Universidad Nacional Mayor de San Marcos
Universidad del Perú. Decana de América
Facultad de Ciencias Matemáticas
Escuela Profesional de Estadística

Random Forest como método alternativo para determinar factores asociados a la anemia en niños y niñas entre 6 a 59 meses en el Perú, según la Encuesta Demográfica y Salud Familiar – ENDES 2022

TRABAJO DE SUFICIENCIA PROFESIONAL

Para optar el Título Profesional de Licenciada en Estadística

AUTOR

Quinia Karina ARIZA RAMIREZ

ASESOR

Dra. Ofelia ROQUE PAREDES

Lima, Perú

2023

Referencia bibliográfica

Ariza, Q. (2023). *Random Forest como método alternativo para determinar factores asociados a la anemia en niños y niñas entre 6 a 59 meses en el Perú, según la Encuesta Demográfica y Salud Familiar – ENDES 2022*. [Trabajo de suficiencia profesional de pregrado, Universidad Nacional Mayor de San Marcos, Facultad de Ciencias Matemáticas, Escuela Profesional de Estadística]. Repositorio institucional Cybertesis UNMSM.

Datos de autor	
Nombres y apellidos	Quinia Karina Ariza Ramirez
Tipo de documento de identidad	DNI
Número de documento de identidad	71592083
URL de ORCID	https://orcid.org/0009-0007-6435-0922
Datos de asesor	
Nombres y apellidos	Ofelia Roque Paredes
Tipo de documento de identidad	DNI
Número de documento de identidad	06243124
URL de ORCID	https://orcid.org/0000 -0001-8280-021X
Datos del jurado	
Presidente del jurado	
Nombres y apellidos	Zoraida Judith Huamán Gutiérrez
Tipo de documento	DNI
Número de documento de identidad	09890094
Miembro del jurado 1	
Nombres y apellidos	Hugo Marino Rodríguez Orellana
Tipo de documento	DNI
Número de documento de identidad	40162362
Datos de investigación	
Línea de investigación	Análisis de Datos y Modelamiento de Problemas de la Sociedad

Grupo de investigación	No aplica.
Agencia de financiamiento	Sin financiamiento.
Ubicación geográfica de la investigación	Universidad Nacional Mayor de San Marcos País: Perú Departamento: Lima Provincia: Lima Distrito: Lima Coordenadas geográficas Latitud: -12.058333 Longitud: -77.083333
Año o rango de años en que se realizó la investigación	Mayo 2022 – Setiembre 2022
URL de disciplinas OCDE	Estadísticas, Probabilidad https://purl.org/pe-repo/ocde/ford#1.01.03

UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS
Universidad del Perú. Decana de América
FACULTAD DE CIENCIAS MATEMÁTICAS
ESCUELA PROFESIONAL DE ESTADÍSTICA

ACTA DE SUSTENTACIÓN DEL TRABAJO DE SUFICIENCIA PROFESIONAL PARA LA OBTENCIÓN DEL TÍTULO PROFESIONAL DE LICENCIADA EN ESTADÍSTICA
(PROGRAMA DE TITULACIÓN PROFESIONAL 2023)

En la UNMSM – Ciudad Universitaria – Facultad de Ciencias Matemáticas, siendo las 15:30 horas del sábado 21 de octubre del 2023, se reunieron los docentes designados como Miembros del Jurado Evaluador (PROGRAMA DE TITULACIÓN PROFESIONAL 2023): Dra. Zoraida Judith Huamán Gutiérrez (PRESIDENTE), Mg. Hugo Marino Rodríguez Orellana (MIEMBRO) y la Dra. Ofelia Roque Paredes (MIEMBRO ASESOR), para la sustentación del Trabajo de Suficiencia Profesional titulado: "**RANDOM FOREST COMO MÉTODO ALTERNATIVO PARA DETERMINAR FACTORES ASOCIADOS A LA ANEMIA EN NIÑOS Y NIÑAS ENTRE 6 A 59 MESES EN EL PERÚ, SEGÚN LA ENCUESTA DEMOGRÁFICA Y SALUD FAMILIAR – ENDES 2022**", presentado por la señorita **Bachiller QUINIA KARINA ARIZA RAMIREZ**, para optar el Título Profesional de Licenciada en Estadística.

Luego de la exposición del Trabajo de Suficiencia Profesional, la Presidente invitó al expositor a dar respuesta a las preguntas formuladas.

Realizada la evaluación correspondiente por los Miembros del Jurado Evaluador, la expositora mereció la aprobación .....Sobresaliente...., con un calificativo promedio de Diecisiete (17)

A continuación, los Miembros del Jurado Evaluador dan manifiesto que la participante **Bachiller QUINIA KARINA ARIZA RAMIREZ**, en vista de haber aprobado la sustentación de su Trabajo de Suficiencia Profesional, será propuesta para que se le otorgue el Título Profesional de Licenciada en Estadística.

Siendo las 16... horas se levantó la sesión firmando para constancia la presente Acta.

