza modelos de predicción es el de Coz (2024). Este estudio analizó los factores de riesgo que contribuyen a la desnutrición crónica infantil, utilizando una metodología cuantitativa con diseño no experimental y correlacional. Aplicando entrevistas estructuradas y análisis estadístico, se identificaron los

factores de mayor riesgo y se evaluó su impacto en la incidencia de la desnutrición infantil.

También, el trabajo realizado por Alejandría (2021) implementó un modelo de análisis predictivo para optimizar el control nutricional de niños menores de cinco años en Agocucho, Cajamarca. Utilizando datos de encuestas y fichas de observación, y herramientas como Microsoft Excel y Azure, se desarrolló y validó un modelo predictivo basado en árboles de decisión.

El estudio basado en datos de Flores y Congacha (2021) utilizó modelos de regresión logística y árboles de clasificación para identificar factores significativos de desnutrición en una muestra de 11.231 niños. Aplicando técnicas de submuestreo e imputación de datos faltantes, se mejoró la precisión de las predicciones, empleando modelos predictivos basados en árboles de clasificación y regresión logística.

Por otro lado, el proyecto de Espinoza-Estrella (2023) analizó la incidencia de la edad materna temprana en la desnutrición crónica infantil mediante el método de diferencia en diferencias (DD) y modelos multivariados. Utilizó datos de diversas encuestas y aplicó análisis factorial y el modelo KNN para predecir el riesgo de desnutrición en niños de madres adolescentes.

En otro trabajo, Granados Mota (2023) utilizó un enfoque cuantitativo no experimental y descriptivo. Este estudio desarrolló un modelo estadístico basado en datos de 1.930 individuos, aplicando modelos de regresión logística binomial para identificar probabilidades de desnutrición infantil. Se analizaron variables como peso, talla, materiales de la vivienda y condiciones de vida.

Por su parte, en el proyecto de Valdez et al. (2023) se aplicaron técnicas de minería de datos y algoritmos de aprendizaje automático para predecir anemia en niños

peruanos. Con datos de 138.369 instancias y utilizando Anova F-test y Chi Cuadrado, se probaron varios algoritmos, encontrando que Naive Bayes fue el más efectivo para la detección temprana de anemia.

Por otra parte, Rosa y Frutos (2022) destacan la creciente importancia de la Ciencia de Datos en la salud en América Latina, señalando tanto oportunidades como desafíos. Resaltan la necesidad de mejorar la interoperabilidad de los sistemas de información y la disponibilidad de datos confiables, y describen iniciativas en Brasil y México que abordan problemas de salud pública mediante el análisis de datos. Sin embargo, se enfrentan a desafíos como el acceso limitado a la atención médica y la falta de estándares técnicos.

Para concluir, Benavides (2023) realizó un estudio basado en datos de la Encuesta Demográfica y de Salud Familiar (ENDES) 2021, que buscó identificar los factores que influyen en la desnutrición crónica en niños de Huancavelica, Perú. Variables como la talla de la madre, el tipo de paredes de la vivienda, los controles prenatales y la presencia de enfermedades resultaron significativas para la desnutrición. El análisis de regresión logística multivariada confirmó que factores como los controles prenatales, el peso al nacer, la inmunización completa, la duración de la lactancia y la presencia de diarrea y tos continuaron siendo significativos para la desnutrición crónica infantil (DCI), con la excepción del parto institucional.

Comparando todos estos estudios con lo propuesto en este trabajo, se observa que todos emplean metodologías cuantitativas y técnicas avanzadas de análisis de datos para identificar factores de riesgo y predecir la desnutrición infantil. Sin embargo, este proyecto se distingue por la integración de múltiples algoritmos de aprendizaje automático, como regresión logística, Random Forest, KNN, Árbol de Clasificación

y *XGBoost*, proporcionando una comparación más amplia y detallada de la eficacia de estos modelos. Además, al enfocarse en una región específica (Distrito Zona 5 en Ecuador) y utilizar datos recientes de 2021 y 2022, el estudio ofrece una perspectiva actualizada y contextualizada que puede guiar intervenciones más precisas y efectivas en la prevención de la desnutrición infantil.

2.3. FUNDAMENTACIÓN LEGAL

Es crucial la base legal del proyecto, porque proporciona el marco legal que apoya y guía todas las acciones y decisiones relacionadas con la investigación. Se tiene como objetivo establecer una base legal que respalde la importancia y la necesidad del estudio, así como garantizar el cumplimiento de las regulaciones actuales en el ámbito de la salud infantil y la protección de los derechos de los niños en Ecuador.

2.3.1. Ley Orgánica de Protección de Datos

"Ámbito de aplicación material. - La presente ley se aplicará al tratamiento de datos personales contenidos en cualquier tipo de soporte, automatizados o no, así como a toda modalidad de uso posterior. La ley no será aplicable a: f) Datos o bases de datos establecidos para la prevención, investigación, detección o enjuiciamiento de infracciones penales o de ejecución de sanciones penales, llevado a cabo por los organismos estatales competentes en cumplimiento de sus funciones legales. En cualquiera de estos casos deberá darse cumplimiento a los estándares internacionales en la materia de derechos humanos y a los principios de esta ley, y como mínimo a los criterios de legalidad, proporcionalidad y necesidad" (*Ley Orgánica de Protección de Datos Personales*, 2021).

En el artículo 7 sobre "El Tratamiento legítimo de datos personas. -El tratamiento será legítimo y lícito si se cumple con alguna de las siguientes condiciones: 7) Para

tratamiento de datos personales que consten en bases de datos de acceso público; y 8) Para satisfacer un interés legítimo del responsable de tratamiento o de tercero, siempre que no prevalezca el interés o derechos fundamentales de los titulares al amparo de lo dispuesto en esta norma"(*Ley Orgánica de Protección de Datos Personales*, 2021).

En el artículo 26 sobre el "Tratamiento de datos sensibles. -Queda prohibido el tratamiento de datos personales sensibles salvo que concurra alguna de las siguientes circunstancias: a) El titular haya dado su consentimiento explícito para el tratamiento de sus datos personales, especificándose claramente sus fines. f) El tratamiento es necesario con fines de archivo en interés público, fines de investigación científica o histórica o fines estadísticos, que debe ser proporcional al objetivo perseguido, respetar en lo esencial el derecho a la protección de datos y establecer medidas adecuadas y específicas para proteger los intereses y derechos fundamentales del titular. g) Cuando el tratamiento de los datos de salud se sujete a las disposiciones contenidas en la presente ley." (Ley orgánica de protección de datos personales, 2021).

2.3.2. Código de la Niñez y Adolescencia

En el artículo 2 sobre "Del derecho de alimentos. - El derecho a alimentos es connatural a la relación parento-filial y está relacionado con el derecho a la vida, la supervivencia y una vida digna. Implica la garantía de proporcionar los recursos necesarios para la satisfacción de las necesidades básicas de los alimentarios que incluye: 1. Alimentación nutritiva, equilibrada y suficiente; 2. Salud integral: prevención, atención médica y provisión de medicinas" (Código de la niñez y adolescencia, 2013).

En el artículo 4 sobre “Titulares del derecho de alimentos. - Tienen derecho a reclamar alimentos: 1. Las niñas, niños y adolescentes, salvo los emancipados voluntariamente que tengan ingresos propios, a quienes se les suspenderá el ejercicio de este derecho de conformidad con la presente norma” (Código de la niñez y adolescencia, 2013).

En el artículo 27 sobre “Derecho a la salud. - Los niños, niñas y adolescentes tienen derecho a disfrutar del más alto nivel de salud física, mental, psicológica y sexual. El derecho a la salud de los niños, niñas y adolescentes comprende:

1. Acceso gratuito a los programas y acciones de salud públicos, a una nutrición adecuada y a un medio ambiente saludable;
2. Acceso permanente e ininterrumpido a los servicios de salud públicos, para la prevención, tratamiento de las enfermedades y la rehabilitación de la salud. Los servicios de salud públicos son gratuitos para los niños, niñas y adolescentes que los necesiten;
3. Acceso a medicina gratuita para los niños, niñas y adolescentes que las necesiten;
4. Acceso inmediato y eficaz a los servicios médicos de emergencia, públicos y privados;
5. Información sobre su estado de salud, de acuerdo con el nivel evolutivo del niño, niña o adolescente;
6. Información y educación sobre los principios básicos de prevención en materia de salud, saneamiento ambiental, primeros auxilios;
7. Atención con procedimientos y recursos de las medicinas alternativas y tradicionales;

8. El vivir y desarrollarse en un ambiente estable y afectivo que les permitan un adecuado desarrollo emocional;
9. El acceso a servicios que fortalezcan el vínculo afectivo entre el niño o niña y su madre y padre; y,
10. El derecho de las madres a recibir atención sanitaria prenatal y postnatal apropiadas" ((*Código de la Niñez y Adolescencia | Ecuador - Guía Oficial de Trámites y Servicios*, 2017)

El articulo 28 sobre el "Derecho a una vida digna. - Los niños, niñas y adolescentes tienen derecho a una vida digna, que les permita disfrutar de las condiciones socioeconómicas necesarias para su desarrollo integral. Este derecho incluye aquellas prestaciones que aseguren una alimentación nutritiva, equilibrada y suficiente, recreación y juego, acceso a los servicios de salud, a educación de calidad, vestuario adecuado, vivienda segura, higiénica y dotada de los servicios básicos." (Código de la niñez y adolescencia, 2013).

CAPÍTULO III
METODOLOGÍA DE LA INVESTIGACIÓN

3.1 TIPO DE INVESTIGACIÓN

3.1.1 Investigación Correlacional

Este tuvo un alcance de investigación correlacional teniendo en cuenta que buscó observar las posibles correlaciones entre una variedad de variables y la prevalencia de la desnutrición infantil en Ecuador sin realizar ninguna manipulación a la información. Esto significó mostrar la identificación de elementos que pueden estar relacionados con la desnutrición para así obtener una mejor comprensión de cómo se desarrolla en la población infantil ecuatoriana, centrándose especialmente en el Distrito Zona 5.

3.2 MÉTODOS UTILIZADOS EN LA INVESTIGACIÓN

La investigación del proyecto englobó los siguientes métodos: Deductivo y analítico.

3.2.1 Método Deductivo

En este proyecto se aplicó el razonamiento deductivo para encontrar un modelo predictivo que identifica la probabilidad de desnutrición en niños basándose en variables demográficas, epidemiológicas y socioeconómicas. Utilizando datos proporcionados por el Ministerio de Salud Pública de los años 2021 y 2022, se identificaron variables clave sobre sus relaciones con la desnutrición infantil. A continuación, se emplearon algoritmos de aprendizaje automático como regresión logística, *random forest, K-Nearest Neighbors* (KNN), árbol de clasificación y *XGBoost* para analizar y validar estas hipótesis. Los resultados del análisis permitieron derivar conclusiones específicas sobre los factores más influyentes en la desnutrición infantil en el Distrito Zona 5, proporcionando una base como información adicional para implementar intervenciones preventivas y terapéuticas.

3.2.2 Método Analítico

El método analítico fue utilizado para descomponer datos e información existente sobre la desnutrición infantil en Ecuador. Se examinó datos, factores determinantes y otros indicadores relevantes. Se obtuvo información por parte del Ministerio de Salud Pública (Zona 5). De ella se extrajo registros de salud y estudios previos sobre la desnutrición infantil. Luego, se realizó un análisis estadístico para identificar patrones y relaciones entre las variables.

3.3 CONSTRUCCIÓN METODOLÓGICA DEL OBJETO DE INVESTIGACIÓN

3.3.1 Técnicas de investigación

3.3.1.1. Entrevistas

Basado en una revisión de artículos académicos y publicaciones científicas en Degefa et al. (2022) y Elhady et al. (2023), las entrevistas se diseñaron mediante un enfoque estructurado. Estas entrevistas se efectuaron para determinar los factores que contribuyen a la desnutrición infantil. El uso de esta técnica permitió recopilar datos sobre las percepciones y experiencias de profesionales de la salud con respecto a la desnutrición infantil. Se formularon preguntas específicas y pertinentes para investigar las variables relacionadas con la desnutrición infantil. De esta manera, se recopilaron datos precisos necesarios para comprender mejor los factores que causan la desnutrición y crear estrategias de prevención e intervención.

3.3.1.2. Análisis exploratorio de datos

Se aplicaron diversas técnicas de análisis de datos. Se realizó una Exploración de Datos (EDA) para identificar la distribución y las relaciones entre las variables. El

propósito de esta técnica fue optimizar la base de datos y prepararla para un análisis posterior en la evaluación de los modelos elegidos.

3.3.1.3. Normalización y escalado

Posteriormente, se aplicaron métodos de normalización y escalado utilizando técnicas como máximo para asegurar que todas las variables numéricas tuvieran la misma escala [-3,3], lo cual es crucial para el rendimiento de los algoritmos de aprendizaje automático. Las técnicas que se utilizaron en el proceso fue la combinación de estandarización y ajuste de escalas para asegurar que los valores estén centrados y distribuidos en el rango deseado mejorando el rendimiento de los modelos para ser evaluados.

3.3.1.4. Análisis de correlación e imputación de valores perdidos

- **Correlación:** Se llevó a cabo un análisis para determinar cómo las variables identificadas (peso, talla, IMC) y la prevalencia de desnutrición infantil están relacionadas entre sí, dependiendo de la naturaleza de los datos y la distribución de las variables.
- **Imputación:** La imputación de datos se utilizó para rellenar los valores faltantes utilizando la mediana y asegurar que los datos de nuestra base estén completos para la evaluación de los modelos.

3.3.1.5. Recodificación y Aprendizaje Automático

- **Recodificación:** Utilizando la técnica de codificación binaria se realizó la recodificación de variables categóricas (cualitativas) transformando a valores binarios para que el modelo los pueda evaluar.
- **Algoritmos de aprendizaje automático:** Se utilizaron varios algoritmos para evaluar y generar un modelo predictivo.

3.3.1.6. Visualización de datos

Se crearon gráficos y diagramas para ilustrar los resultados facilitando la interpretación de los hallazgos.

3.3.2 Instrumentos de investigación

Para llevar a cabo el análisis y desarrollo del modelo predictivo de desnutrición infantil se utilizaron diversas herramientas de software y bibliotecas especializadas en análisis de datos.

- Python: El lenguaje de programación Python fue la base para la implementación del estudio, permitiendo la integración de múltiples librerías.
- Pandas: Librería para la manipulación y limpieza de datos.
- NumPy: Librería para operaciones numéricas avanzadas.
- Scikit-learn: Librería para la aplicación de algoritmos de aprendizaje automático y técnicas de preprocesamiento.
- Matplotlib y Seaborn: Librería para la creación de visualizaciones detalladas que facilitan la interpretación de los resultados.

3.3.2.1 Algoritmos de Aprendizaje Automático

Se utilizaron varios algoritmos de Aprendizaje Automático para construir y evaluar el modelo predictivo de desnutrición infantil. Los algoritmos escogidos para esta investigación fueron:

- Regresión Logística
- *Random Forest*
- *K-Nearest Neighbors* (KNN)
- Árbol de Clasificación

- *XGBoost*,

Cada uno de estos algoritmos ofreció ventajas específicas que, en conjunto, permitieron aportar información para un modelo predictivo válido para la ayuda de la desnutrición infantil.

3.4 ELABORACIÓN DEL MARCO TEÓRICO

Para el desarrollo del marco teórico del proyecto, se realizó una revisión de la literatura científica utilizando palabras clave como "desnutrición infantil" y "modelos predictivos" en bases de datos académicas como PubMed, Scopus y el motor de búsqueda Google Scholar. Se definieron conceptos clave relacionados con la desnutrición infantil y se exploraron técnicas de Ciencia de Datos aplicables, como regresión logística, árboles de clasificación y *XGBoost*, entre otros.

3.5 RECOLECCIÓN DE LA INFORMACIÓN

Para llevar a cabo este estudio se realizó una solicitud formal al Ministerio de Salud Pública para obtener los datos necesarios como se puede observar en el Anexo 3. La petición incluyó la solicitud de registros médicos y datos demográficos de niños menores de cinco años del Distrito Zona 5, correspondientes a los años 2021 y 2022. Estos datos fueron fundamentales para el análisis y el desarrollo del modelo predictivo.

Se utilizaron bases de datos que se encontraban divididas por meses correspondientes a los años 2021 y 2022, dando un total de 24 bases de datos. Estas bases de datos se encuentran en la plataforma de Registro de Atención en Salud del Ministerio de Salud Pública, distrito Zona 5, como parte del programa de monitoreo y evaluación periódica de pacientes pediátricos implementado en Ecuador. Abarcan variables críticas como Edad (años), Talla (cm) y Peso (kg), proporcionando datos

históricos necesarios para el desarrollo y validación del modelo predictivo de desnutrición infantil. Se realizó una petición con la ayuda de la institución Universidad Técnica Estatal de Quevedo dirigida al distrito Zona 5 de Quevedo para que nos facilitaran la base de datos, ya que en la plataforma solo ellos pueden visualizarla y no está accesible al público.

El Ministerio de Salud Pública gestiona este programa, que recopila información detallada sobre la salud infantil para crear una base de datos sólida y actualizada. La información se recopila de manera sistemática en intervalos regulares y abarca una variedad de indicadores de salud de los pacientes pediátricos.

3.6 PROCESAMIENTO Y ANÁLISIS

3.6.1 Selección de indicadores relevantes y aplicación de técnicas de estadísticas.

3.6.1.1. Parámetros médicos

La obtención de parámetros médicos se realizó mediante entrevistas a profesionales en el área de salud. Este informe recoge las opiniones y experiencias de diversos profesionales de la salud sobre la desnutrición infantil. Las entrevistas fueron realizadas para obtener una perspectiva amplia y detallada acerca de la situación actual, los desafíos, y las prácticas en el manejo de la desnutrición infantil en Ecuador.