Dra. Zoraida Judith Huamán Gutiérrez
PRESIDENTE

Mg. Hugo Marino Rodríguez Orellana
MIEMBRO

Dra. Ofelia Roque Paredes
MIEMBRO ASESOR

CERTIFICADO DE SIMILITUD

Yo, Ofelia Roque Paredes en mi condición de asesora acreditada con Resolución Decanal N° 001610-2023-D-FCM/UNMSM del Trabajo de Suficiencia Profesional, cuyo título es "RANDOM FOREST COMO MÉTODO ALTERNATIVO PARA DETERMINAR FACTORES ASOCIADOS A LA ANEMIA EN NIÑOS Y NIÑAS ENTRE 6 A 59 MESES EN EL PERÚ, SEGÚN LA ENCUESTA DEMOGRÁFICA Y SALUD FAMILIAR – ENDES 2022", presentado por la bachiller QUINIA KARINA ARIZA RAMIREZ, para optar el título de Licenciada en Estadística.

Certifico que se ha cumplido con lo establecido en la Directiva de Originalidad y de Similitud de Trabajos Académicos, de Investigación y Producción Intelectual. Según la revisión, análisis y evaluación mediante el software de similitud textual, el documento evaluado cuenta con el porcentaje de **12%** de similitud, nivel **PERMITIDO** para continuar con los trámites correspondientes y para su **publicación en el repositorio institucional**.

Se emite el presente certificado en cumplimiento de lo establecido en las normas vigentes, como uno de los requisitos para la obtención del título correspondiente.

DNI: 06243124

Dra. Ofelia Roque Paredes

Huella Digital

RESUMEN

A nivel mundial, en el Perú, la anemia es un problema de salud pública que afecta a niños y niñas, adolescentes, mujeres embarazadas y adultos mayores. Presenta diversas causas; sin embargo, la más frecuente es por la deficiencia en el consumo de hierro, y las consecuencias se observan a corto y largo plazo. Existe muchos estudios sobre el déficit en la atención, crecimiento retardado, disminución en la respuesta inmunológica y un bajo cognitivo, como consecuencias de la anemia, en niños y niñas de 6 a 59 meses.

La anemia es afectada por muchos factores, y este estudio tiene como objetivo mostrar un método alternativo para establecer los factores más importantes asociados a la anemia en niños y niñas de 6 a 59 meses, usando la Encuesta Demográfica y de Salud Familiar en el Perú del año 2022.

En el desarrollo del presente trabajo de investigación, se encontraron resultados sobre los factores de la anemia usando el método machine learning “Random Forest”, y se identificó que los factores sociodemográficos como el índice de riqueza, el parentesco con el jefe del hogar, el nivel de estudios del apoderado y el lugar de residencia están asociados con la anemia en niños y niñas entre 6 a 59 meses en el Perú.

El método machine learning “Random Forest” es una alternativa eficiente para encontrar resultados sobre factores que presentan relación con la anemia en niños y niñas de 6 a 59 meses, usando la Encuesta Demográfica y de Salud Familiar, ENDES 2022. En medida que los datos de la ENDES serán evaluados con un método no convencional, se considera que se trata de un aporte que proporciona nuevas alternativas de análisis.

ABSTRACT

Worldwide, anemia is a public health problem that affects children, adolescents, pregnant women and older adults in Peru. It has several causes; however, the most frequent is due to iron deficiency, and the consequences are observed in the short and long term. There are many studies on attention deficit, delayed growth, decreased immune response and cognitive decline as consequences of anemia in children aged 6 to 59 months.

Anemia is affected by many factors, and this study aims to show an alternative method to establish the most important factors associated with anemia in children aged 6 to 59 months, using the Demographic and Family Health Survey in Peru in 2022.

In the development of this research work, results were found on the factors of anemia using the "Random Forest" machine learning method, and it was identified that sociodemographic factors such as wealth index, relationship with the head of household, level of education of the proxy and place of residence are associated with anemia in children between 6 to 59 months in Peru.

The machine learning method "Random Forest" is an efficient alternative to find results on factors related to anemia in children aged 6 to 59 months, using the Demographic and Family Health Survey, ENDES 2022. Since the ENDES data will be evaluated with a non-conventional method, it is considered a contribution that provides new analysis alternatives.