3.6.1.2. Unificación, Revisión y preparación de datos

- Unificación: Se realizó la unificación de las 24 bases de datos individuales correspondientes a los registros mensuales de los años 2021 y 2022. El objetivo fue crear una base de datos consolidada que permita un análisis integral y continuo de los datos de los pacientes. Cada una de estas bases contenía entre

280,175 y 414,487 registros de pacientes infantiles, permitiendo un análisis detallado a nivel individual. Esta información, proporcionada por el Ministerio de Salud Pública de la Zona 5, abarcó las provincias de Santa Elena, Guayas, Los Ríos, Galápagos y Bolívar. La amplitud de los datos disponibles es aceptable para permitir la predicción adecuada de individuos en riesgo de desnutrición infantil.

- Revisión y preparación: Para la evaluación de algoritmos con el objetivo de generar un modelo de predicción, se comenzó revisando la base de datos proporcionada por el Ministerio de Salud Pública del Distrito Zona 5 de Ecuador. Fue necesario verificar la calidad de los datos, corregir valores faltantes y atípicos. Primero, se identificaron y analizaron los valores perdidos en el proceso de limpieza de datos, los errores de registro o la falta de seguimiento pueden causar datos insuficientes. Para tratar estos valores, se utilizaron métodos de imputación como la medida de tendencia central la mediana para garantizar que la base de datos fuera lo más completa y precisa posible, como lo indican Jakobsen et al. (2017). Si la cantidad de información faltante era significativa, se procedía a eliminar los registros con datos insuficientes.

3.6.1.3.Codificación y transformación

Con un enfoque basado en la Ciencia de Datos, se llevó a cabo un tratamiento de los datos para garantizar su adecuación al análisis. Esto incluyó:

- Normalización y estandarización. Se aplicó a para evitar la sensibilidad de los algoritmos a las escalas de las características, comparando las métricas con diferente rango permitiendo una comparación justa.

- Transformación de datos categóricos en variables ficticias para su uso en los modelos predictivos. Este proceso permitió estandarizar los datos y reducir posibles sesgos en el análisis utilizando la técnica codificación binaria.

3.6.1.4. Exploración de datos y análisis estadístico

En la Figura 1 se muestra el análisis exploratorio de datos que se realizó para comprender mejor la distribución y las características del conjunto de datos. Este análisis incluyó una evaluación estadística de las correlaciones entre las variables, proporcionando una visión clara de las relaciones y patrones presentes en los datos.

Figura 1 *Proceso de análisis correlacional*

Elaborado por: La autora (2024)

3.6.1.5. Identificación de Variables

Se identificaron varias variables que influyen en la desnutrición infantil. Estas variables se incluyeron, pero no se limitaron a:

- Variables demográficas: Edad, género.
- Variables de salud: Peso, talla, índice de masa corporal (IMC)
- Variables socioeconómicas y ambientales: Tipo de atención medica recibida.

3.6.1.6. Recolección de Datos

Los datos relacionados con estas variables se recopilaron de diversas fuentes, como:

- Entrevistas a profesionales de salud. Ayudaron a tener una idea de los datos cualitativos sobre factores asociados con la desnutrición infantil, registros médicos y de salud pública. Datos proporcionados por el Ministerio de Salud Pública del Ecuador.

3.6.1.7. Selección y cálculo de coeficiente de correlación

Para determinar y visualizar el coeficiente de correlación se utilizaron medidas como:

- Coeficiente de correlación de Pearson: El resultado de este análisis fue la determinación de la fuerza y la dirección de la relación entre las variables.
- Cálculo de coeficiente de correlación: Se utilizó un lenguaje de programación de alto nivel que permite hacer análisis estadísticos como Python para calcular la correlación existente entre las variables.

Los resultados del análisis de correlación permitieron identificar las variables que están más estrechamente relacionadas con la desnutrición infantil en Ecuador. Se prestó especial atención a las correlaciones significativas y a las variables que pueden tener un impacto considerable en el número de casos de desnutrición, específicamente en el Distrito Zona 5. Estas conclusiones ayudaron a comprender mejor las causas de la desnutrición infantil.

3.6.1.8. Visualización de datos

La visualización de las relaciones entre las variables se observó utilizando técnicas de matrices de correlación y mapas de calor mediante colores que indicaron la fuerza y la dirección de la relación.

3.6.1.9. Aplicación de Smote

Se aplicó la técnica de *Smote* para abordar el problema de desequilibrio de clases en el conjunto de datos proporcionados. Esta estrategia permitió aumentar la cantidad de datos de la clase subrepresentada sin simplemente duplicar registros, lo cual contribuyó a evitar el sobreajuste y a mejorar la precisión y la capacidad de generalización de los modelos entrenados.

3.6.2. Modelo Predictivo

3.6.2.1. División de base de datos

Galli (2023) señala que, sin la división de datos para evaluar, sería difícil detectar si el modelo está sobreajustando (*overfitting*) los datos de entrenamiento, lo que llevaría a un rendimiento pobre en datos reales. La división en conjuntos de entrenamiento y prueba se realizó para evaluar el desempeño de cada algoritmo para generar el modelo predictivo de manera objetiva. Este procedimiento permitió medir cómo de bien el modelo generaliza a datos no vistos, lo cual es esencial para predecir correctamente la desnutrición infantil en nuevas muestras que no fueron utilizadas durante el entrenamiento. Sin esta división, sería difícil detectar si el modelo está sobreajustando (*overfitting*) los datos de entrenamiento, lo que llevaría a un rendimiento pobre en datos reales.

3.6.2.2. Selección de Algoritmos

Se examinaron los algoritmos de aprendizaje automático seleccionando los adecuados para el problema de predicción de desnutrición infantil, tales como regresión logística, árboles de decisión, *random forest*, máquinas de soporte vectorial (SVM), KNN y *XGboost.*

Para justificar la elección de los algoritmos en la investigación sobre la predicción de la desnutrición infantil, se consideraron varios modelos basados en su eficacia y aplicabilidad en estudios recientes. La regresión logística es ampliamente utilizada por su capacidad para manejar problemas de clasificación binaria y ha demostrado ser efectiva en contextos médicos y sociales, como se evidencia en estudios como el de Asmare y Agmas (2022) sobre la coexistencia de desnutrición y anemia en niños menores de cinco años en Gambia y Rwanda. Dries et al. (2020) describe que los árboles de decisión y *Random Forest* son preferidos por su facilidad de interpretación y capacidad para manejar grandes conjuntos de datos con múltiples variables, como se discute en la revisión de técnicas robustas de regresión logística. Starbuck (2023) menciona que *XGBoost* ha sido reconocido por su alta precisión y eficiencia computacional, especialmente en grandes volúmenes de datos, lo cual es crucial para mejorar la exactitud de las predicciones.

Además, Asmare & Agmas (2022) referencia que el modelo *K-Nearest Neighbors* (KNN) y las *Support Vector Machines* (SVM) son útiles para capturar relaciones no lineales complejas, y han sido aplicados exitosamente en estudios epidemiológicos recientes para identificar factores de riesgo asociados con la desnutrición infantil.

3.6.2.3. Entrenamiento del Modelo

Los datos se dividieron en conjuntos de entrenamiento (80 %) y prueba (20 %) para entrenar los modelos seleccionados. Esta división permite evaluar la eficacia de los modelos en datos no vistos previamente y asegurar que su rendimiento generalice bien a nuevos datos. Al utilizar el conjunto de entrenamiento para ajustar los parámetros del modelo, ayuda a la precisión predictiva, mientras que el conjunto de prueba se reserva para una evaluación imparcial del rendimiento del modelo una vez que ha sido entrenado. Este enfoque es fundamental para validar la precisión del modelo predictivo.

3.6.3. Evaluación de Modelos

3.6.3.1. Métricas de Evaluación

Como se discute en la revisión de métricas de evaluación en algoritmos de aprendizaje automático publicada por Naidu et al. (2023), estas métricas son ampliamente utilizadas y recomendadas para evaluar modelos de clasificación en diversas aplicaciones, incluyendo estudios médicos y análisis de desnutrición infantil. Cada modelo fue evaluado mediante métricas de rendimiento específicas, como matriz de confusión, precisión, recall, F1-score y área bajo la curva ROC (AUC-ROC), para determinar su eficacia en la predicción de la desnutrición infantil. Los resultados obtenidos se presentaron a través de visualizaciones claras y detalladas, facilitando la interpretación y comunicación de los hallazgos. Estas técnicas combinadas permitieron un análisis exhaustivo y fundamentado de los datos, proporcionando una base sólida para las conclusiones y recomendaciones del estudio.

- **Matriz de Confusión**: Se generó la tabla que mostró las verdaderas etiquetas frente a las predicciones del modelo.
- **Precisión (Accuracy):** La métrica de precisión indicó la proporción de predicciones correctas sobre el total de predicciones.
- **Recall (Sensibilidad)**: La métrica indicó la proporción de verdaderos positivos sobre el total de positivos reales.
- **F1-Score**: se utilizó para evaluar el rendimiento de los modelos de clasificación, especialmente en situaciones con clases desbalanceadas. El F1-Score es la media armónica de la precisión y el *recall,* lo que lo convierte en una métrica equilibrada para medir la efectividad del modelo.
- **Área bajo la curva ROC (AUC-ROC)**: Esta medida se utilizó para evaluar el rendimiento del modelo en diferentes umbrales de clasificación, proporcionando una visión completa de su capacidad discriminativa. El AUC-ROC es especialmente útil en nuestro proyecto porque permite comparar la eficacia de los modelos en la identificación de casos positivos de desnutrición infantil, independientemente de los umbrales específicos de decisión. Un AUC-ROC más alto indica un mejor rendimiento del modelo, mostrando que es más efectivo en distinguir entre clases positivas y negativas.

3.6.3.2. Selección del Mejor Modelo

La selección de un modelo para la predicción de la desnutrición infantil, después de evaluar múltiples algoritmos, implicó en comparar las métricas de rendimiento y considerar otros factores relevantes para determinar el modelo más adecuado como el tiempo de ejecución.

CAPÍTULO IV

RESULTADOS Y DISCUSIÓN

4.1. INDICADORES RELEVANTES CON BASE EN LOS PARÁMETROS MÉDICOS ESTABLECIDOS Y APLICANDO TÉCNICAS ESTADÍSTICAS

En este apartado se presenta el proceso y resultados de la identificación y utilización de indicadores clave basados en parámetros médicos establecidos fundamentalmente para la construcción de un modelo predictivo. Para identificar los indicadores pertinentes, en este estudio se partió con entrevistas a profesionales de la salud y se aplicaron técnicas estadísticas.

4.1.1. Parámetros médicos

Al realizar las entrevistas a profesionales del área de salud se obtuvieron los siguientes resultados:

Una de las preguntas que se realizó fue la definición de desnutrición infantil. Los profesionales la definieron como una condición compleja que resulta de una ingesta insuficiente de nutrientes esenciales y necesarios para el crecimiento y desarrollo adecuado de los niños. Esta condición es influenciada por una variedad de factores, incluidos los socioeconómicos, que limitan el acceso a alimentos de calidad y servicios de salud adecuados. La falta de recursos económicos y educativos se identifica como un factor agravante, exacerbando la vulnerabilidad de los niños afectados.

También se estableció cuál es la edad en la que los niños son considerados con desnutrición. La mayoría de los profesionales de la salud coincidieron en que los niños menores de 5 años son los más vulnerables a la desnutrición, dado que esta etapa es crítica para el crecimiento y desarrollo. Sin embargo, también se reconoce la importancia de monitorear la nutrición en niños hasta los 9 años, ya que los efectos

de la desnutrición pueden manifestarse en etapas posteriores de la infancia y afectar el desarrollo a largo plazo. Los profesionales también comentaron sobre los principales desafíos en la detección y manejo de la desnutrición infantil, que incluyen la falta de educación nutricional entre padres y cuidadores, y la carencia de herramientas precisas y accesibles para evaluar el estado nutricional. La identificación de casos subclínicos y leves representa una dificultad significativa, impidiendo una intervención oportuna.

Otra de las preguntas apuntó a cuáles son los indicadores más útiles para la detección y predicción de la desnutrición infantil. Se conoció que los indicadores incluyen el peso, la talla, el índice de masa corporal (IMC) y el perímetro braquial. Estas mediciones antropométricas proporcionan una visión integral del estado nutricional del niño y son fundamentales para evaluar el riesgo de desnutrición y monitorizar el progreso del crecimiento y desarrollo.

Se obtuvo conocimiento también sobre la existencia de guías y estándares internacionales que abordan la desnutrición infantil, principalmente elaborados por la Organización Mundial de la Salud (OMS). Estas guías son ampliamente utilizadas en las instituciones de salud para el manejo de la desnutrición infantil, proporcionando un marco de referencia para su evaluación y tratamiento. Sin embargo, se reconoce que en algunas regiones hay una carencia de políticas específicas que regulen y controlen efectivamente esta condición.

4.1.2. Base de datos

4.1.2.1. Unificación

El proceso incluyó la importación y unificación de las 24 bases de datos, lo cual resultó en la creación de un único conjunto de datos integrado. Como resultado de

este proceso, se obtuvo un *dataframe* compuesto por 5,372,113 observaciones y un total de 110 variables. El principal objetivo de esta integración es disponer de un único conjunto de datos que sirva para entrenar los modelos predictivos de manera más eficiente.

4.1.2.2.Revisión y preparación de los datos

Después de realizar la unificación de las 24 bases de datos se procedió a la revisión de los datos para la detección de datos duplicados, valores faltantes o ausente utilizando la herramienta de Python para realizar la evaluación de algoritmos y encontrar un modelo de predicción para la desnutrición infantil. Una vez detectados se procedió a eliminar los registros con valores duplicados eliminando la filas y columnas completas por su alto porcentaje. Para corregir los valores faltantes se detalla a continuación el tratamiento que se utilizó en el estudio.

4.1.2.2. Tratamiento de valores faltantes

A continuación, se detalla el análisis y los pasos ejecutados para realizar la estimación de los modelos y obtener la información completa, eliminando así los valores perdidos. En la Tabla 2 se muestra un resumen de la cantidad de cada una de las variables de un conjunto de datos, información necesaria sobre el estado de los datos antes de proceder con el análisis y la construcción del modelo predictivo. Específicamente, indica el porcentaje de datos perdidos para cada variable.

Tabla 2 *Variables con valores perdidos*

Variable	**Porcentaje de valores perdidos**
ATEMED_TIP_ATE	0.85 %

PROF_SEXO	1.19 %
PCTE_PESO	7.33 %
PCTE_TALLA	8.62 %
PCTE_ULT_IMC	20.59 %

Elaborado por: La autora (2024)

Se puede notar que las variables con los valores perdidos más significativos son PCTE_ULT_IMC, que corresponde al Índice de Masa Corporal (IMC), con un 20.59 % de datos faltantes. Le siguen PCTE_TALLA con un 8.62 % y PCTE_PESO con un 7.33 % de valores perdidos.

4.1.2.3. Tratamiento de valores perdidos

Para variables con un porcentaje moderado de valores perdidos como PCTE_PESO, PCTE_TALLA y PCTE_ULT_IMC se utilizó como técnica de imputación la mediana.

En el caso de alto porcentaje de valores perdidos, se debe evaluar si es viable imputar estos valores o si es preferible eliminar los registros faltantes si la imputación no es adecuada.

En la Tabla 3 se observa que, una vez realizada la eliminación de los valores perdidos, los datos se encuentran completos y listos para proceder con la codificación de las variables.

Tabla 3 *Resultado de tratamiento de variables con valores atípicos.*

Variable	Porcentaje de Valores Perdidos
ATEMED_TIP_ATE	0 %
PROF_SEXO	0 %
PCTE_PESO	0 %
PCTE_TALLA	0 %
PCTE_ULT_IMC	0 %

Elaborado por: La autora (2024)

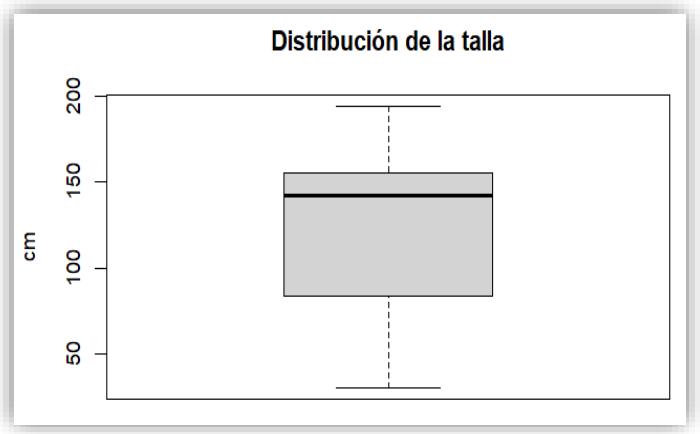
El proceso de tratamiento de valores faltantes resultó en un conjunto de datos completo y fiable, como se evidencia en la tabla 3, donde las 5 variables ahora muestran 0 % de valores perdidos. Esto aseguró que el modelo predictivo desarrollado sea fiable, permitiendo una detección temprana y efectiva de la desnutrición infantil. Una vez realizado el proceso de datos duplicados, valores faltantes se obtuvo una base de datos con 3,297,745 y una cantidad de 10 variables.

4.1.2.4. Identificación y tratamiento de valores atípicos

Para el siguiente paso antes de la estimación de los modelos, se realizó un análisis de valores atípicos con el objetivo de evitar sesgos en las estimaciones. Para ello, se utilizaron gráficos de caja.

En la Figura 2, se puede observar que, en cuanto a la talla, no se detectó la presencia de un sesgo marcado en la variable ni de valores perdidos. El diagrama de caja mostró que la mayoría de las tallas se encontraban entre aproximadamente 75 cm y 175 cm, con una mediana alrededor de los 140 cm. La dispersión de los datos fue relativamente uniforme y no se observaron valores atípicos significativos en la distribución de la talla.

Figura 2 *Diagrama de caja Talla*

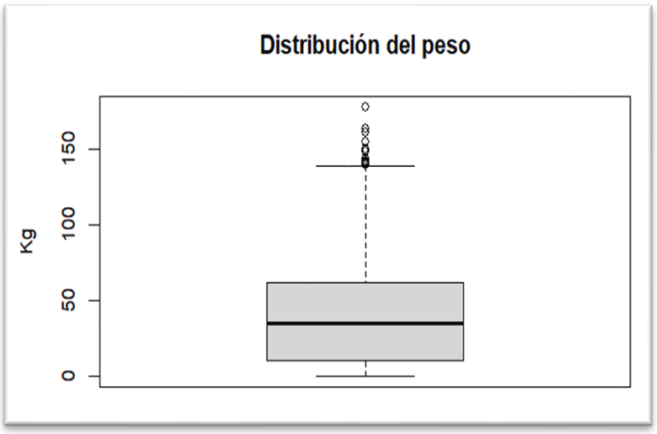


Elaborado por: La autora (2024)

El diagrama de caja de la Figura 3 muestra que la mayoría de los pesos se encontraban entre aproximadamente 25 kg y 75 kg, con una mediana de 50 kg. Sin embargo, se observaron varios valores atípicos por encima de los 100 kg, lo que indica la presencia de individuos con pesos significativamente mayores que el rango típico del conjunto de datos. La dispersión de los datos fue relativamente uniforme, con una clara identificación de los valores atípicos en la parte superior.

En resumen, según la Figura 3 no se observó una tendencia o existencia excesiva de valores atípicos en el peso, por lo que, a priori, el peso se considera adecuado. Además, al ser una variable estrechamente asociada con la nutrición de los menores, se procuró mantener esta característica inalterada.

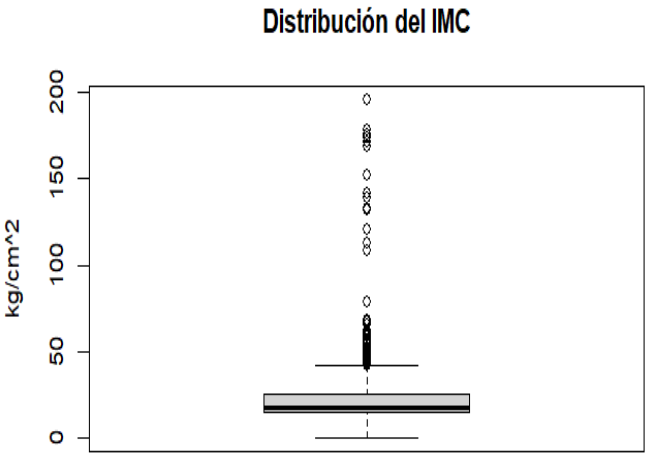
Figura 3 *Diagrama de caja Peso*



Elaborado por: La autora (2024)

Al observar la figura 4 se nota que en el IMC había muchos valores anómalos, lo que sesgaba la distribución de los datos. Se realizó una imputación de valores atípicos para corregir esta variable. La media no era una medida confiable para la imputación porque estos valores extremos sesgaban la distribución. Por lo tanto, para imputar los datos atípicos, se utilizó la mediana, que es menos sensible a los valores extremos.

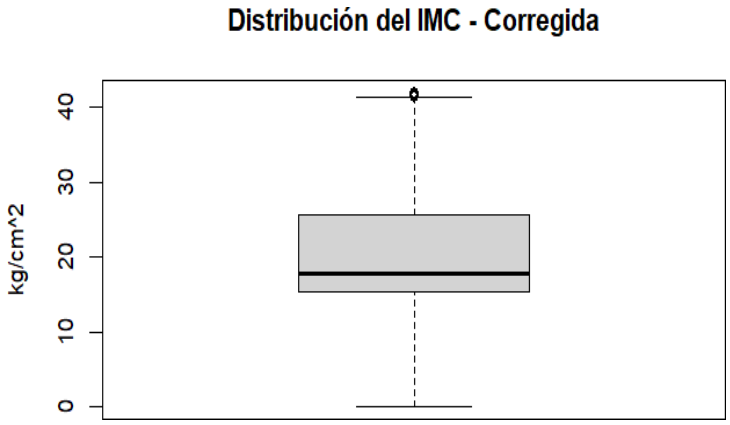
Figura 4 *Diagrama de caja con valores atípicos*



Elaborado por: La autora (2024)

Después de realizar esta imputación, se creó un nuevo diagrama de caja, como se muestra en la Figura 5, que representa la distribución corregida del IMC. La cantidad de valores atípicos se redujo significativamente con la corrección, lo que resultó en una distribución más uniforme y menos dispersa. La mayoría de los valores de IMC oscilaban entre 10 kg/m² y 30 kg/m², con una mediana de alrededor de 20 kg/m² y solo un valor excepcional por encima de 40 kg/m². Esto demostró que la variabilidad del IMC se había controlado y el sesgo se había eliminado, mejorando la calidad de los datos para el análisis posterior.

Figura 5 *Diagrama de caja IMC corregido*



Elaborado por: La autora (2024)

4.1.2.5. Codificación y transformación de datos

Para asegurar que todas las variables numéricas tengan la misma escala se procedió a realizar la normalización de variables. Además, se procedió a convertir las variables categóricas en variables binarias para su uso en modelos predictivos.

4.1.2.6. Normalización de variables

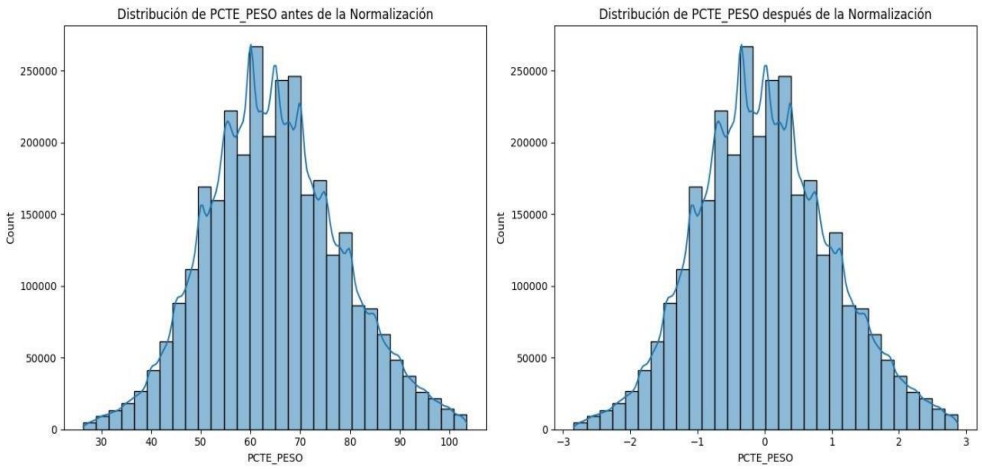
Se llevó a cabo el proceso de normalización con el objetivo de ajustar los valores de las características de las variables a una escala común, garantizando así el

rendimiento óptimo de los modelos predictivos. A continuación, se detallan las distribuciones de las variables antes y después de la normalización.

Distribución PCTE_PESO

En la Figura 6, el gráfico de la izquierda muestra la distribución de PCTE_PESO en su escala original, con valores que oscilan entre aproximadamente 30 y 100 kg. La distribución tiene una forma aproximadamente normal, con una alta densidad de datos alrededor de los 50 a 70 kg, indicando que la mayoría de los valores de peso se encuentran en este rango. La dispersión de los datos es mayor en los extremos, mostrando una disminución en la frecuencia de valores muy bajos y altos de peso.

Figura 6 *Distribución de las variables PCTE_PESO antes y después de la normalización*



Elaborado por: La autora (2024)

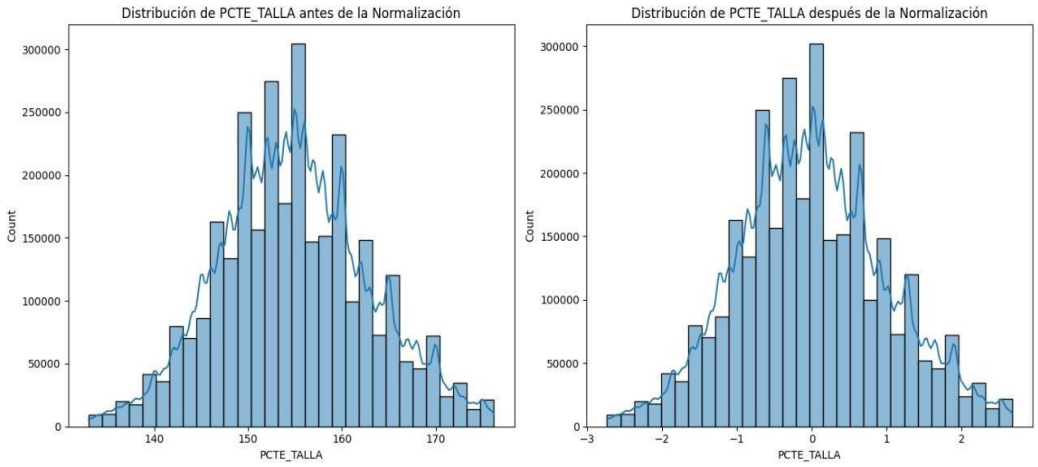
El gráfico de la derecha muestra la distribución de PCTE_PESO después de la normalización. Los valores han sido transformados a una escala centrada en torno a una media de 0 y una desviación estándar de 1. La forma de la distribución se mantiene similar a la original, pero ahora los datos están centrados alrededor de 0, con la mayoría de los valores oscilando entre -3 y 3. La normalización permite que los algoritmos de aprendizaje automático interpreten las características de manera

más uniforme, mejorando la eficiencia y precisión del modelo al evitar que algunas características dominen sobre otras debido a diferencias en la escala.

Distribución de PCTE_TALLA

En la Figura 7, el gráfico de la izquierda muestra la distribución de PCTE_TALLA en su escala original, con valores que oscilan entre aproximadamente 130 y 180 cm. La distribución tiene una forma aproximadamente normal, con una alta densidad de datos alrededor de los 150 a 160 cm, indicando que la mayoría de los valores de talla se encuentran en este rango. Además, la dispersión de los datos es mayor en los extremos, mostrando una disminución en la frecuencia de valores muy bajos y altos de talla.

Figura 7 *Distribución de las variables PCTE_TALLA antes y después de la normalización*



Elaborado por: La autora (2024)

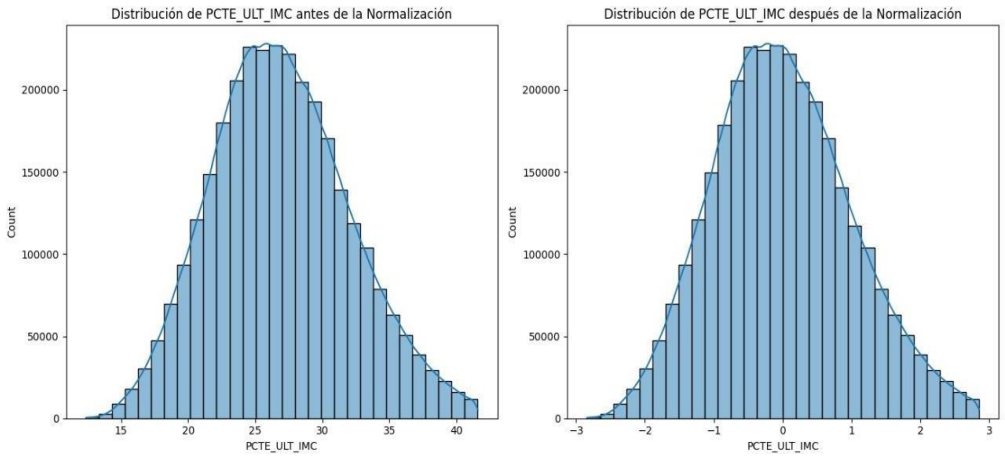
Al aplicar la normalización, en la Figura 7, el gráfico de la derecha muestra la distribución de PCTE_TALLA después de la normalización. Los valores han sido transformados a una escala centrada en torno a una media de 0 y una desviación estándar de 1. La forma de la distribución se mantiene similar a la original, pero ahora los datos están centrados alrededor de 0, con la mayoría de los valores

oscilando entre -3 y 3. La normalización permite que los algoritmos de aprendizaje automático interpreten las características de manera más uniforme, mejorando la eficiencia y precisión del modelo al evitar que algunas características dominen sobre otras debido a diferencias en la escala.

Distribución de PCTE_ULT_IMC

En la Figura 8, el gráfico de la izquierda muestra la distribución de PCTE_ULT_IMC en su escala original, con valores que oscilan entre aproximadamente 15 y 40. La distribución tiene una forma aproximadamente normal, con una alta densidad de datos alrededor de los 20 a 30, indicando que la mayoría de los valores de IMC se encuentran en este rango. La dispersión de los datos es mayor en los extremos, mostrando una disminución en la frecuencia de valores muy bajos y altos de IMC.

Figura 8 *Distribución de las variables PCTE_UTL_IMC antes y después de la normalización*



Elaborado por: La autora (2024)

En la Figura 8 se muestra la distribución original de PCTE_ULT_IMC. El gráfico de la derecha muestra la distribución de PCTE_ULT_IMC después de la normalización. Los valores han sido transformados a una escala centrada en torno a

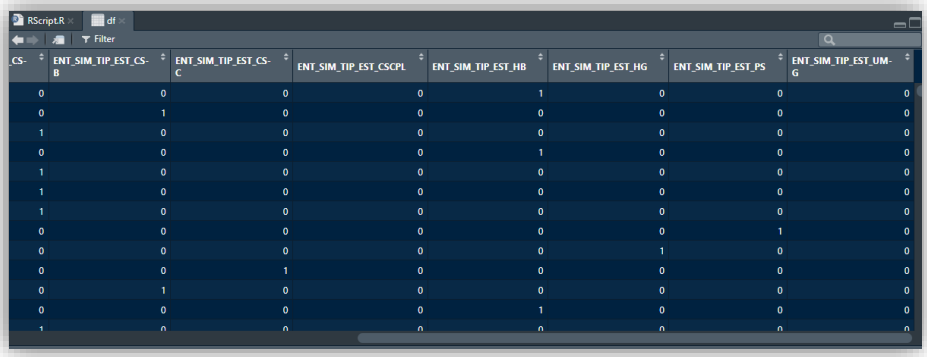
una media de 0 y una desviación estándar de 1. La forma de la distribución se mantiene similar a la original, pero ahora los datos están centrados alrededor de 0, con la mayoría de los valores oscilando entre -3 y 3.

4.1.2.7. Transformación de datos categóricos

Dentro de este proceso, después de la selección de variables, fue necesario realizar una recodificación. Dado que los modelos predictivos no pueden interpretar información de tipo texto, se requería transformar las variables categóricas en valores binarios. En esta codificación, un valor de 1 representa la presencia de una característica y un valor de 0 su ausencia. Este paso fue crucial porque los modelos predictivos solo comprenden y se entrenan con datos numéricos. Por lo tanto, las variables categóricas, tanto multinomiales como dicotómicas, se transformaron en vectores binarios.

En la Figura 9 se muestra que, para las variables categóricas con más de dos categorías, se creó una nueva variable para cada una de las categorías, siguiendo el mismo procedimiento que para las variables dicotómicas. Cuando una observación contenía la característica específica, se le asignaba un valor de 1; de lo contrario, se le asignaba un valor de 0.

Figura 9 Matriz de correlación de variables



Elaborado por: La autora (2024)

Al realizar el procedimiento de transformación de variables categóricas, la base de datos resultante contiene una mayor cantidad de variables debido a la creación de variables binarias para la recodificación de categorías. La Tabla 4 proporciona una descripción detallada de cada una de estas variables, facilitando la interpretación del significado y contexto de la base de datos unificada.

Tabla 4 *Descripción de variables*

Variables	Descripción
ATEMED_TIP_ATE	Tipo de atención médica recibida
PROF_SEXO	Indicador de género del profesional
PCTE_PESO	Peso del paciente
PCTE_TALLA	Altura del paciente
IMC	Índice de Masa Corporal del paciente
desnutrición	Estado patológico
ENT_SIM_TIP_EST_CS-A	Tipo de estado de salud relacionado con el código A
ENT_SIM_TIP_EST_CS-B	Tipo de estado de salud relacionado con el código B

ENT_SIM_TIP_EST_CS-C	Tipo de estado de salud relacionado con el código C.
ENT_SIM_TIP_EST_CSCPL	Estado de salud relacionado con el código CSCPL.
ENT_SIM_TIP_EST_HB	Estado de salud relacionado con el código HB
ENT_SIM_TIP_EST_HG	Estado de salud relacionado con el código HG
ENT_SIM_TIP_EST_PS	Estado de salud relacionado con el código PS
ENT_SIM_TIP_EST_UM-G	Estado de salud relacionado con el código UM-G

Elaborado por: La autora (2024)

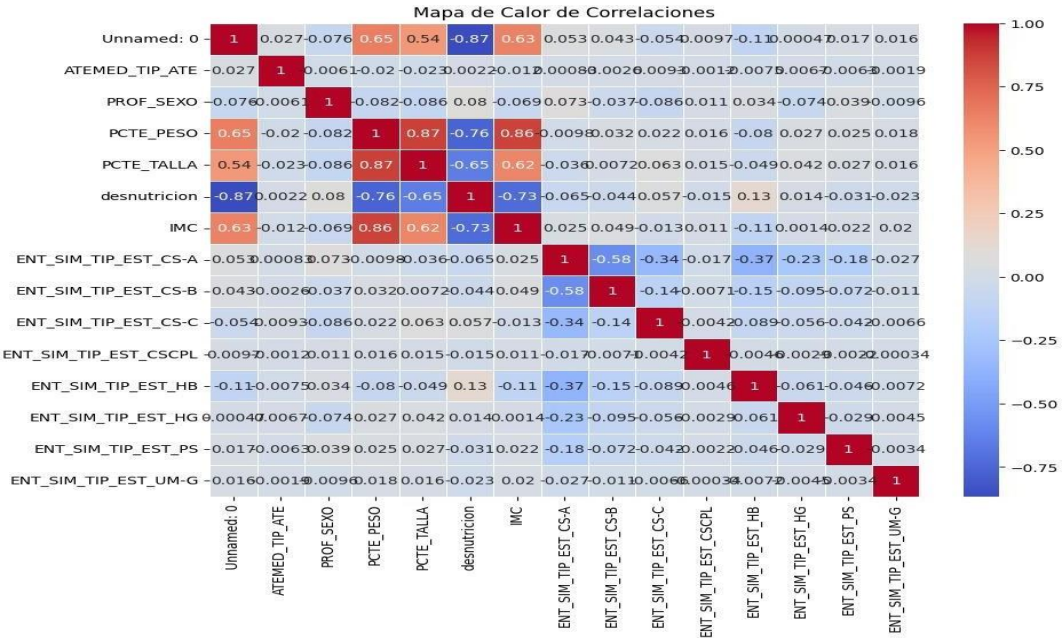
4.1.2.8. Análisis exploratorio de datos

Para optimizar la base de datos, se realizó un análisis exploratorio de datos (EDA). Este proceso permitió comprender mejor la distribución de los datos, identificar patrones relevantes y asegurar la calidad de la información mediante la detección y manejo de valores perdidos y atípicos. Este proceso permitió no solo la integración de la información, sino también la reducción de datos redundantes, proporcionando así una base de datos más limpia y manejable. A continuación, se realiza un análisis exploratorio a los datos obtenidos por el Ministerio de Salud Pública de la Zona 5.