ÍNDICE

I. INTRODUCCIÓN 9

II. DESCRIPCIÓN DE LA ACTIVIDAD 10

2.1. Instituto Nacional de Estadística 10

2.1.1. Misión 11

2.1.2. Visión 11

2.2. Programa Nacional de Alimentación Escolar Qali Warma 11

2.2.1. Misión 11

2.3. Organigrama del Ministerio de Desarrollo e Inclusión Social 12

2.4. Organigrama del Programa Nacional de Alimentación Escolar Qali Warma 13

2.5. Problemática 14

2.5.1. Causas que provocan la anemia 16

2.5.2. Consecuencias de la anemia 16

2.5.3. Cómo determinar si un niño padece anemia 17

2.5.4. Clasificación de la anemia según el INEI 17

2.5.5. Alcances de la investigación usando el modelo Random Forest 18

2.5.6. Formulación del problema 19

2.5.7. Formulación de los problemas específicos 19

2.5.8. Objetivo principal 19

2.5.9. Objetivos específicos 19

III. MARCO TEÓRICO 21

3.1. Bases teóricas 21

3.1.1. Árboles de decisión 21
3.1.2. Random Forest 21
3.1.3. Formación del algoritmo del modelo “Random Forest” 23
3.1.4. Ventajas 24
3.1.5. Desventajas 25
3.1.6. Medidas de las variables de importancia 25
3.1.7. Comparación de modelos - Curva ROC 27
3.1.8. Tabla de clasificación 27
3.1.9. Curva ROC 29
3.1.10. Anemia 29
3.1.11. Antecedentes internacionales 31
3.1.12. Antecedentes nacionales 34
IV. METODOLOGÍA 35
4.1. Diseño de investigación y tipo de estudio 35
4.2. Población y muestra 35
4.3. Método para recolección de datos 36
4.4. Instrumento de medición 36
4.5. Variables 37
4.6. Procesamiento de la base de datos 38
V. RESULTADOS 40
5.1. Análisis descriptivo de factores asociados a la anemia en el Perú 40
5.2. Porcentaje de sensibilidad y especificidad con el modelo Random Forest 43
5.3. Importancia de variables 44

5.4. Porcentaje de sensibilidad y especificidad con el modelo Random Forest por área 45

VI. CONCLUSIONES 47

VII. RECOMENDACIONES 48

VIII. REFERENCIAS 49

ÍNDICE DE FIGURAS

Figura 1 *Organigrama del MIDIS* 12

Figura 2 *Organigrama del PNAEQW* 13

Figura 3 *Porcentaje dominante de anemia en niños y niñas menores de 5 años, 2019* 15

Figura 4 *Anemia en la población infantil: causas y consecuencias* 17

Figura 5 *Error de clasificación según número de cuenta árboles* 22

Figura 6 *Error de clasificación según número de árboles* 26

Figura 7 *Matriz de confusión* 28

Figura 8 *Curva ROC* 29

Figura 9 *Beneficios del Programa de Alimentación según indicador de anemia en el Perú, 2022* 40

Figura 10 *Beneficios del Programa Nacional de Alimentación según el lugar de residencia, 2022* 41

Figura 11 *Índice de riqueza de los entrevistados en el Perú, 2022* 42

Figura 12 *Área bajo la curva (AUC)* 44

ÍNDICE DE CUADROS

Cuadro 1 *Clasificación de la anemia* 18

Cuadro 2 *Resumen de la muestra* 35

Cuadro 3 *Resumen de variables sociodemográficas para la aplicación del Random Forest* 37

Cuadro 4 *Área de residencia de niños y niñas de 6 a 59 meses en el Perú, 2022* 42

Cuadro 5 *Lugar de residencia de niños y niñas de 6 a 59 meses en el Perú, 2022* 43

Cuadro 6 *Indicador del modelo Random Forest* 43

Cuadro 7 *Importancia de variables en el modelo Random Forest* 44

Cuadro 8 *Importancia de variables por área según en el modelo Random Forest* 45

Cuadro 9 *Importancia de variables en el modelo Random Forest* 45

I. INTRODUCCIÓN

En este trabajo de suficiencia, se aplicará el modelo machine learning “Random Forest” usando como base los datos de la Encuesta Demográfica y de Salud Familiar, ENDES 2022.

El capítulo II describe la reseña de las actividades, el organigrama del Instituto Nacional de Estadística (INEI) y el Programa Nacional de Alimentación Escolar Qali Warma, así como el resumen de la evaluación de la problemática social usando como referencia a la ONU y el INEI.

En el capítulo III se detalla las bases teóricas, así como los antecedentes a nivel nacional e internacional sobre la anemia y las aplicaciones sobre los modelos machine learning “Random Forest”.

El capítulo IV describe el tipo y diseño de investigación, población, muestra, método de recolección de datos, instrumento de medición y las variables que se van a evaluar en la base de datos, así como el detalle del procesamiento de datos.

En el capítulo V y VI, describe los resultados descriptivos y variables de importancia evaluados bajo el modelo de Random Forest obtenidos de las bases de datos de la INEI, de la Encuesta Demográfica y de Salud Familiar.

En los capítulos VII y VIII se describen las conclusiones y recomendaciones obtenidas en el presente trabajo de investigación.

II. DESCRIPCIÓN DE LA ACTIVIDAD

Las actividades se realizaron en el Programa Nacional de Alimentación Escolar Qali Warma, sin embargo, para el análisis y elaboración del trabajo de suficiencia se consultaron los datos del Instituto Nacional de Estadística (INEI). Por consiguiente, se describe una breve reseña del INEI.

2.1. Instituto Nacional de Estadística

El Instituto Nacional de Estadística (INEI), entidad responsable del Sistema Estadístico Nacional, se encarga de brindar y dar seguimiento a los indicadores de resultados y cifras estadísticas.

Una de las actividades del INEI es recopilar información con el Cuestionario Individual de la Encuesta Demográfica y de Salud Familiar (ENDES), para medir los resultados del lineamiento de la política nacional – primera infancia, y así brindar cifras estadísticas para la toma de decisiones.

Durante los últimos años, el INEI brinda información de los indicadores de resultados sobre "salud materna e infantil", "fecundidad", "mortalidad" e "indicadores de resultados para el monitoreo y evaluación de los programas presupuestales (PPR)" en el Perú [indicadores del programa presupuestal articulado nutricional, salud materno neonatal, acceso de la población (6 a 59 meses de edad) a la identidad, violencia a la mujer, etc.].

La recopilación de información por la ENDES inició en el año 1995; sin embargo, desde el 2003 se realiza de forma frecuente a nivel nacional. Cada año la entidad brinda principales

resultados acerca de la calidad de vida de la población, gastos e ingresos en los hogares, y la pobreza monetaria en el país e indicadores del PPR.

2.1.1. Misión

Según el INEI (s.f.) "producir y difundir información estadística oficial […] con el propósito de contribuir al diseño, monitoreo y evaluación de políticas públicas y al proceso de toma de decisiones de los agentes socioeconómicos, el sector público y la comunidad en general".

2.1.2. Visión

"Somos un organismo líder a nivel nacional e internacional, que utiliza los más altos estándares metodológicos y tecnológicos para la producción y difusión de estadísticas oficiales que contribuyan eficazmente en el diseño de políticas públicas para el desarrollo del país" (INEI, s.f.).

2.2. Programa Nacional de Alimentación Escolar Qali Warma

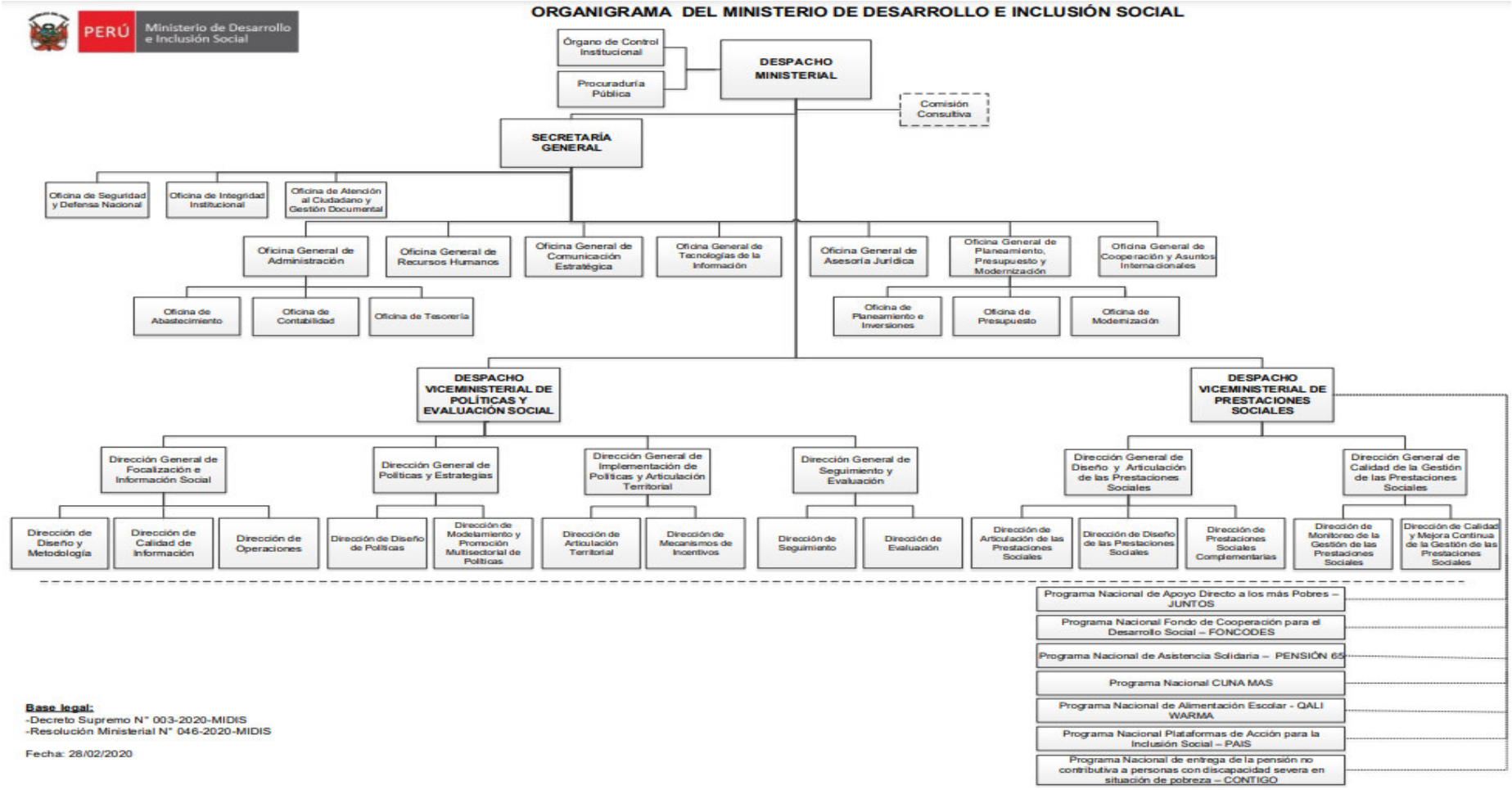
2.2.1. Misión

"Ser un programa eficiente, eficaz y articulado, que fomenta el desarrollo humano a través del servicio alimentario de calidad, en gestión conjunta con la comunidad local". "Programa Nacional de Alimentación Escolar Qali Warma (PNAE - QALI WARMA) es una entidad adscrita al Ministerio de Desarrollo e Inclusión Social" (Gobierno del Perú, s.f.).