4.1.2.9.Análisis de correlación

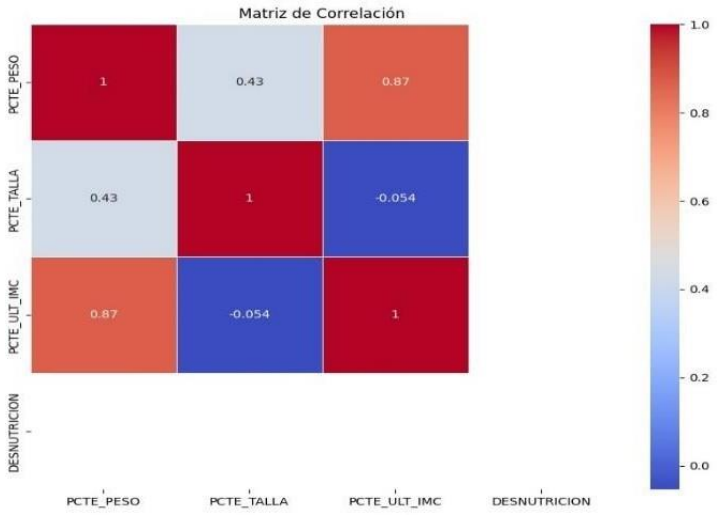
En la Figura 10 se ilustra la matriz de correlación entre las variables del conjunto de datos mediante un mapa de calor. Esta visualización permite observar las correlaciones entre las diferentes variables del conjunto de datos de manera clara.

Figura 10 *Matriz de correlación de las variables*



Elaborado por: *La autora (2024)*

Figura 11 *Matriz de correlación de variables seleccionadas.*



Elaborado por: La autora (2024)

En la Figura 11 se visualiza un mapa de calor que proporciona una visión rápida de las relaciones entre las variables del conjunto de datos. Las variables relacionadas con las medidas físicas del paciente (peso, talla, IMC) tienden a mostrar correlaciones fuertes entre sí, mientras que las variables categóricas relacionadas con la atención médica y el estado de salud muestran correlaciones más bajas. Esta información es útil para identificar relaciones importantes entre variables.

Tabla 5 *Resultado de variables seleccionadas*

Variables
PCTE_PESO
PCTE_TALLA
PCTE_UTL_IMC
DESNUTRICION

Elaborado por: La autora (2024)

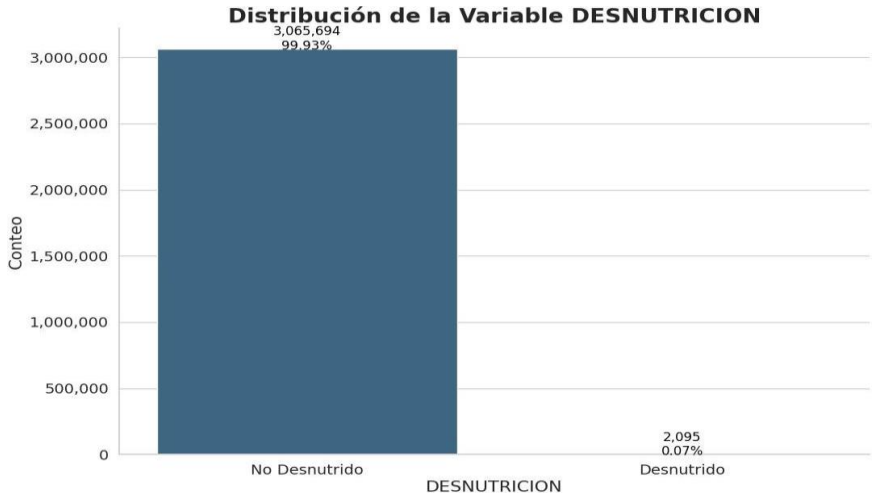
En la Tabla 5, se observa que los resultados de la matriz de correlación sugirieron que tanto la talla como el peso eran variables importantes para un modelo predictivo de desnutrición infantil. La fuerte correlación entre el peso y el IMC, así como la moderada correlación entre la talla y el IMC, indicaron que el peso era el predictor más influyente del IMC, aunque la talla también contribuyó significativamente y no debía ser descartada.

4.1.3. Balanceo de datos

Para garantizar el entrenamiento de los modelos predictivos, fue importante realizar una revisión del balanceo de datos de la variable a predecir, en este caso, la desnutrición infantil. El balanceo de datos ayudó a la distribución equitativa de las

clases dentro de la variable dependiente, en otras palabras, que estén representadas de manera proporcional para evitar sesgos en el modelo. Este paso es influyente en el análisis de desnutrición infantil, donde la representación precisa de casos tanto de desnutrición como de no desnutrición es fundamental para desarrollar modelos predictivos.

Figura 12 *Datos desbalanceados de la variable a predecir Desnutrición*



Elaborado por: La autora (2024)

La Figura 12 muestra que hay un desbalance significativo en la variable DESNUTRICION. Se puede observar que la clase "No Desnutrido" tiene 3,065,694 casos, lo que representa el 99.93 % de los datos. Adicional, se observa a la clase "Desnutrido" tiene solo 2,095 casos, lo que representa el 0.07 % de los datos. Al realizar la revisión se puede observar que la variable objetivo desnutrición está extremadamente desbalanceada. Este desbalance puede ser problema al entrenar modelos predictivos, ya que los modelos pueden sesgarse hacia la clase mayoritaria (No Desnutrido) y tener un rendimiento pobre en la clase minoritaria (Desnutrido). Esto puede llevar a conclusiones erróneas y decisiones subóptimas en la gestión y prevención de la desnutrición infantil.

Este proceso, conocido como balanceo de datos, mostró el número de observaciones en cada categoría de la variable respuesta. Como se observa en la Figura 12, los datos estaban desbalanceados, con una preponderancia significativa de una sola categoría. Este desbalance provocó problemas de sobreentrenamiento en los modelos predictivos, afectando su capacidad para generalizar y hacer predicciones precisas en nuevas muestras.

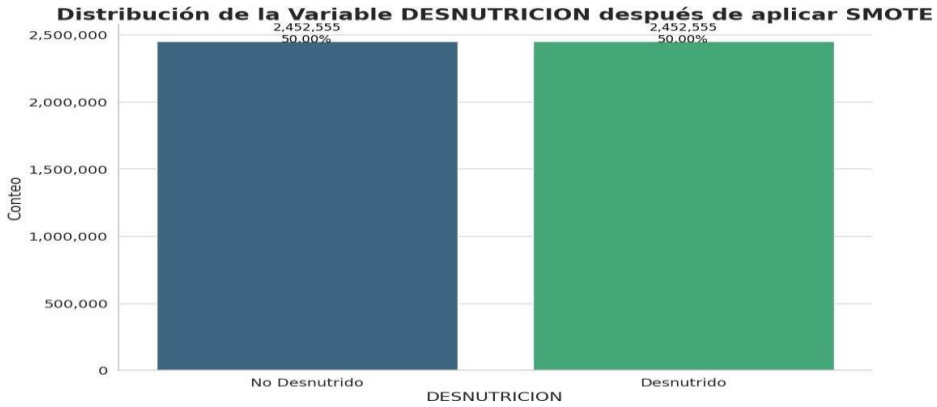
4.1.4. Aplicación de SMOTE

Para mitigar el impacto del desbalance de clases, se aplicó la técnica de *smote* (Técnica de Sobremuestreo de Minorías Sintéticas) al conjunto de entrenamiento. *Smote* generó ejemplos sintéticos de la clase minoritaria basados en las características de las muestras existentes, equilibrando así la distribución de las clases. Como muestra la Figura 13, después de aplicar *smote*, la distribución de la variable desnutrición se balanceó perfectamente, donde se observa que:

No Desnutrido: 2,452,555 casos (50 %)

Desnutrido: 2,452,555 casos (50 %)

Figura 13 *Datos balanceados de la variable a predecir Desnutrición*



Elaborado por: La autora (2024)

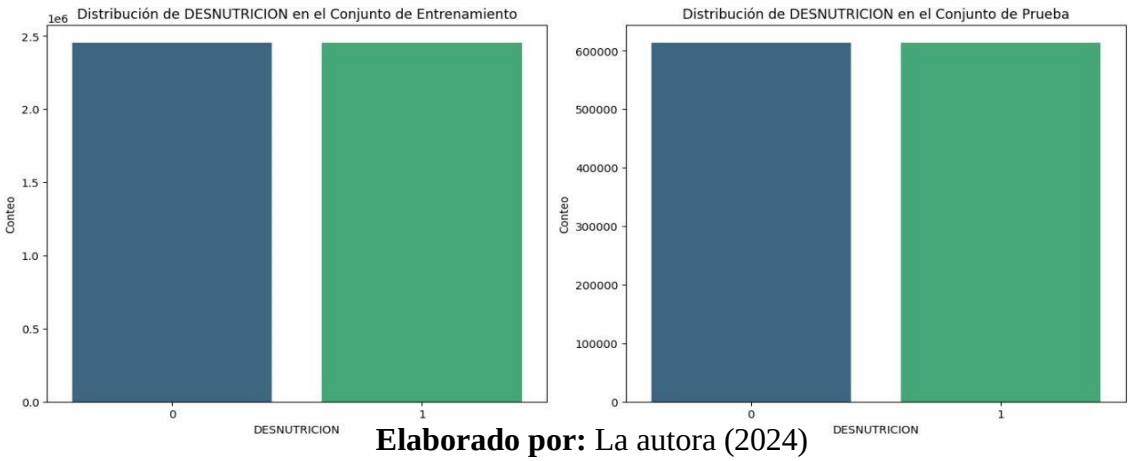
La Figura 13 muestra la nueva distribución balanceada de las clases. Ambos grupos, "No Desnutrido" y "Desnutrido", tienen el mismo número de observaciones,

2,452,555 casos cada uno, representando el 50 % del total de datos en el conjunto de entrenamiento.

La aplicación de *SMOTE* fue efectiva para balancear las clases de la variable desnutrición, lo que es fundamental para mejorar la precisión y la capacidad predictiva de los modelos de aprendizaje automático en el análisis de la desnutrición infantil. Este enfoque garantizó que ambos grupos, desnutridos y no desnutridos, sean tratados con igual importancia, permitiendo desarrollar un modelo para la toma de decisiones en la gestión de la desnutrición infantil en el Ecuador, Distrito Zona 5. Este procedimiento aseguró que los datos utilizados para entrenar el modelo sean representativos y que el modelo pueda ser evaluado de manera efectiva en un conjunto de datos no visto durante el entrenamiento, proporcionando una estimación precisa de su rendimiento en datos reales.

Para confirmar que los conjuntos de entrenamiento y prueba estén balanceados, se generó gráficos de barras que muestran la distribución de la variable DESNUTRICION en ambos conjuntos. Estos gráficos indicaron que ambos conjuntos mantienen un balance similar, lo que fue crucial para el desarrollo y evaluación de un modelo predictivo robusto.

Figura 14 *Gráfico de distribución de variable Desnutrición*



Elaborado por: La autora (2024)

La Figura 14 muestra los gráficos de distribución que verifican que la división de la base de datos ha mantenido el balance de las clases, asegurando que el modelo predictivo tenga suficiente información representativa de ambas clases durante el entrenamiento y la evaluación.

4.1.5. Discusión objetivo 1

La identificación y utilización de indicadores clave para la construcción de un modelo predictivo de desnutrición infantil, basado en parámetros médicos establecidos y técnicas estadísticas aplicadas. Los hallazgos se obtuvieron a partir de entrevistas a profesionales de la salud y la unificación y análisis de múltiples bases de datos.

Las entrevistas a los profesionales de salud confirmaron que la desnutrición infantil es una condición compleja influenciada por una variedad de factores socioeconómicos, limitando el acceso a alimentos de calidad y servicios de salud adecuados. Los profesionales destacaron que los niños menores de 5 años son los más vulnerables, aunque también se reconoció la importancia de monitorear la nutrición en niños hasta los 9 años.

Onis et al. (2007) señalan que los indicadores más útiles para la detección y predicción de la desnutrición infantil incluyen peso, talla, índice de masa corporal (IMC). Estos hallazgos están en línea con estudios previos que han identificado estas medidas antropométricas como fundamentales para evaluar el estado nutricional de los niños.

Ergul Aydin y Kamisli Ozturk, (2024) indican que este enfoque es consistente con metodologías recomendadas en estudios de manejo de datos faltantes. La integración de 24 bases de datos resultó en un conjunto de datos con 5,372,113

observaciones y 110 variables. Este proceso permitió crear una base de datos robusta para entrenar los modelos predictivos. La revisión y preparación de datos incluyó la eliminación de duplicados y el tratamiento de valores faltantes, utilizando la imputación de la mediana para variables con valores moderadamente perdidos, como el peso, la talla y el IMC.

En estudios recientes, Easily (2024) y Kalaivani & Ranichitra (2024) sugieren métodos como la imputación utilizando modelos de regresión, que son menos sensibles a los *outliers*, mejorando así la robustez del modelo. El análisis de valores atípicos mediante gráficos de caja reveló una distribución adecuada para las variables talla y peso, pero identificó valores anómalos en el IMC. La imputación de valores atípicos con la mediana mejoró la distribución de datos, alineándose con prácticas recomendadas para minimizar el sesgo en análisis predictivos.

Llego (2023) indica que la transformación de distribuciones a una media de 0 y desviación estándar de 1 sigue siendo una técnica para mejorar la calidad del modelo, especialmente en contextos de Aprendizaje Automático, asegurando que todas las variables contribuyan de manera equitativa al modelo. La normalización de las variables fue crucial para asegurar una escala común y mejorar el rendimiento del modelo predictivo.

Kalaivani y Ranichitra (2024) indican que la recodificación de variables categóricas en valores binarios facilita su uso en modelos predictivos, siguiendo prácticas comunes en el preprocesamiento de datos. Estas prácticas siguen siendo relevantes, y se han desarrollado herramientas y bibliotecas modernas en Python y R que simplifican este proceso, permitiendo una transformación eficiente y precisa de los datos. Las bibliotecas como pandas y scikit-learn en Python, así como dplyr y caret

en R, proporcionan funciones integradas para realizar estas transformaciones de manera eficiente y precisa, mejorando la preparación de datos para el análisis predictivo.

Easily (2024) afirma que el análisis exploratorio de datos (EDA) permitió identificar patrones relevantes y asegurar la calidad de la información. La matriz de correlación mostró que el peso y la talla eran importantes predictores del IMC y, por ende, de la desnutrición infantil. Estos hallazgos están respaldados por estudios que destacan la importancia de las medidas antropométricas en la evaluación del estado nutricional. El uso de EDA no solo facilita la identificación de patrones en los datos, sino que también ayuda a detectar y corregir errores, garantizando así la integridad y precisión del análisis posterior.

La aplicación de *SMOTE* para balancear las clases de la variable DESNUTRICION fue una estrategia efectiva para desarrollar modelos predictivos robustos y precisos. Este enfoque asegura que ambos grupos, desnutridos y no desnutridos, sean tratados con igual importancia en el proceso de modelado, mejorando la precisión del modelo y proporcionando una base más confiable para la toma de decisiones en la gestión de la desnutrición infantil en el Ecuador, Distrito Zona 5.

Estudios recientes han demostrado la eficacia de *SMOTE* en el tratamiento de conjuntos de datos desbalanceados. Por ejemplo, Hassannataj Joloudari et al. (2023) mostraron que el uso combinado de *SMOTE* con redes neuronales convolucionales (CNN) puede mejorar significativamente la precisión en la clasificación de datos desbalanceados, alcanzando una precisión del 99.08 % en sus experimentos. Este resultado resalta la capacidad de *SMOTE* para mejorar el rendimiento de los modelos al equilibrar la distribución de clases minoritarias y mayoritarias.

Asimismo, la revisión de Galli (2023) confirma que *SMOTE* es una técnica valiosa en el *toolkit* de aprendizaje automático para tratar con conjuntos de datos desbalanceados. Al generar datos sintéticos para la clase minoritaria, *SMOTE* ayuda a los modelos a aprender de manera más equitativa, lo que resulta en predicciones más fiables y justas.

4.2. MODELOS Y ALGORITMOS DE APRENDIZAJE AUTOMÁTICO EXISTENTES PARA DATOS RELACIONADOS CON LA DESNUTRICIÓN INFANTIL EN ECUADOR

El presente apartado describe la búsqueda, el entrenamiento, comparación y evaluación de los modelos de predicción con el objetivo de obtener los mejores resultados en términos de precisión y optimización de recursos computacionales, tanto durante el entrenamiento como en las predicciones con cada modelo. A continuación, se redacta los resultados obtenidos tratando de alcanzar este objetivo.