2.3. Organigrama del Ministerio de Desarrollo e Inclusión Social

Figura 1

Organigrama del MIDIS



Nota. Información obtenida del organigrama del MIDIS [Aprobado con D.S. N° 003-2020-MIDIS y R.M. N° 046-2020-MIDIS]

2.4. Organigrama del Programa Nacional de Alimentación Escolar Qali Warma

Figura 2

Organigrama del PNAEQW

Nota. Información obtenida del organigrama del PNAEQW [Aprobado Manual de Operaciones 2017, aprobado con RM 283-2017-MIDIS]

2.5. Problemática

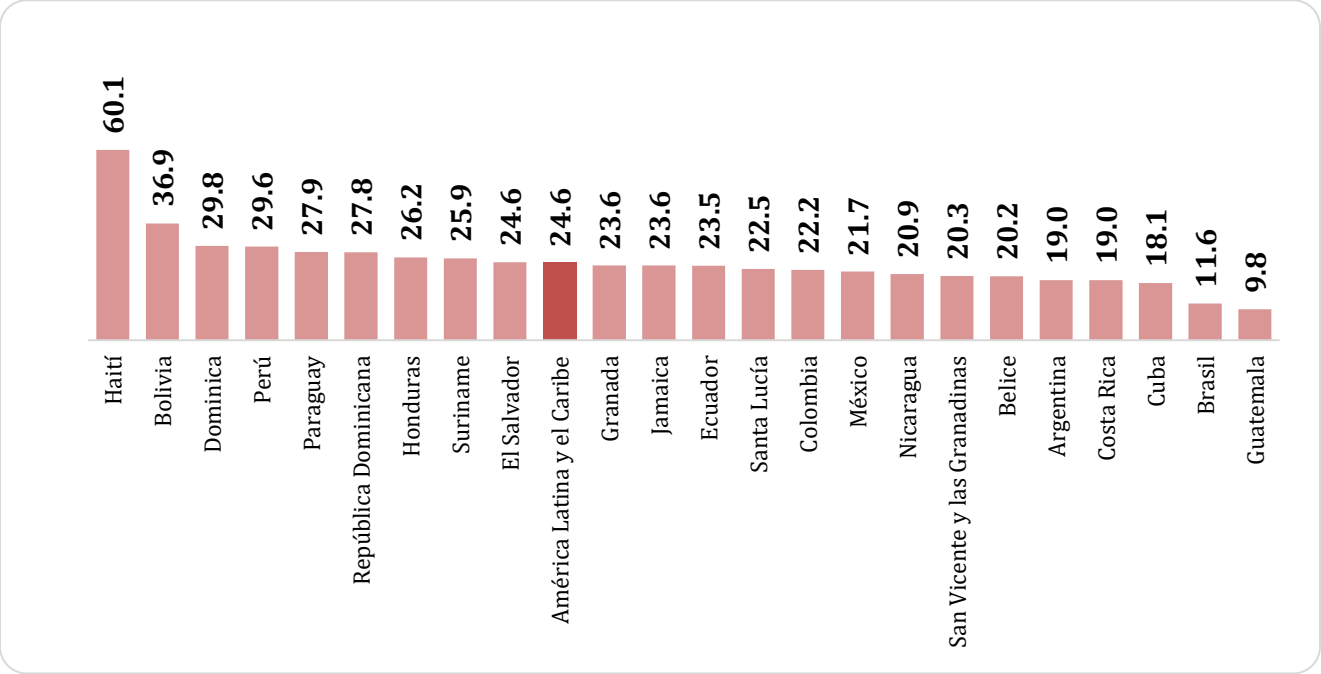
El Gobierno, en los últimos años, viene realizando diferentes medidas para que se mejore la salud de los niños y niñas en relación con la anemia. Sin embargo, la anemia es un problema que a la fecha sigue afectando a muchos niños y niñas, y es causada por diversos factores.

Mundialmente el 20% tanto de niños como de niñas, cuyo rango de edad se encuentra entre 6 a 59 meses, son afectados por la anemia (OMS, 2023).

En el año 2023, en América Latina y el Caribe, los indicadores de prevalencia de anemia promedio en la infancia (menores a 5 años) es del 24.6%. Los países con mayor prevalencia de anemia son Haití (60.1%), Bolivia (36.9%), República Dominicana (36.9%), Perú (29.6%) (OMS, 2023).

Figura 3

Porcentaje dominante de anemia en niños y niñas menores de 5 años, 2019



Nota. Datos tomados del Banco Mundial. Elaboración propia.

En el Perú, de acuerdo con el INEI (2023), la prevalencia de la anemia en los últimos dos años se incrementó en un 5.2%, siendo el incremento de niños con anemia para el 2021 de 28.9% y para el 2022 de 33.6%.

La evaluación exhaustiva de la prevalencia de la anemia muestra que para el 2022, el área rural evidenció un 42.4% de niños con anemia y el área urbana un 30.2%. A nivel departamental, Puno (66.3%) presenta mayor prevalencia de anemia; así mismo, los departamentos de Madre de Dios, Loreto, Pasco, Ucayali, Junín, Cusco, Huancavelica y Apurímac presentan un porcentaje entre 40 a 57.9% (INEI, 2023).

Es así que la OMS y el INEI nos presentan evidencia cuantitativa de la problemática de la salud pública sobre la prevalencia e incremento de la anemia en los últimos dos años.

2.5.1. Causas que provocan la anemia

Diversas causas conllevan a padecer de anemia; sin embargo, una de las causas más frecuentes que conllevan a que niños y niñas padezcan de anemia es la insuficiencia de hierro en el cuerpo humano. Cerca del 60% de niños que la población con anemia es por la inadecuada ingesta de hierro (Zavaleta y Astete-Robilliard, 2017).

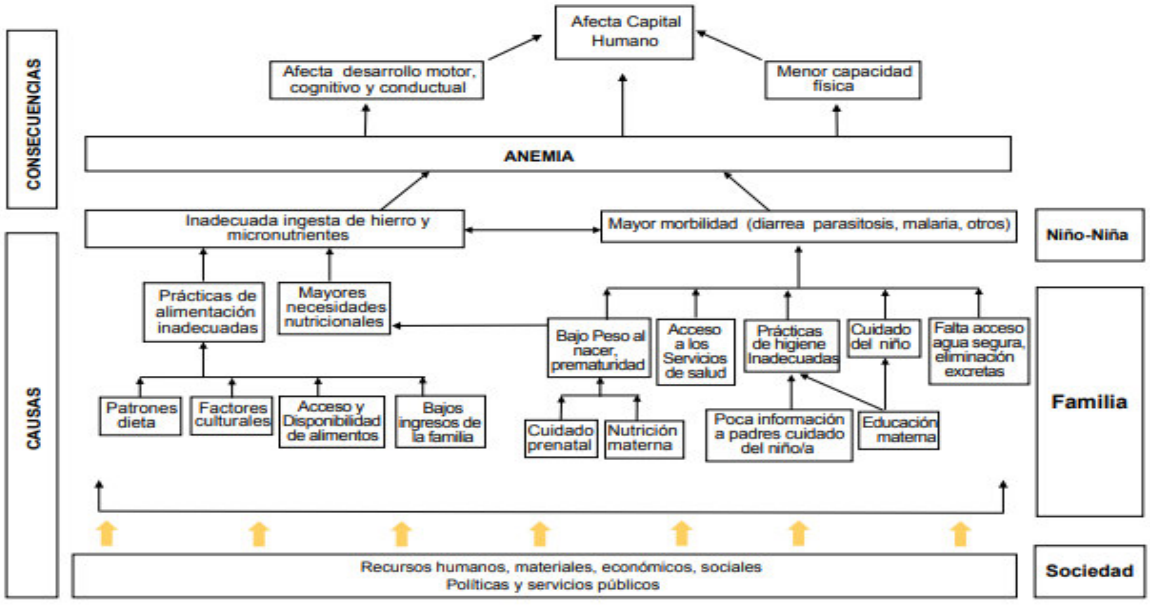
Así mismo, diversos estudios relacionados con la anemia evidencian que esta se asocia al cuidado de la salud del niño (lugar de residencia, área, región, nivel socioeconómico bajo, pobreza monetaria, madre y/o padre adolescente, bajo nivel educativo, fiebre reciente y falta de control prenatal) y a múltiples fenómenos sociodemográficos.

2.5.2. Consecuencias de la anemia

La anemia puede generar consecuencias inmediatas en niños y niñas como el crecimiento retardado, fatiga (síntomas), debilidad, palidez, irritabilidad, así como deficiencia en la atención y efectos a largo plazo como: bajo desarrollo mental (déficit cognitivo o bajo desempeño cognitivo), respuesta emocional lenta, déficit de atención e hiperactividad, pérdida de productividad en adultos mayores y pérdidas al Estado. Otra consecuencia es la pérdida del 0.62% del Producto Bruto Interno (PBI), entre el 2009 y 2010 significó cerca del 40% del presupuesto para el sector de la salud (Zavaleta y Astete-Robilliard, 2017).

Figura 4

Anemia en la población infantil: causas y consecuencias



Nota. Tomado de la *Revista Peruana de Medicina Experimental y Salud Pública* (p. 717).

2.5.3. Cómo determinar si un niño padece anemia

El sistema HemoCue, la cual mide la anemia, es una técnica confiable y usada para la detección de la anemia en diferentes países, así como Perú. Desde 1996, se viene realizando la medición de la hemoglobina con el método HomeCue (INEI, 2023).

2.5.4. Clasificación de la anemia según el INEI

Según el INEI (2023), para los niños y niñas menores de 5 años, existen tres niveles de clasificación: severa, moderada y leve.