4.2.2. Proceso de División de la Base de Datos

Para evaluar y entrenar el modelo predictivo de desnutrición infantil, primero se procedió a realizar una división de la base de datos balanceada en dos conjuntos: un conjunto de entrenamiento y un conjunto de prueba, en una proporción de 80 % y 20 % respectivamente como se muestra. La división se realizó utilizando la función *train_test_split de scikit-learn*, asegurando que la proporción de las clases en la variable objetivo (DESNUTRICION) se mantuviera constante en ambos conjuntos mediante una división estratificada.

Un division de 80/20 proporciona que haya suficientes ejemplos tanto para entrenar como para probar el modelo, reduciendo la variabilidad en la estimación del rendimiento del modelo y proporcionando un balance adecuado entre el ajuste del

modelo y su validación como lo indica Sachin (2024).

En la división de la base de datos las características utilizadas fueron PCTE_PESO, PCTE_TALLA, y PCTE_ULT_IMC. Además, la variable objetivo fue DESNUTRICION, codificada como 1 para desnutrido y 0 para no desnutrido.

El conjunto de entrenamiento se dividió en *"X_train"* que Contiene 80 % de las observaciones de las características balanceadas y "*Y_train*" que contiene 80 % de las observaciones de la variable objetivo-balanceada. Además, la división del conjunto de prueba se dió en *X_test* contiene 20 % de las observaciones de las características balanceadas y "*Y_test*" contiene 20 % de las observaciones de la variable objetivo-balanceada. En la Tabla 6 se puede observar las dimensiones de los conjuntos aplicadas para la división de datos en entrenamiento y prueba.

Tabla 6 *Resultado de dimensiones para división de datos entrenamiento y prueba*

Conjunto	Características	Observaciones
Entrenamiento (X_train, y_train)	(2680596, 3)	2,680,596
Prueba (X_test, y_test)	(670149, 3)	67,0149

Elaborado por: La autora (2024)

4.2.3. Selección de modelos de Aprendizaje automático

La selección de los modelos de aprendizaje utilizados en este estudio se fundamenta en su capacidad para manejar grandes volúmenes de datos, su efectividad en la predicción y su interpretación de resultados como lo indica Pérez y Luis (2014). Se seleccionaron varios modelos de aprendizaje automático, incluyendo Regresión Logística, Máquinas de Vectores de Soporte (SVM), Bosques Aleatorios (*Random Forest*), Vecinos más Cercanos (KNN), Árboles de decisión y XGBoost.

4.2.3.1. Modelo de Regresión Logística

El modelo fue evaluado utilizando un conjunto de datos de prueba, como se muestra en la Figura 15. Los 2,763 pertenecían a la clase 0 y 2,877 a la clase 1. Se utilizó la función *classification_report* de scikit-learn para calcular las métricas de precisión, recall y F1-score. La precisión indica la proporción de verdaderos positivos sobre el total de predicciones positivas; el recall mide la proporción de verdaderos positivos sobre el total de ejemplos positivos reales, y el F1-score es la media armónica de precisión y recall. Además, se calculó el soporte, que refleja el número de ocurrencias de cada clase en el conjunto de datos de prueba.

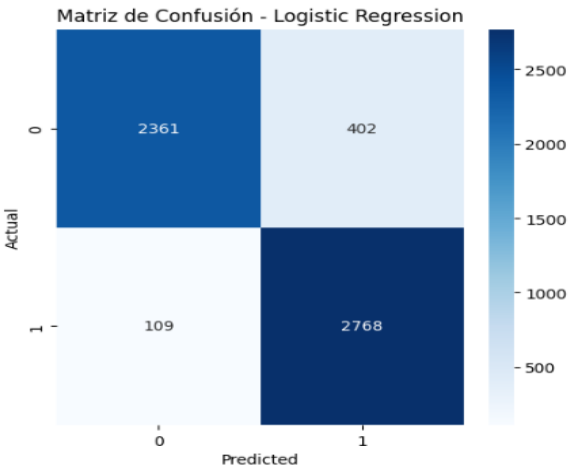
Figura 15 *Porcentaje de acierto para Regresión Logística*

Classification Report:

	precision	recall	f1-score	support
0	0.96	0.85	0.90	2763
1	0.87	0.96	0.92	2877
accuracy			0.91	5640
macro avg	0.91	0.91	0.91	5640
weighted avg	0.91	0.91	0.91	5640

Elaborado por: La autora (2024)

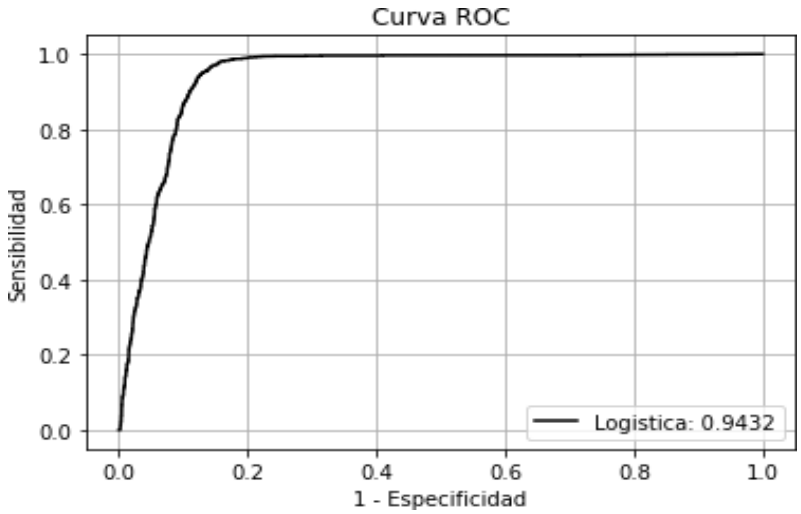
Figura 16 *Matriz de confusión de Regresión logística*



Elaborado por: La autora (2024)

El modelo presentó un rendimiento sólido y equilibrado en la clasificación de datos como muestra curva ROC. Para la clase 0, alcanzó una precisión del 96 %, un recall del 85 % y un F1-score de 0.90. Para la clase 1, la precisión fue del 87 %, con un recall del 96 % y un F1-score de 0.92. La precisión global del modelo fue del 91 %, con promedios macro y ponderados de precisión, recall y F1-score de 0.91, indicando un rendimiento consistente en todas las clases. El alto recall para la clase 1 destacó la capacidad del modelo para identificar la mayoría de los ejemplos positivos.

Figura 17 Porcentaje curva Roc



Elaborado por: La autora (2024)

4.2.3.2.Modelo SVM

La Figura 18 ilustra el modelo de Máquina de Vectores de Soporte (SVM) evaluado en este estudio. Muestra un desempeño notablemente robusto y equilibrado en la tarea de clasificación. La precisión global del modelo es del 91 %, lo que indica una alta eficacia general en sus predicciones.

Figura 18 *Porcentaje de acierto para SVM*

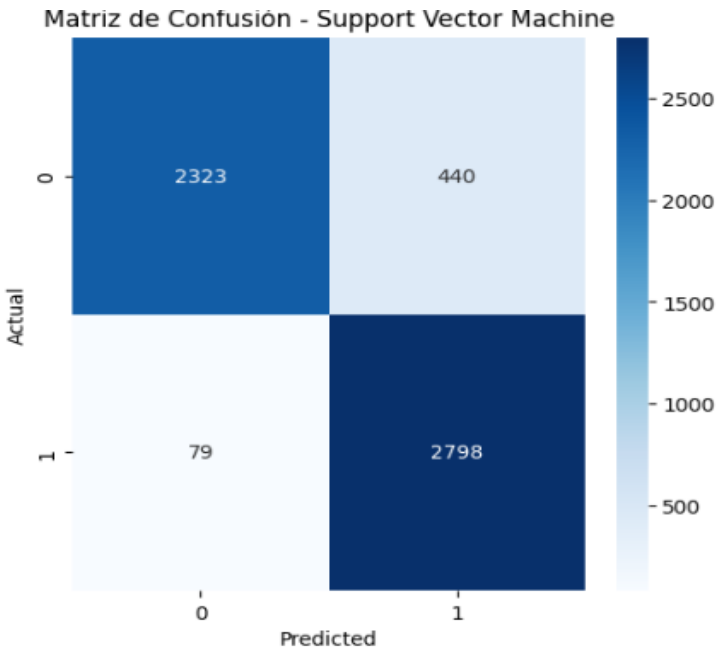
```
Classification Report:
              precision    recall  f1-score   support

           0       0.97      0.84      0.90      2763
           1       0.86      0.97      0.92      2877

    accuracy                           0.91      5640
   macro avg       0.92      0.91      0.91      5640
weighted avg       0.91      0.91      0.91      5640
```

Elaborado por: La autora (2024)

Figura 19 *Matriz de confusión de SVM*

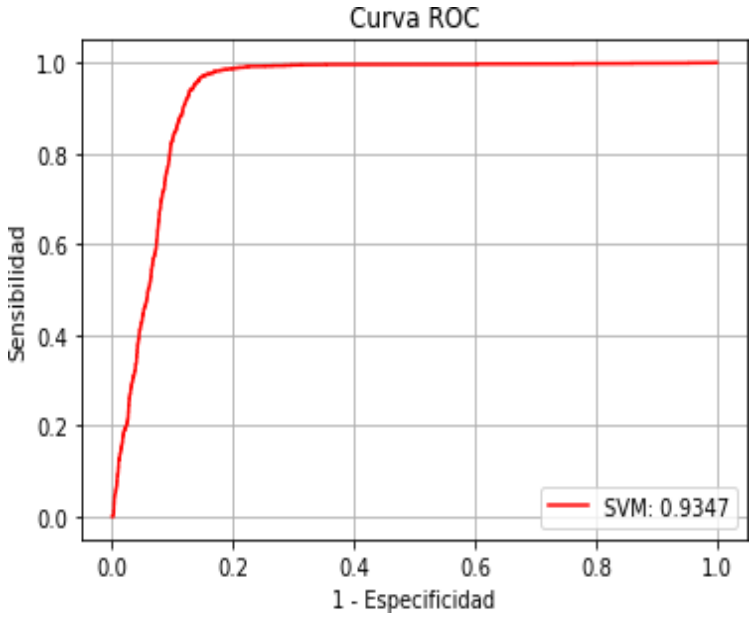


Elaborado por: La autora (2024)

Utilizando los datos de la Figura 19, el modelo mostró un rendimiento equilibrado y fiable en la clasificación de datos. Para la clase 0, alcanzó una precisión del 97 %, un recall del 84 % y un F1-score de 0.90, indicando un buen equilibrio entre precisión y recall, aunque con margen para mejorar la sensibilidad. Para la clase 1, la precisión fue del 86 %, con un recall del 97 % y un F1-score de 0.92, demostrando

una capacidad robusta para detectar ejemplos positivos. Las métricas globales y promedios, tanto macro como ponderados, fueron consistentes, con valores de 0.90 a 0.92 para precisión, recall y F1-score, sugiriendo un rendimiento equilibrado y fiable en ambas clases. Este desempeño es crucial para aplicaciones prácticas donde la precisión y la capacidad de detección son igualmente importantes.

Figura 20 *Porcentaje Curva ROC SVM*



Elaborado por: La autora (2024)

Para el modelo SVM evaluado en este estudio, como se visualiza en la Figura 20, la curva Roc muestra un AUC de 0.9347, lo que significa una alta efectividad en la clasificación, siendo capaz de diferenciar bien entre las clases positiva y negativa. Es altamente efectivo y confiable para la tarea de clasificación en el conjunto de datos analizado. La alta precisión en la clase 0 y el excelente recall en la clase 1 destacaron la capacidad del modelo para realizar predicciones precisas y capturar ejemplos positivos. Estas características hacen del modelo SVM una herramienta adecuada y balanceada para aplicaciones donde tanto la precisión como la sensibilidad son esenciales.

4.2.3.3. Modelo Árbol de Decisión (Decision Tree)

El modelo de Árbol de Decisión fue evaluado utilizando un conjunto de datos de prueba divididos en dos clases: 2,763 ejemplos de la clase 0 y 2,877 de la clase 1. El modelo evaluado, como se muestra en la Figura 21 y en la Figura 22, tuvo un rendimiento equilibrado y efectivo en la clasificación de datos. Para la clase 0, alcanzó una precisión del 88 %, un recall del 89 % y un F1-score de 0.88, reflejando un buen equilibrio entre precisión y recall. Para la clase 1, el modelo presentó una precisión del 89 %, un recall del 88 % y un F1-score de 0.89, indicando un desempeño robusto en la identificación de ejemplos positivos.

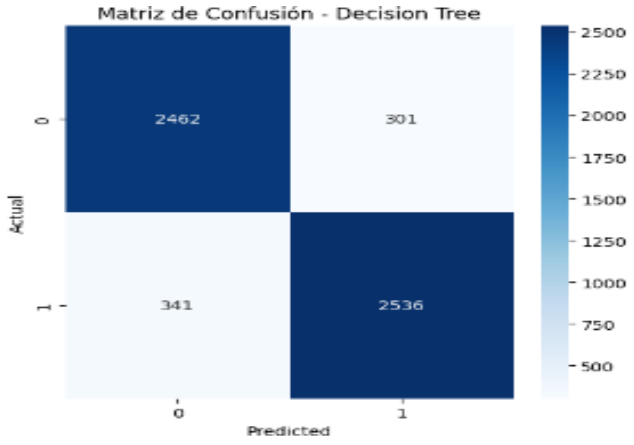
Figura 21 *Porcentaje de acierto para Árbol de Decisión*

Classification Report:

	precision	recall	f1-score	support
0	0.88	0.89	0.88	2763
1	0.89	0.88	0.89	2877
accuracy			0.89	5640
macro avg	0.89	0.89	0.89	5640
weighted avg	0.89	0.89	0.89	5640

Elaborado por: La autora (2024) ´

Figura 22 *Matriz de Confusión Árbol de decisión*

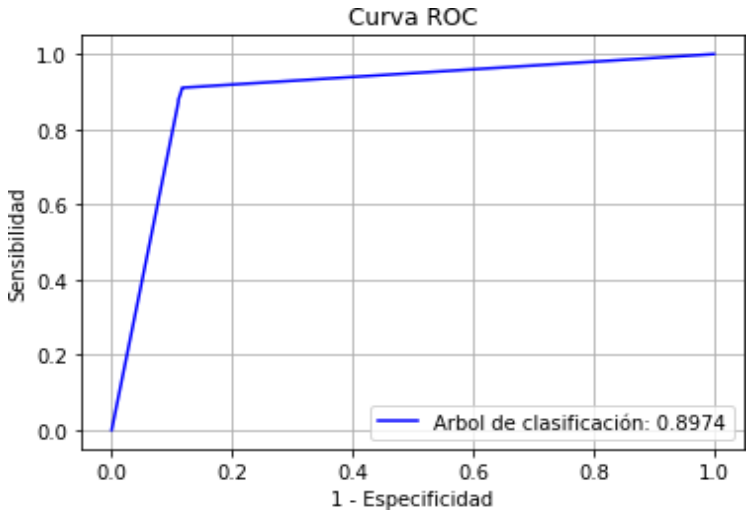


Elaborado por: La autora (2024)

Las métricas globales, incluyendo precisión, recall y F1-score, fueron consistentemente del 89 %, subrayando la efectividad general del modelo. Estas métricas sugieren que el modelo es confiable y mantiene un rendimiento sólido y equilibrado en ambas clases, haciendo del modelo una herramienta adecuada para aplicaciones donde tanto la precisión como la sensibilidad son cruciales.

En la Figura 23 la curva ROC y el AUC de 0.8974 indican que el modelo de Árbol de Clasificación tiene un buen rendimiento en la discriminación entre las clases positivas y negativas, aunque no es perfecto. Esta información es útil para evaluar la eficacia del modelo y considerar posibles mejoras o ajustes.

Figura 23 Porcentaje Curva ROC Árbol de decisión



Elaborado por: La autora (2024)

4.2.3.4. Modelo Random Forest

El modelo *Random Forest* fue evaluado utilizando un conjunto de datos de prueba empleado los datos de la matriz de confusión que ilustra la Figura 26, divididos en dos clases: 2,763 ejemplos de la clase 0 y 2,877 de la clase 1 como muestra la Figura 24.

Figura 24 *Porcentaje de acierto para Random Forest*

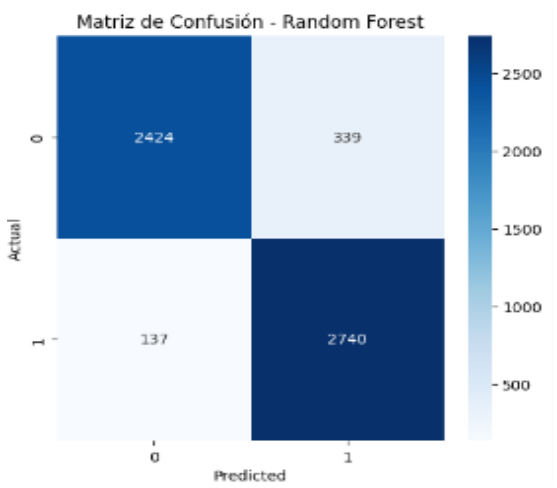
```
Classification Report:
              precision    recall  f1-score   support

           0       0.95      0.88      0.91      2763
           1       0.89      0.95      0.92      2877

    accuracy                           0.92      5640
   macro avg       0.92      0.91      0.92      5640
weighted avg       0.92      0.92      0.92      5640
```

Elaborado por: La autora (2024)

Figura 25 *Matriz de Confusión para Random Forest*

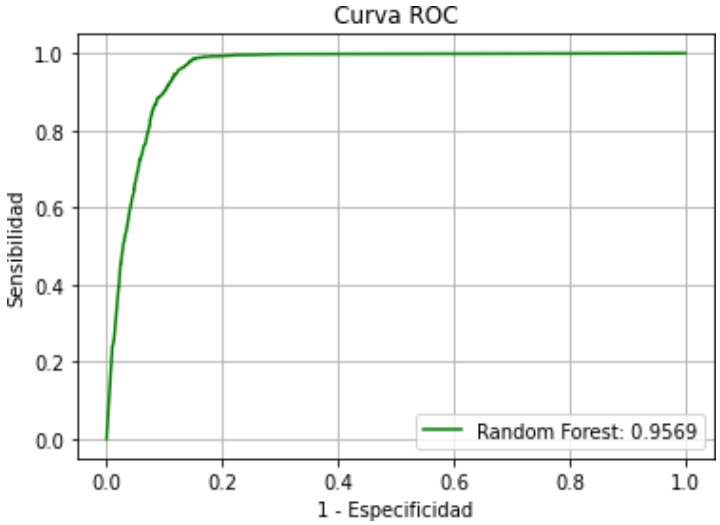


Elaborado por: La autora (2024)

El modelo *Random Forest* demostró un rendimiento sólido y equilibrado en la clasificación de datos. Para la clase 0, alcanzó una precisión del 95 %, un recall del 88 % y un F1-score de 0.91. Para la clase 1, presentó una precisión del 89 %, un recall del 95 % y un F1-score de 0.92. Las métricas globales del modelo incluyeron una precisión del 91 %, con promedios macro y ponderados de precisión, recall y F1-score entre 0.91 y 0.92, indicando un rendimiento consistente y equilibrado en ambas clases.