Cuadro 1

Clasificación de la anemia

Clasificación	Hemoglobina
Severa	$< 7{,}0\ g/dl$
Moderada	$[7\ a\ 9.9\ g/dl]$
	$[10.0\ a\ 11.9\ g/dl]$
Leve	$[10.0\ a\ 10.9\ g/dl]$*
	Nota (*): Solo para mujeres embarazadas y niños(as).

Nota. Elaboración propia. Adaptado del INEI (2023).

Las causas de la anemia se generan por factores diversos, las cuales conllevan a padecer de esta enfermedad. El objetico del presente trabajo muestra una aplicación de modelos de machine learning "Random Forest" para identificar los factores que influyen en el problema de salud pública (anemia) en niños y niñas entre 0 a 59 meses, haciendo uso de la base de datos de ENDES – INEI del año 2022.

2.5.5. Alcances de la investigación usando el modelo Random Forest

En el sector salud y en diversos ámbitos sociales, para una adecuada toma de decisiones, la investigación busca contribuir con una alternativa de solución para el cálculo o identificación de aquellos factores asociados con la anemia (Random Forest).

2.5.6. Formulación del problema

- ¿Es posible el uso de los "Random Forest" como método alternativo para determinar los factores que están asociados con la anemia en niñas y niños de 6 a 59 meses en el Perú, usando la ENDES 2022?

2.5.7. Formulación de los problemas específicos

- ¿El "Random Forest" es un método eficiente para determinar los factores que se asocian con la anemia en niñas y niños entre los 6 a 59 meses en el Perú, usando la ENDES 2022?
- ¿Qué factores están asociados con la anemia en niñas y niños entre 6 a 59 meses, usando "Random Forest"?
- ¿Existen diferencias según área de residencia entre los factores identificados de la anemia, usando "Random Forest"?

2.5.8. Objetivo principal

- Aplicar el "Random Forest" como un método alternativo para determinar los factores asociados con la anemia en niñas y niños entre 6 a 59 meses en el Perú, 2022.

2.5.9. Objetivos específicos

- Evaluar si la técnica "Random Forest" es un método eficiente para determinar los factores que se asocian con la anemia en niños y niñas entre 6 a 59 meses, en el Perú, 2022.
- Identificar los factores asociados con la anemia en niños y niñas entre 6 a 59 meses, usando "Random Forest".

- Identificar las diferencias entre los factores asociados con la anemia según el área de residencia, usando “Random Forest”.

III. MARCO TEÓRICO

3.1. Bases teóricas

3.1.1. Árboles de decisión

Según Véliz (2018) los árboles de decisión (algoritmos de aprendizaje supervisado) clasifican y predicen, usando variables predictoras numéricas o categóricas, y pueden ser robustas. Se usa árbol de clasificación cuando la variable respuesta es cualitativa (categórica), y se usa árbol de regresión cuando es cuantitativa (continua).

3.1.2. Random Forest

Es un método que combina muchas "construcciones" simples de modelos "bosques aleatorios" para obtener un modelo único y poderoso.

Según ha señalado Breiman (2005), estos bosques se caracterizan por ser una combinación de árboles de decisión del vector aleatorio que realiza el muestreo de manera independiente con una distribución equitativa para los árboles de decisión, para posteriormente predecir las variables de interés.

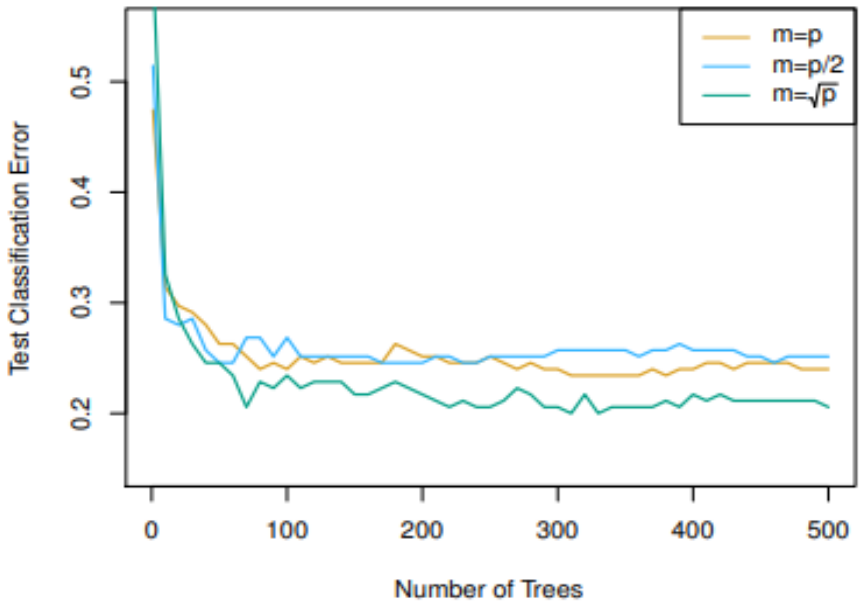
Para James et al. (2021), los bosques aleatorios presentan una mayor ventaja respecto a los árboles embolsados por medio de una aleatoriedad de la construcción de árboles de decisión, correlacionándose con otros árboles. En el embolsado, se construye un número de árboles usando las muestras de entrenamiento Bootstrap, sin embargo, en el procedimiento para la construcción de dichos árboles, cada vez que se considera una división en un árbol, en la muestra aleatoria de

m predictores se escogen candidatos divididos del conjunto completo de p predictores. La división puede usar solo uno de esos m predictores.