La Figura 26 muestra un AUC de 0.9569, indicando una excelente capacidad de discriminación entre clases, destacando su efectividad y confiabilidad para aplicaciones prácticas donde la precisión y la capacidad de detección son cruciales.

Figura 26 *Porcentaje Curva ROC para Random Forest*



Elaborado por: La autora (2024)

4.2.3.5. Modelo K-Nearest Neighbors (KNN)

En la Figura 27 se visualiza que, el modelo KNN fue evaluado utilizando un conjunto de datos de prueba divididos en dos clases: 2763 ejemplos de la clase 0 y 2877 de la clase 1.

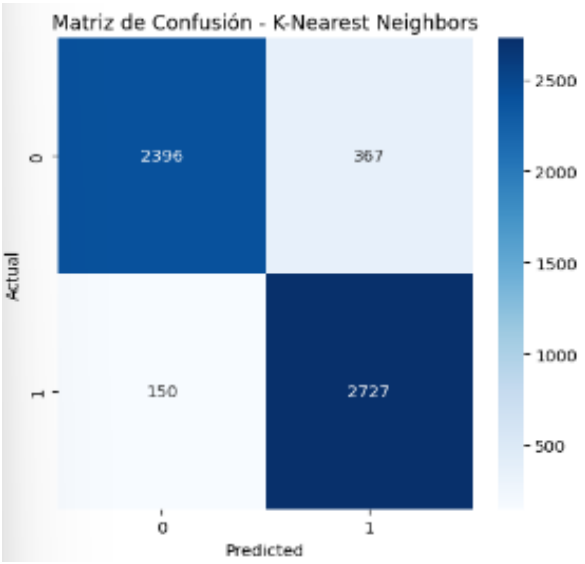
Figura 27 *Porcentaje de acierto para KNN*

Classification Report:

	precision	recall	f1-score	support
0	0.94	0.87	0.90	2763
1	0.88	0.95	0.91	2877
accuracy			0.91	5640
macro avg	0.91	0.91	0.91	5640
weighted avg	0.91	0.91	0.91	5640

Elaborado por: La autora (2024)

Figura 28 *Matriz de Confusión KNN*

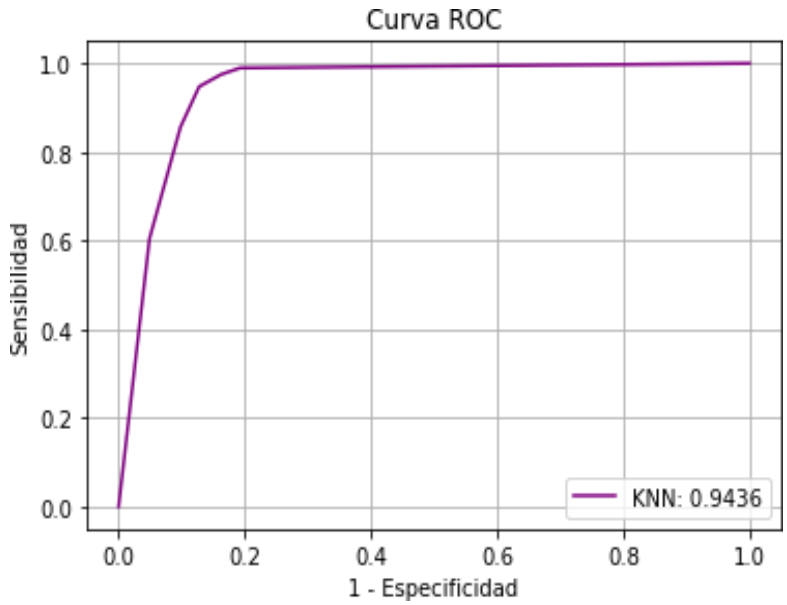


Elaborado por: La autora (2024)

Como muestra la Figura 28, el modelo KNN demostró un rendimiento sólido y equilibrado en la clasificación de datos. Luego de evaluar el modelo utilizando los datos de la matriz de confusión, se observa que para la clase 0 alcanzó una precisión del 94 %, con un recall del 87 % y un F1-score de 0.90, sugiriendo un buen equilibrio entre precisión y recall. Para la clase 1, la precisión fue del 88 %, el recall del 95 % y el F1-score de 0.91, indicando un rendimiento robusto y equilibrado.

Como se observa en la Figura 29, las métricas globales, incluyendo precisión, recall y F1-score, fueron consistentemente del 91 %, subrayando un rendimiento consistente en ambas clases. La alta capacidad del modelo para distinguir entre clases, reflejada en un AUC de 0.9436, destaca su efectividad y confiabilidad para aplicaciones prácticas donde tanto la precisión como la capacidad de detección son cruciales. Estas métricas resaltan la capacidad del modelo KNN para proporcionar predicciones precisas y fiables, lo que es esencial para su aplicación en escenarios de salud pública y otras áreas donde la clasificación precisa es fundamental.

Figura 29 *Porcentaje Curva ROC para KNN*



Elaborado por: La autora (2024)

4.2.3.6. Modelo XGBoost

El modelo *XGBoost* fue evaluado en un conjunto de datos de prueba compuesto por 2763 ejemplos de la clase 0 y 2877 de la clase 1 mostrados en la Figura 30. Este modelo utilizó un algoritmo de *boosting* (Minimizar errores) para construir un conjunto de árboles de decisión secuenciales, donde cada árbol corrige los errores del anterior, lo que resulta en un modelo más preciso y robusto.

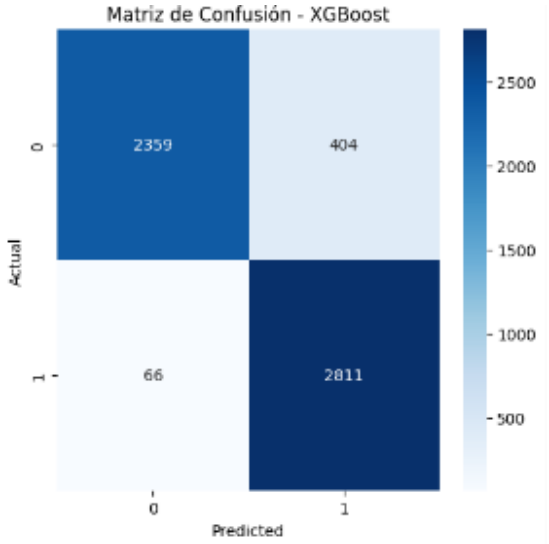
Figura 30 *Porcentaje de acierto para XGBoost*

Classification Report:

	precision	recall	f1-score	support
0	0.97	0.85	0.91	2763
1	0.87	0.98	0.92	2877
accuracy			0.92	5640
macro avg	0.92	0.92	0.92	5640
weighted avg	0.92	0.92	0.92	5640

Elaborado por: La autora (2024)

Figura 31 *Matriz de Confusión XGBoost*

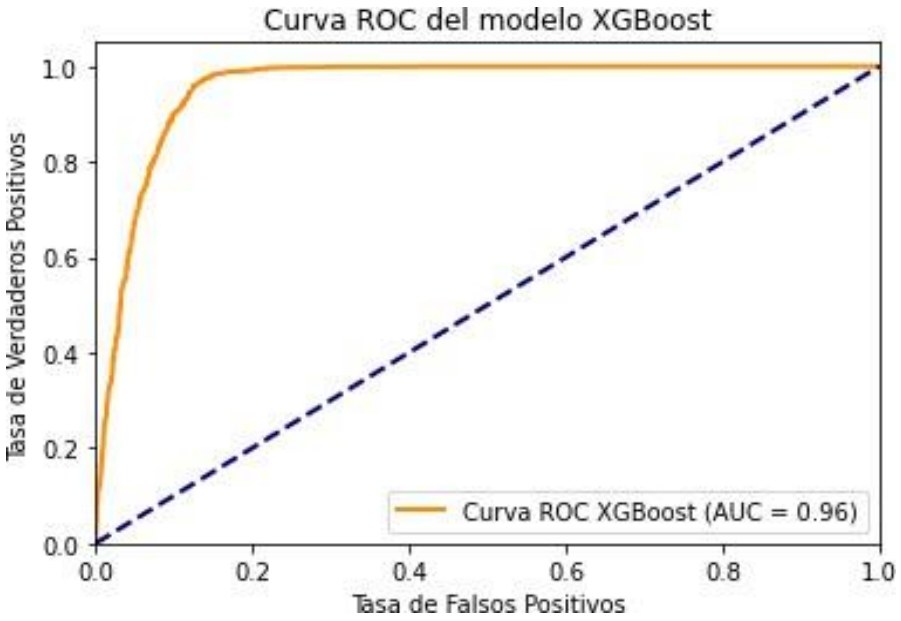


Elaborado por: La autora (2024)

El modelo *XGBoost* mostró un rendimiento sólido y equilibrado en la clasificación de datos. Luego de utilizar los datos de la matriz de confusión, como se muestra en la Figura 31, la clase 0 alcanzó una precisión del 97 %, un recall del 85 % y un F1-score de 0.91, reflejando un buen equilibrio entre precisión y recall. Para la clase 1, la precisión fue del 87 %, con un recall del 98 % y un F1-score de 0.92, indicando un rendimiento robusto y equilibrado. Las métricas globales del modelo incluyeron Las métricas globales, incluyendo precisión, recall y F1-score, fueron consistentemente del 92 %, sugiriendo un rendimiento consistente en ambas clases.

La curva ROC, como se visualiza en la Figura 32, con un AUC de 0.96, destacó la capacidad del modelo para distinguir entre clases positivas y negativas, mostrando una alta tasa de verdaderos positivos y una baja tasa de falsos positivos. Esto refuerza la confiabilidad y efectividad del modelo *XGBoost* en aplicaciones prácticas.

Figura 32 *Porcentaje Curva XGBoost*



Elaborado por: La autora (2024)

Tabla 7 *Resumen de evaluación de modelos*

Modelo	Accuracy	Sensibilidad	Especificidad	F1 - Score	Tiempo de entrenamiento (s)
Logístico	0,91	0,91	0,91	0,91	3,004
Support Vector Machine	0,91	0,91	0,91	0,91	67,003
Random Forest	0,91	0,92	0,91	0,91	2,708
KNN	0,91	0,91	0,91	0,91	0,084
Árbol de clasificación	0,89	0,89	0,89	0,89	0,163
XGBoost	0,92	0,92	0,92	0,92	2,819

Elaborado por: La autora (2024)

Como muestra la Tabla 7, entre los modelos evaluados, *XGBoost* se destacó como el más robusto y preciso, seguido por *Random Forest* y SVM. La elección del modelo más adecuado dependerá del contexto específico y de las necesidades en términos de precisión y eficiencia computacional. Estos resultados subrayan la importancia de utilizar técnicas avanzadas de aprendizaje automático para abordar problemas críticos de salud pública como la desnutrición infantil en Ecuador.

4.2.4. Discusión del objetivo 2

Fisher y C. Rosella, (2022) destacan la importancia de seleccionar y evaluar modelos de aprendizaje automático para abordar la desnutrición infantil en Ecuador, enfatizando la capacidad de estos modelos para manejar grandes volúmenes de datos y proporcionar predicciones precisas y fiables. La selección de estos modelos se fundamenta en estudios previos y literatura reciente que demuestran la efectividad de ciertos algoritmos en la clasificación de problemas de salud pública. La evaluación cuidadosa de estos modelos es crucial para implementar soluciones efectivas que mejoren la salud y el bienestar de los niños en el país.

Se evaluaron los siguientes modelos: Regresión Logística, Máquinas de Vectores de Soporte (SVM), Bosques Aleatorios (*Random Forest*), Vecinos más Cercanos (*KNN*), Árboles de Decisión y *XGBoost*. La evaluación se realizó utilizando un conjunto de datos balanceado, dividido en 80 % para entrenamiento y 20 % para prueba mediante la función *train_test_split* de *scikit-learn* con división estratificada. La división estratificada es una técnica comúnmente utilizada en la investigación de Aprendizaje Automático para asegurar que la proporción de clases en la variable objetivo se mantenga constante en los conjuntos de entrenamiento y prueba. Esto es

crucial para evitar sesgos en la evaluación del rendimiento del modelo, especialmente cuando se trabaja con datos desbalanceados. Según Morgenstern et al. (2020), la estratificación ayuda a mantener la representatividad de las clases minoritarias y mayoritarias en los conjuntos de datos, lo que mejora la robustez y la validez de los resultados del modelo.

Los modelos se seleccionaron por su capacidad para manejar datos de alta dimensionalidad y su eficacia en tareas de clasificación. Estudios recientes han demostrado que estos algoritmos proporcionan un buen equilibrio entre precisión y capacidad de generalización, cruciales en aplicaciones de salud pública.

El uso de modelos de aprendizaje automático como la Regresión Logística, SVM, Random Forest, KNN, Árboles de Decisión y *XGBoost* en la evaluación de problemas de salud pública ha sido bien documentado en la literatura. Un estudio realizado en Corea por Kim et al. (2022) sobre síndromes metabólicos demostró que los modelos de bosques aleatorios y *XGBoost* eran superiores en términos de AUC y F1-score, destacando su efectividad y precisión. Los resultados se evaluaron utilizando métricas como precisión, recall, F1-score y el área bajo la curva ROC (AUC).

Comparando estos resultados con otros trabajos similares, un estudio sobre desnutrición infantil en Etiopía comparó cinco algoritmos de aprendizaje automático, incluyendo *XGBoost*, Random Forest, KNN y regresión logística. En este estudio, Fisher y Rosella (2022) evaluaron el modelo *XGBoost* mostrando una alta precisión y siendo particularmente efectivo para predecir la desnutrición, alcanzando un AUC de 0.85 para todas las características. Esto destaca la robustez

del modelo *XGBoost* para la predicción de condiciones de salud, similar a los resultados obtenidos en el presente proyecto en Ecuador.

4.3. MODELO CON RENDIMIENTO FAVORABLE CON MAYOR GRADO DE PRECISIÓN EN LAS PREDICCIONES DE LA DESNUTRICIÓN INFANTIL MEDIANTE LOS RESULTADOS DEL CONJUNTO DE DATOS DE VALIDACIÓN.

Se identificó el modelo de aprendizaje automático que ofreció el mejor rendimiento en términos de precisión para predecir la desnutrición infantil. Para lograrlo, se evaluaron varios algoritmos de clasificación utilizando un conjunto de datos de validación. Esta etapa permitió determinar cuál de los modelos probados proporcionaba las predicciones más precisas y fiables.

La evaluación de los modelos se realizó mediante diversas métricas de rendimiento, incluyendo precisión, sensibilidad, especificidad, F1-Score y el área bajo la curva ROC (AUC-ROC). La selección del modelo óptimo se basó tanto en su capacidad para predecir correctamente los casos de desnutrición como en su habilidad para generalizar y manejar datos nuevos con eficacia. Este proceso garantizó que el modelo elegido fuera el más adecuado para la identificación y prevención de la desnutrición infantil.

4.3.1. Interpretaciones

4.3.1.1.Regresión Logística

La regresión logística demostró un rendimiento sólido en la clasificación con una precisión, sensibilidad, especificidad y F1-score de 0.91. Estas métricas indican que

el modelo fue altamente eficaz en predecir correctamente tanto las clases positivas como las negativas. Sin embargo, su tiempo de entrenamiento de 3.004 segundos fue moderado en comparación con otros modelos, sugiriendo que, aunque el modelo tiene una eficiencia computacional aceptable, no destacó por ser el más rápido.

4.3.1.2. SVM (Support Vector Machine)

El modelo SVM también mostró un rendimiento consistente con todas las métricas de evaluación en 0.91, indicando su capacidad para clasificar correctamente tanto las clases positivas como las negativas en una alta proporción. Sin embargo, su tiempo de entrenamiento de 67.003 segundos fue considerablemente más largo en comparación con otros modelos, lo que sugiere una eficiencia computacional más baja.

4.3.1.3. Random Forest

Random Forest mostró un rendimiento sólido con una precisión, especificidad y F1-score de 0.91, y una sensibilidad ligeramente más alta de 0.92, indicando su eficacia en detectar casos positivos. Además, su tiempo de entrenamiento de 2.708 segundos fue moderado, sugiriendo una eficiencia computacional aceptable.

4.3.1.4. K-Nearest Neighbors (KNN)

El modelo KNN mostró un rendimiento sólido con todas las métricas de evaluación en 0.91, destacando por su tiempo de entrenamiento extremadamente rápido de solo 0.084 segundos, lo que indica una eficiencia computacional muy alta. Esta característica lo hace

especialmente atractivo para conjuntos de datos grandes donde la velocidad de entrenamiento es crucial.

4.3.1.5. Árbol de Clasificación

El modelo de árbol de clasificación mostró un rendimiento ligeramente inferior en comparación con otros modelos, con todas las métricas de evaluación en 0.89, sugiriendo que podría ser menos preciso en la clasificación. Sin embargo, su tiempo de entrenamiento extremadamente bajo de 0.163 segundos indicó una eficiencia computacional muy alta.

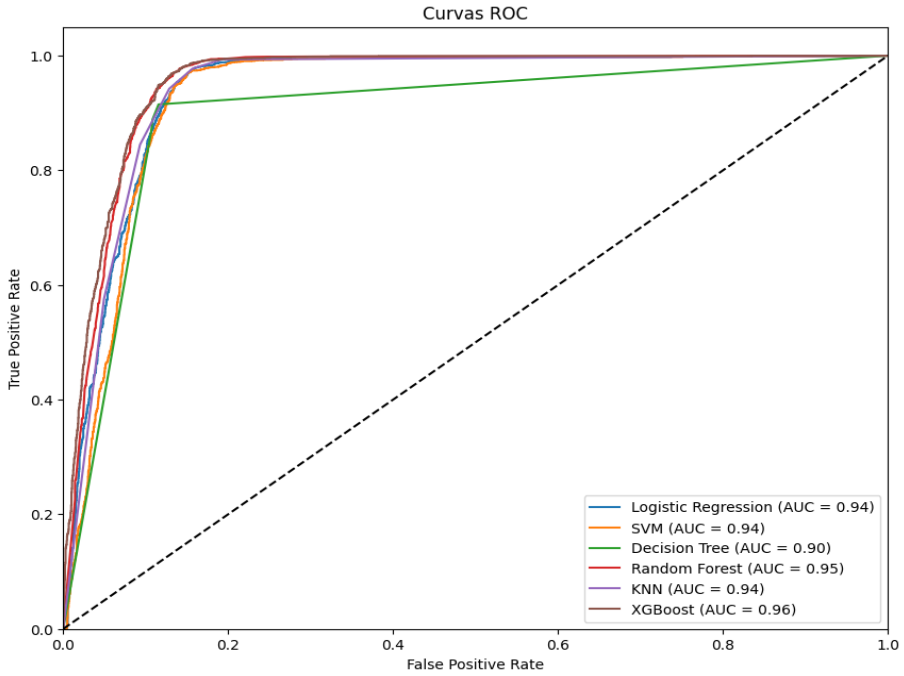
4.3.1.6. XGBoost

XGBoost destacó como el modelo con el mejor rendimiento general, con una precisión, sensibilidad, especificidad y F1-score de 0.92, indicando su capacidad para predecir correctamente tanto las clases positivas como las negativas en una alta proporción. Su tiempo de entrenamiento de 2.819 segundos fue moderado, sugiriendo una eficiencia computacional aceptable, especialmente considerando su rendimiento superior.

4.3.1.7. Curva ROC

En la Figura 33, para una comparación más visual y utilizando otra medida comparativa, se observó la curva ROC para cada uno de los modelos. Los resultados mostraron que todos los modelos tenían una capacidad predictiva similar, con más del 90 % de predicciones correctas. Sin embargo, a pesar de su alta capacidad predictiva, el modelo de Support Vector Machine (SVM) resultó ser ineficiente y costoso en términos computacionales. Alternativas más ligeras, como el modelo *XGBoost* y la regresión logística, demostraron ser mejores opciones para las predicciones, no solo por su menor costo computacional sino también por su capacidad predictiva comparable.

Figura 33 *Porcentaje de la curva ROC*



Elaborado por: La autora (2024)

4.3.1.8. Tiempo de Entrenamiento

El tiempo de entrenamiento varió significativamente entre los modelos. KNN fue el más rápido, con solo 0.084 segundos, destacándose por su alta eficiencia temporal. En contraste, SVM fue el más lento, con un tiempo de entrenamiento de 67.003 segundos, lo que puede ser una desventaja en aplicaciones donde el tiempo es crítico. Otros modelos como Regresión Logística, *Random Forest* y *XGBoost* tuvieron tiempos de entrenamiento razonables, todos inferiores a 3.004 segundos, haciéndolos adecuados para situaciones que requieren un buen equilibrio entre rendimiento y tiempo de procesamiento.

4.3.2. Discusión del Objetivo 3

El objetivo 3 se centró en determinar el modelo con el mejor rendimiento para predecir la desnutrición infantil mediante el análisis de un conjunto de datos de

validación. Se evaluaron varios modelos de clasificación, incluyendo Regresión Logística, Support Vector Machine (SVM), *Random Forest*, *K-Nearest Neighbors* (KNN), Árbol de Clasificación y *XGBoost*, utilizando métricas como precisión, sensibilidad, especificidad, F1-Score y el área bajo la curva ROC (AUC-ROC).

Los resultados indican que la mayoría de los modelos alcanzaron una precisión del 90 %, con la excepción del Árbol de Clasificación, que obtuvo un 89 %. *XGBoost* se destacó con una precisión del 92 %, sugiriendo una mejor capacidad para manejar la variabilidad en los datos. En términos de sensibilidad y especificidad, *Random Forest* y *XGBoost* mostraron una ligera ventaja con un 92 %, superando al resto de los modelos.

El F1-Score de *XGBoost* también fue superior, con un 92 %, en comparación con el 91 % de los demás modelos, excepto el Árbol de Clasificación que obtuvo un 89 %. En cuanto al tiempo de entrenamiento, KNN fue el más rápido con solo 0.084 segundos, mientras que SVM fue el más lento con 67.003 segundos.

XGBoost se posicionó como el modelo más robusto y eficiente, recomendado para problemas de clasificación complejos y grandes conjuntos de datos. *Random Forest* también se mostró como una opción viable, equilibrando bien precisión y tiempo de entrenamiento. En contraste, el Árbol de Clasificación presentó limitaciones en precisión, haciéndolo menos adecuado. En resumen, *XGBoost* combinó alta precisión, sensibilidad, especificidad, F1-Score y un tiempo de entrenamiento aceptable, siendo la mejor opción para la predicción de la desnutrición infantil.

CAPÍTULO V

CONCLUSIONES Y RECOMENDACIONES

5.1 CONCLUSIONES

- El tratamiento de datos, que incluyó la limpieza e imputación, fue útil para mejorar la calidad y precisión de los modelos predictivos. La eliminación de valores perdidos y la recodificación de variables categóricas permitieron que los modelos trabajaran con información completa y fiable. El análisis de correlación y la información obtenida por parte de entrevistas realizadas a los especialistas con conocimiento sobre desnutrición infantil indicó que variables como el peso y la talla, así como el IMC, son predictores significativos del estado nutricional de los niños. La fuerte relación entre estas variables y la desnutrición refuerza su inclusión en cualquier modelo predictivo desarrollado para este propósito.

- Al aplicar algoritmos de aprendizaje automático existentes para analizar datos relacionados con la desnutrición infantil en el distrito zona 5 de Ecuador, se logró estimar el grado de precisión en las predicciones a través del entrenamiento y la validación de los resultados. Los modelos evaluados incluyeron Regresión Logística, Random Forest, *K-Nearest Neighbors* (KNN), Árbol de Clasificación y *XGBoost*. Cada uno de estos modelos mostró capacidades únicas para manejar los datos.

- Los modelos de aprendizaje automático han demostrado ser herramientas efectivas para predecir la desnutrición infantil. Se pudo concluir que, entre los evaluados, *XGBoost* se destacó por su alta precisión y capacidad predictiva, siendo el más adecuado para su implementación en el área de salud debido a su manejo eficiente de datos desbalanceados y su robustez ante el sobreajuste.

Otros modelos, como la regresión logística, Random Forest, SVM, KNN y los árboles de decisión, también mostraron efectividad, cada uno con ventajas específicas, como la rapidez de KNN y la robustez de Random Forest. Sin embargo, estos modelos presentaron ciertas limitaciones en comparación con *XGBoost*. En general, los modelos basados en árboles, como *Random Forest* y *XGBoost*, ofrecieron una combinación óptima de precisión y capacidad de generalización. Estos hallazgos proporcionan una base para futuras investigaciones y aplicaciones prácticas en el campo de la salud pública, facilitando intervenciones mejoradas en la prevención y tratamiento de la desnutrición infantil.

5.2 RECOMENDACIONES

- Se recomienda incluir variables adicionales como antecedentes familiares y nivel socioeconómico para mejorar la precisión del modelo predictivo de desnutrición infantil, además de realizar validaciones cruzadas periódicas y actualizaciones de datos para mantener su fiabilidad. Es crucial seguir colaborando con especialistas en nutrición infantil para ajustar el modelo según los avances más recientes y establecer mecanismos prácticos para su implementación y seguimiento en programas de salud pública.

- Se recomienda seleccionar el modelo con mejor rendimiento basado en las métricas de precisión y otras métricas clave obtenidas durante la validación como el tiempo. Además, es esencial realizar ajustes finos a los hiperparámetros de los modelos más prometedores y evaluar su desempeño en diferentes subconjuntos de datos para garantizar su robustez y generalización en el contexto específico de la desnutrición infantil en el distrito zona 5 de Ecuador.
- Se recomienda implementar XGBoost como el modelo principal para la predicción de desnutrición infantil en el área de salud, aprovechando su alta precisión y robustez. Además, se sugiere considerar la combinación de modelos a través de técnicas de ensamblado (ensemble) como el stacking o boosting para mejorar aún más la capacidad predictiva, lo que podría ser un enfoque valioso en trabajos futuros. Es crucial seguir monitoreando el desempeño del modelo en la práctica y ajustando sus parámetros según sea necesario para asegurar su eficacia en diversos escenarios y conjuntos de datos.

CAPÍTULO VI

BIBLIOGRAFÍA

Alejandria, J. J. N. (2021). ANÁLISIS PREDICTIVO EN EL CONTROL DE NUTRICIÓN DE LOS NIÑOS MENORES DE 5 AÑOS DEL PUESTO DE SALUD DE AGOCUCHO, DE CAJAMARCA 2020. 88.

Antunes, H. S. (2019). Inteligência artificial e responsabilidade civil: Enquadramento. Revista de Direito da Responsabilidade, 1.

Argota-Pérez, G., Argota-Pérez, Y., Álvarez-Becerra, R. M., & Reyes-Diaz, M. G. (2022). DECISIÓN FORMATIVA COMO ELEMENTO DE LA INVESTIGACIÓN CIENTÍFICA DESDE LA CONCEPTUALIZACIÓN ESTADÍSTICA Y CIENCIA DE DATOS: LO OBVIO, NO TAN OBVIO. Paideia XXI, 12(1), Article 1. https://doi.org/10.31381/paideia.v12i1.4841

Armesto Formoso, D. (2011). Pruebas diagnósticas: Curvas ROC. Revista Electrónica de Biomedicina, 1, 77-82. https://dialnet.unirioja.es/servlet/articulo?codigo=8886478

Asmare, A. A., & Agmas, Y. A. (2022). Determinants of coexistence of stunting, wasting, and underweight among children under five years in the Gambia; evidence from 2019/20 Gambian demographic health survey: Application of multivariate binary logistic regression model. BMC Public Health, 22(1), 1621. https://doi.org/10.1186/s12889-022-14000-3

AstraEd. (2022, julio 16). ¿Sigue siendo científico de datos el trabajo más sexy del siglo XXI? https://blog.astraed.co/sigue-siendo-cientifico-de-datos-el-trabajo-mas-sexy-del-siglo-xxi/

Basu, A. (2023). Quality Comparison of Upcycled Guitars and Standard Guitars using

Fourier Transforms and Phyphox Tools. International Journal of Applied Physics, 10, 1-5. https://doi.org/10.14445/23500301/IJAP-V10I3P101

Beltrán, C., & Barbona, I. (2019). Regresión Logística y Árboles de Clasificación. Un estudio de simulación para la comparación en el caso de grupos balanceados y desbalanceados. http://hdl.handle.net/2133/14285

Benavides, R. (2023). Factores asociados a la desnutrición crónica infantil en menores de 5 años en la región Huancavelica, 2021 (p. 84). https://cybertesis.unmsm.edu.pe/backend/api/core/bitstreams/eb8af7fa-3b00-4a7d-91bf-3335b6d2a146/content

Browne, C., Matteson, D. S., McBride, L., Hu, L., Liu, Y., Sun, Y., Wen, J., & Barrett, C. B. (2021). Multivariate random forest prediction of poverty and malnutrition prevalence. PLOS ONE, 16(9), e0255519. https://doi.org/10.1371/journal.pone.0255519

Cairo, A. (2017). Visualización de datos: Una imagen puede valer más que mil números, pero no siempre más que mil palabras. Profesional de la información, 26(6), Article 6. https://doi.org/10.3145/epi.2017.nov.02

Código de la Niñez y Adolescencia | Ecuador—Guía Oficial de Trámites y Servicios. (2017). https://www.gob.ec/index.php/regulaciones/codigo-ninez-adolescencia

Córdova, I. E. A., Acosta, N. M., Armijos, A. H., & Castro, P. J. (2018). La matriz de consistencia: Una metodología de investigación para desarrollar el estado del arte para emprendimientos artesanales enfocados en las TIC´s. INNOVA Research Journal, 3(8.1), Article 8.1. https://doi.org/10.33890/innova.v3.n8.1.2018.773

Coz, G. M. (2024). “COMPARACIÓN DE LOS MODELOS LOGIT Y PROBIT PARA LA ESTIMACIÓN DE LA POBREZA Y DESNUTRICIÓN INFANTIL EN BASE A LA ENAHO 2018”. 85.

Cueva Moncayo, M. F., Pérez Padilla, C. A., Ramos Argilagos, M., & Guerrero Caicedo, R. (2021). La desnutrición infantil en Ecuador. Una revisión de literatura. Bol. malariol. salud ambient, 556-564. http://iaes.edu.ve/iaespro/ojs/index.php/bmsa/article/view/364

de Onis, M., Onyango, A. W., Borghi, E., Siyam, A., Nishida, C., & Siekmann, J. (2007). Development of a WHO growth reference for school-aged children and adolescents. Bulletin of the World Health Organization, 85(9), 660-667. https://doi.org/10.2471/BLT.07.043497

Degefa, K., Tadesse, A., Ackley, C., Madrid, L., Assefa, N., Breines, M., Sivalogan, K., Maixenchs, M., & Blevins, J. (2022). Using traditional healers to treat child malnutrition: A qualitative study of health-seeking behaviour in eastern Ethiopia. BMC Public Health, 22(1), 873. https://doi.org/10.1186/s12889-022-13323-5

Dries, C., Stefan, V. A., & Tim, V. (2020). Robust and sparse logistic regression | Advances in Data Analysis and Classification. https://link.springer.com/article/10.1007/s11634-023-00572-4

Durán-Pincay, Y. E., Andrade-Santos, R. M., Aveiga-Bartolomé, Y. L., & Molina-Zambrano, D. S. (2022). Análisis Situacional de la desnutrición infantil a nivel de Latinoamérica. MQRInvestigar, 6(3), Article 3. https://doi.org/10.56048/MQR20225.6.3.2022.1205-1225

Easily, L. S. (2024, marzo 31). Outlier Detection and Treatment: A Comprehensive Guide. LEARN STATISTICS EASILY. https://statisticseasily.com/outlier-detection-and-treatment/

El Estado Mundial de la Infancia 2019: Niños, alimentos y nutrición | UNICEF. (2019, octubre 15). https://www.unicef.org/lac/informes/el-estado-mundial-de-la-infancia-2019-ni%C3%B1os-alimentos-y-nutrici%C3%B3n

Elhady, G. W., Ibrahim, S. kamal, Abbas, E. S., Tawfik, A. M., Hussein, S. E., & Salem, M. R. (2023). Barriers to adequate nutrition care for child malnutrition in a low-resource setting: Perspectives of health care providers. Frontiers in Public Health, 11. https://doi.org/10.3389/fpubh.2023.1064837

Ergul Aydin, Z., & Kamisli Ozturk, Z. (2024). Filter-based feature selection methods in the presence of missing data for medical prediction models. Multimedia Tools and Applications, 83(8), 24187-24216. https://doi.org/10.1007/s11042-023-15917-6

Espinosa-Zúñiga, J. J. (2020). Aplicación de algoritmos Random Forest y XGBoost en una base de solicitudes de tarjetas de crédito. Ingeniería, investigación y tecnología, 21(3). https://doi.org/10.22201/fi.25940732e.2020.21.3.022

Espinoza-Estrella, B. (2023). DESNUTRICIÓN CRÓNICA EN NIÑOS MENORES DE CINCO AÑOS EN ECUADOR 2005-2019. INCIDENCIAS DESDE LAS MADRES ADOLESCENTES. Revista Economía, 75(121), Article 121. https://doi.org/10.29166/economa.v75i121.4472

Fenta, H. M., Zewotir, T., & Muluneh, E. K. (2021). A machine learning classifier approach for identifying the determinants of under-five child undernutrition in Ethiopian

administrative zones. BMC Medical Informatics and Decision Making, 21(1), 291. https://doi.org/10.1186/s12911-021-01652-1

Fernández-Martínez, L. C., Sánchez-Ledesma, R., Godoy-Cuba, G., Pérez-Díaz, O., Estevez-Mitjans, Y., Fernández-Martínez, L. C., Sánchez-Ledesma, R., Godoy-Cuba, G., Pérez-Díaz, O., & Estevez-Mitjans, Y. (2022). Factores determinantes en la desnutrición infantil en San Juan y Martínez, 2020. Revista de Ciencias Médicas de Pinar del Río, 26(1). http://scielo.sld.cu/scielo.php?script=sci_abstract&pid=S1561-31942022000100005&lng=es&nrm=iso&tlng=es

Figueroa, D. K. C., & Ruiz, M. E. P. (2023). Desnutrición crónica infantil y sus efectos en el crecimiento y desarrollo. RECIAMUC, 7(2), Article 2. https://doi.org/10.26820/reciamuc/7.(2).abril.2023.677-686

Fisher, S., & C. Rosella, L. (2022). Priorities for successful use of artificial intelligence by public health organizations: A literature review | BMC Public Health | Full Text. BMC Public Health. https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-022-14422-z

Fiuza Pérez, M. D., & Rodríguez Pérez, J. C. (2000). La regresión logística: Una herramienta versátil. Nefrología, 20(6), 495-500. http://www.revistanefrologia.com/es-la-regresion-logistica-una-herramienta-articulo-X0211699500035664

Flores, P., & Congacha, G. (2021). FACTORES ASOCIADOS A LA DESNUTRICIÓN CRÓNICA INFANTIL EN ECUADOR. ESTUDIO BASADO EN MODELOS DE REGRESIÓN Y ÁRBOLES DE CLASIFICACIÓN. Perfiles, 1(26), 21-33. https://doi.org/10.47187/perf.v1i26.132

Galli, S. (2023a, marzo 28). Overcoming Class Imbalance with SMOTE: How to Tackle Imbalanced Datasets in Machine Learning. Train in Data's Blog. https://www.blog.trainindata.com/overcoming-class-imbalance-with-smote/

Galli, S. (2023b, marzo 28). Overcoming Class Imbalance with SMOTE: How to Tackle Imbalanced Datasets in Machine Learning. Train in Data's Blog. https://www.blog.trainindata.com/overcoming-class-imbalance-with-smote/

Giraldo Mejía, J. C., & Vargas Agudelo, F. A. (2011). Aplicación de la Técnica Regresión Logística de la minería de datos en el proceso de Descubrimiento de Conocimiento (KDD) en bases de datos operativas o transaccionales. Universidad Inca Garcilaso de la Vega. http://repositorio.uigv.edu.pe/handle/20.500.11818/901

Granados Mota, D. J. (2023). MODELO DE REGRESIÓN MULTIVARIADO APLICADO A LAS CONDICIONES DE DESNUTRICIÓN INFANTIL, PARA MEJORAR LAS PROPUESTAS DE DESARROLLO SOCIAL EN EL SUROCCIDENTE DE GUATEMALA [Masters, Universidad de San Carlos de Guatemala]. https://postgrado.ingenieria.usac.edu.gt/

Hasan, Md. M., Popp, J., & Oláh, J. (2020). Current landscape and influence of big data on finance. Journal of Big Data, 7(1), 21. https://doi.org/10.1186/s40537-020-00291-z

Hassannataj Joloudari, J., Marefat, A., Ali Nematollahi, M., Sunday Oyelere, S., & Hussain, S. (2023). Ciencias aplicadas Aprendizaje eficaz de desequilibrio de clases basado en SMOTE y redes neuronales convolucionales. https://www.mdpi.com/2076-3417/13/6/4006

Hernández, F., & Usuga, O. (2024). Manual de R. https://fhernanb.github.io/Manual-de-

R/

Huang, S., Cai, N., Pacheco, P. P., Narrandes, S., Wang, Y., & Xu, W. (2018). Applications of Support Vector Machine (SVM) Learning in Cancer Genomics. Cancer Genomics & Proteomics, 15(1), 41-51. https://doi.org/10.21873/cgp.20063

Huiracocha, L., Robalino, G., Huiracocha, M., García, J., Pazán, C., & Angulo, A. (2012). Retrasos del desarrollo psicomotriz en niños y niñas urbanos de 0 a 5 años: Estudio de caso en la zona urbana de Cuenca, Ecuador. MASKANA, 3(1), 13-28. https://doi.org/10.18537/mskn.03.01.02

IBM. (2023, noviembre 7). ¿Qué es KNN? | IBM. https://www.ibm.com/mx-es/topics/knn

Instituto Nacional de Estadística y Censos, I. N. de E. y C. (2023, septiembre 5). PRIMERA ENCUESTA ESPECIALIZADA REVELA QUE EL 20.1% DE LOS NIÑOS EN ECUADOR PADECEN DE DESNUTRICIÓN CRÓNICA INFANTIL – Instituto Nacional de Estadística y Censos. https://www.ecuadorencifras.gob.ec/institucional/primera-encuesta-especializada-revela-que-el-20-1-de-los-ninos-en-ecuador-padecen-de-desnutricion-cronica-infantil/

Jiménez, W. A. A. (2012). La infancia como sujeto de derechos según UNICEF. Aportes para una lectura crítica y de extrañamiento. Pedagogía y Saberes, 37, 89-101. https://www.redalyc.org/articulo.oa?id=614064827008

Kalaivani, B., & Ranichitra, A. (2024). Unveiling the Impact of Outliers: An Improved Feature Engineering Technique for Heart Disease Prediction. En P. P. Joby, M. S.

Alencar, & P. Falkowski-Gilski (Eds.), IoT Based Control Networks and Intelligent Systems (pp. 469-478). Springer Nature. https://doi.org/10.1007/978-981-99-6586-1_32

Kim, J., Mun, S., Lee, S., Jeong, K., & Baek, Y. (2022). Prediction of metabolic and pre-metabolic syndromes using machine learning models with anthropometric, lifestyle, and biochemical factors from a middle-aged population in Korea. BMC Public Health, 22(1), 664. https://doi.org/10.1186/s12889-022-13131-x

Ley Orgánica de Protección de Datos Personales (379637). (2021). Asamblea Nacional del Ecuador. https://www.asambleanacional.gob.ec/es/multimedios-legislativos/63464-ley-organica-de-proteccion-de-datos

Llego, M. A. (2023, octubre 11). Detecting and Handling Outliers in Python: Methods and Strategies - llego.dev. https://llego.dev/posts/outlier-detection-handling-python-guide/

M. Locks, L., Pooja R, P., Osei Akoto K., David S, S., Debendra P., A., Nancy J., H., Victoria J., Q., & Jennifer N., N. (2015). Using formative research to design a context-specific behaviour change strategy to improve infant and young child feeding practices and nutrition in Nepal. 11, 882-896. https://doi.org/10.1111/mcn.12032

Manosalvas, M. (2019). La política del efectivismo y la desnutrición infantil en el Ecuador. Perfiles latinoamericanos, 27(54). https://doi.org/10.18504/pl2754-013-2019

Marcos Valdez, A. J., Navarro Ortiz, E. G., Quinteros Peralta, R. E., Tirado Julca, J. J., Valentin Ricaldi, D. F., & Calderon Vilca, H. D. (2023). Machine Learning for the Prediction of Anemia in Children Under 5 Years of Age by Analyzing their Nutritional Status Using Data Mining. Computación y Sistemas, 27(3). https://doi.org/10.13053/cys-

27-3-4315

Mark, H. E., Costa, G. D. da, Pagliari, C., & Unger, S. A. (2020). Malnutrition: The silent pandemic. BMJ, 371, m4593. https://doi.org/10.1136/bmj.m4593

Mero, C., & Dario, J. (2021). Factores socioeconómicos, educativos y su impacto en la desnutrición en niños/as de dos Centros Infantiles del Cantón Francisco Orellana, 2021. https://repositorio.ucv.edu.pe/handle/20.500.12692/72919

Morgenstern, J. D., Buajitti, E., O'Neill, M., Piggott, T., Goel, V., Fridman, D., Kornas, K., & Rosella, L. C. (2020). Predicting population health with machine learning: A scoping review. BMJ Open, 10(10), e037860. https://doi.org/10.1136/bmjopen-2020-037860

Munawar, H. S., Ullah, F., Qayyum, S., & Shahzad, D. (2022). Big Data in Construction: Current Applications and Future Opportunities. Big Data and Cognitive Computing, 6(1), Article 1. https://doi.org/10.3390/bdcc6010018

Naidu, G., Zuva, T., & Sibanda, E. M. (2023). A Review of Evaluation Metrics in Machine Learning Algorithms. En R. Silhavy & P. Silhavy (Eds.), Artificial Intelligence Application in Networks and Systems (pp. 15-25). Springer International Publishing. https://doi.org/10.1007/978-3-031-35314-7_2

Nieto, Y., García-Díaz, V., Montenegro, C., & Crespo, R. G. (2019). Supporting academic decision making at higher educational institutions using machine learning-based algorithms. Soft Computing, 23(12), 4145-4153. https://doi.org/10.1007/s00500-018-3064-6

Ortega, L. G. A. (2019). Desnutrición infantil, una mirada desde diversos factores.

Investigación Valdizana, 13(1), Article 1. https://doi.org/10.33554/riv.13.1.168

Pérez, R., & Luis, J. (2014). Técnicas de aprendizaje automático para la detección de intrusos en redes de computadoras. Revista Cubana de Ciencias Informáticas, 8(4), 52-73. http://scielo.sld.cu/scielo.php?script=sci_abstract&pid=S2227-18992014000400003&lng=es&nrm=iso&tlng=es

Prol Castelo, G., Urda, B., Núñez Carpintero, I., Cirillo, D., & Valencia, A. (2022). La inteligencia artificial en biomedicina: Oportunidades y desafíos. AmbioCiencias: revista de divulgación, 20, 7-21. https://dialnet.unirioja.es/servlet/articulo?codigo=8788516

Rivera, J. (2019). La malnutrición infantil en Ecuador: Una mirada desde las políticas públicas. Revista Estudios de Políticas Públicas, 5(1), 89-107. https://dialnet.unirioja.es/servlet/articulo?codigo=7390665

Rosa, J. M., & Frutos, E. L. (2022). Ciencia de datos en salud: Desafíos y oportunidades en América Latina. Revista Médica Clínica Las Condes, 33(6), 591-597. https://doi.org/10.1016/j.rmclc.2022.09.007

Sachin. (2024, junio 2). What Is The 80/20 Rule In Machine Learning? https://www.techmediatoday.com/what-is-the-80-20-rule-in-machine-learning/

Silva Hernández, F., & Martínez Prats, G. (2021). Estrategias de innovación docente mediante las TIC. 3C TIC: Cuadernos de desarrollo aplicados a las TIC, 10(4), 89-103. https://doi.org/10.17993/3ctic.2021.104.89-103

Silva, T. P., Carvalho, M. de N., & Takeshita, W. M. (2021). Estado da arte da Inteligência Artifical (IA) na radiologia odontológica: Revisão sistemática. ARCHIVES OF

HEALTH INVESTIGATION, 10(7), Article 7. https://doi.org/10.21270/archi.v10i7.5069

Starbuck, C. (2023). Logistic Regression. En C. Starbuck (Ed.), The Fundamentals of People Analytics: With Applications in R (pp. 223-238). Springer International Publishing. https://doi.org/10.1007/978-3-031-28674-2_12

Suryawan, A., Jalaludin, M. Y., Poh, B. K., Sanusi, R., Tan, V. M. H., Geurts, J. M., & Muhardi, L. (2022). Malnutrition in early life and its neurodevelopmental and cognitive consequences: A scoping review. Nutrition Research Reviews, 35(1), 136-149. https://doi.org/10.1017/S0954422421000159

The State of Food Security and Nutrition in the World 2020. (2020). FAO, IFAD, UNICEF, WFP and WHO. https://doi.org/10.4060/ca9692en

The State of Food Security and Nutrition in the World 2021. (2021). FAO, IFAD, UNICEF, WFP and WHO. https://doi.org/10.4060/cb4474en

UNICEF. (2016). Para cada niño el mejor comienzo Primera infancia 2010- 2016.pdf (p. 24). https://www.unicef.org/argentina/media/626/file/Primera%20infancia%202010-%202016.pdf

UNICEF. (2021). Desnutrición Crónica Infantil | UNICEF. UNICEF PARA CADA INFANCIA. https://www.unicef.org/ecuador/desnutrici%C3%B3n-cr%C3%B3nica-infantil

Anexos 1. Certificado de anti-plagio

	UNIVERSIDAD TECNICA ESTATAL QUEVEDO FACULTAD DE POSGRADO	

PARA Dr. Byron Oviedo Bayas
Decano de Posgrado
DE Dr. Orlando Erazo Moreta
ASUNTO Informe Proyecto de Investigación
FECHA 3 de agosto del 2024

Adjunto al presente sírvase encontrar el documento final del proyecto de investigación titulado: MODELO PREDICTIVO DE DESNUTRICION INFANTIL DEL ECUADOR DISTRITO ZONA 5, elaborado por el ING. ANGÉLICA NOEMÍ CARRIÓN GONZÁLEZ posgradista de la MAESTRÍA EN CIENCIA DE DATOS. El proyecto de investigación fue elaborado bajo mi dirección según lo asignado en el contrato Nro. UTEQ-RUTEQ- 2023-4194-M de fecha 7 de mayo de 2024, el mismo que cumple el informe de la herramienta COMPILATIO, el cual avala los niveles de originalidad, en un 97 % del trabajo investigativo.


CERTIFICADO DE ANÁLISIS
magister

ACarrion_Proyecto de investigación Modelo predictivo de desnutricion plagio

3% Textos sospechosos

2% Similitudes
< 1% similitudes entre comillas
0% entre las fuentes mencionadas
< 1% Idiomas no reconocidos

Nombre del documento: ACarrion_Proyecto de investigación Modelo predictivo de desnutricion plagio.pdf
ID del documento: 9de14bbf0997965d4f74dc9d1438d47eda5f0672
Tamaño del documento original: 1,8 MB

Depositante: ORLANDO RAMIRO ERAZO MORETA
Fecha de depósito: 3/8/2024
Tipo de carga: interface
fecha de fin de análisis: 3/8/2024

Número de palabras: 19.911
Número de caracteres: 144.875

Ubicación de las similitudes en el documento:

Atentamente,

Firmado electrónicamente por:
ORLANDO RAMIRO ERAZO MORETA

Dr. Orlando Erazo Moreta
Director de Proyecto de Investigación

Anexo 2. Solicitud de base de datos al ministerio de salud publica distrito zona 5.

UTEQ

Universidad Técnica Estatal de Quevedo
La primera universidad agropecuaria del Ecuador

MINISTERIO DE SALUD PÚBLICA
DIRECCIÓN DISTRITAL 12D03
RECIBIDO
Fecha: 23/02/24 Hora: 13:25
Nombre: CRISTHIAN ALCIVAR
Anexos:
MSP-CZ5S-LR-12D03-QM-20.....E

Oficio Nro. UTEQ-POSG-2024-0004

Quevedo, 23 de febrero de 2024

Asunto: Colaboración para un proyecto de investigación

Señores
Dr. Fleshman Jiménez
Director Distrital
CENTROS DE SALUD Y DIRECCIÓN DISTRITAL 12D03 DEL MINISTERIO DE SALUD PÚBLICA DEL ECUADOR
En su Despacho

De mi consideración:

En mi calidad de Coordinador del Programa de Maestría en Ciencia de Datos en la Facultad de Posgrado de la Universidad Técnica Estatal de Quevedo, me dirijo a ustedcon el fin de solicitar su colaboración para un proyecto de investigación de vital importancia.

La estudiante *Angélica Noemí Carrión González*, identificada con el número de cédula 0940902788, se encuentra actualmente desarrollando su proyecto de investigación titulado "*Modelo Predictivo de Desnutrición Infantil en el Ecuador*", bajo la dirección del *Ing. Orlando Ramiro Erazo Moreta, PhD.*

El propósito fundamental de este proyecto es generar un modelo predictivo que utilice técnicas de análisis de datos y aprendizaje automático para anticipar la desnutrición infantil en niños ecuatorianos, identificando los factores de riesgo asociados. Para avanzar en este proyecto, solicitamos amablemente la colaboración del Ministerio de Salud Pública del Ecuador.

Para el estudio se requiere una base de datos, dentro de la cual contengan variables relevantes para el estudio como: Nombres, edad, peso, talla, percentiles pediátricos referente a peso y talla, de preferencia de los últimos dos o tres años y en formato de Excel (ficheros con extensión xls).

Junto a la participación del *Dr. Jimmy Barros Segovia*, Director Médico en Centro Healthy, dará relevancia a la realización del proyecto de investigación por su contribución invaluable que puede ofrecer al proyecto, proporcionando información esencial para la toma de decisiones dirigidas a la prevención de la desnutrición infantil en Ecuador. La desnutrición, al privar a los niños de los nutrientes necesarios para su adecuado crecimiento y desarrollo, afectanegativamente su calidad de vida presente y futura, representando una amenaza significativa para su desarrollo físico y cognitivo. Al identificar áreas o grupos de población con mayor riesgo de desnutrición, podemos asignar recursos de manera más eficiente, enfocándolos en programas de nutrición específicos y asegurando que lleguen a quienes más lo necesitan.

Por lo tanto, solicitamos su intermediación y respaldo para facilitar la formalización de esta solicitud y los procedimientos necesarios para establecer una colaboración con su institución. Su experiencia y liderazgo son cruciales para garantizar que nuestra solicitud sea debidamente considerada.

Además, me gustaría destacar que esta colaboración no solo beneficiará el proyecto de investigación en curso, sino que también enriquecerá el programa de posgrado al proporcionar acceso a datos y la experiencia de expertos en el campo de la desnutrición infantil. Estamos dispuestos a trabajar en estrecha colaboración para construir los modelos predictivos necesarios y alcanzar nuestros objetivos de investigación.

Agradecemos de antemano su atención y disposición para apoyarnos en este importante asunto. Quedo a su entera disposición para discutir cualquier detalle adicional y seguir los procedimientos necesarios.

Con sentimientos de distinguida consideración.

Atentamente,

ANGEL IVAN TORRES QUIJIJE

Ing. Angel Iván Torres Quijije
COORDINADOR MAESTRÍAS: CIENCIA DE DATOS, AUTOMATIZACIÓN Y CONTROL INDUSTRIAL

Anexo 3 Preguntas de entrevista a profesionales en el área de salud sobre la desnutrición infantil

Nombre:

Especialidad:

Institución laboral:

1. ¿Qué es la desnutrición infantil para Usted?

---

2. ¿Hasta qué edad se puede hablar de desnutrición Infantil?

---

3. ¿Cuáles son los desafíos más comunes que se enfrenta en la detección de la desnutrición infantil en su entorno clínico?

---

4. ¿Qué tipo de datos considera más útiles para predecir o identificar posibles casos de desnutrición infantil?

---

5. ¿Cuáles son los indicadores clínicos que considera más relevantes para conocer la existencia de casos de desnutrición infantil?

---

6. ¿Existen normas o estándares a nivel local, nacional o internacional que regulen el manejo de la desnutrición infantil en su institución? ¿Cuáles son?

---

7. ¿Qué variables o factores considera más importantes incluir en un modelo de predicción de desnutrición infantil?

Firma