En cada una de las divisiones, se toma una muestra de m predictores, y con frecuencia se elige m ≈ √p, es decir, el número de predictores seleccionados en cada división es aproximadamente igual.

Figura 5

Error de clasificación según número de cuenta árboles



Nota. Ejemplo de variable de importancia usando el índice de Gini. Tomado de "*An introduction to statistical learning: with applications in R"*, de Gareth James et al., 2021.

La idea esencial de los bosques aleatorios en el embolsado de árboles es realizar un promedio de una gran cantidad de modelos ruidosos, pero aproximadamente imparciales y, en consecuencia, se reduce la varianza. Los árboles son considerados como los candidatos ideales para el embolsado, ya que pueden realizar interacciones complejas con los árboles.

3.1.3. Formación del algoritmo del modelo "Random Forest"

El algoritmo se utiliza frecuentemente cuando las variables de entrada son altas. Los parámetros del algoritmo corresponden al número de variables que se cultivan en el bosque y el número de características que se elegirá en cada nodo al momento de construir individualmente.

Se toma en cuenta los siguientes pasos:

- Paso 1: Generar cada árbol, b=1 a B

a) Extraer una muestra bootstrap Z* (para datos de entrenamiento)
b) Construir árboles Random Forest T_b (repetir pasos hasta obtener nodo mínimo n_{min})
   b.1) Seleccionar m variables al azar (m=1,..,p)
   b.2) Elegir m variables (m: punto de división)
   b.3) Dividir nodo (2 dos) $\{T_b\}_1^B$

- Paso 2: Para hacer una predicción en un nuevo punto x:

$$Regresión: f_{pf}^{2B}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$$

Clasificación: Sea $\hat{C}\,b\,(x)$, predicción de b ésimo (Random Forest), entonces $\hat{C}\,rf\;B\,(x) = voto\ mayoritario\ \{\hat{C}\,b\,(x)\}_1^B$

3.1.4. Ventajas

Las ventajas identificadas en los modelos Ramdom Forest son los siguientes:

- Selección de variables predictores de automáticamente.
- Se pueden usar como modelos de regresión y clasificación.
- Los árboles son capaces de manejar variables predictoras del tipo numérico y categórico sin tener que transformar a una variable dummy (también conocida como one-hot-encoding). Este criterio se apoya en la implementación del algoritmo y desarrollo de cada librería.
- Requiere menos tratamiento (limpieza de base de datos) y preprocesamiento de datos respecto a otras técnicas de aprendizaje estadístico (sin la necesidad de estandarizar datos).
- Es menor la influencia de datos con ruido y datos outliers.
- Identifica las variables predictoras más importantes de forma rápida y eficiente.
- El Out-of-Bag (OOB) Error estima el error de validación y no recurre a métodos computacionalmente como la validación cruzada. Para los casos de series de tiempo no es aplicable.
- Poseen buena escalabilidad y se aplican para una gran cantidad de observaciones.

3.1.5. Desventajas

Las desventajas identificadas en los modelos Ramdom Forest son los siguientes:

- Pérdida de interpretación al combinar múltiples árboles para generar un modelo de Random Fosrest.
- Al momento de categorizar las variables que son continuas para la división de los nodos se pierde información.
- Para la extrapolación solo usan las variables predictoras observadas de los datos de entrenamiento.

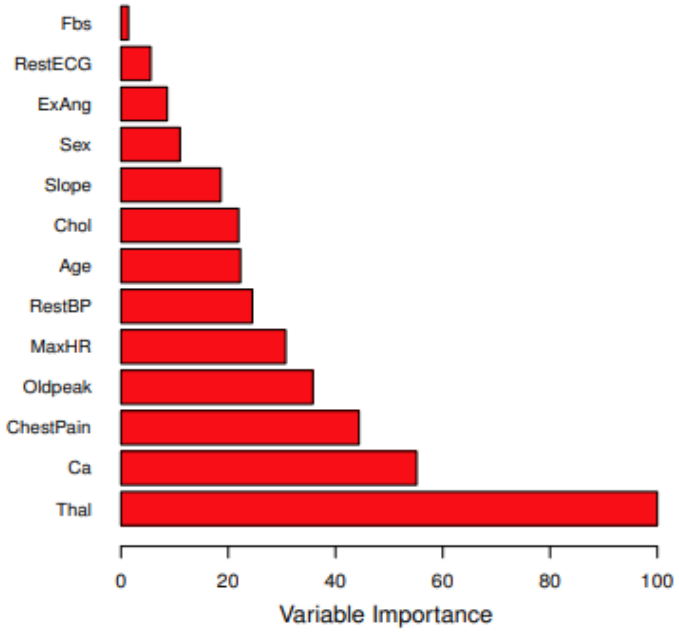
3.1.6. Medidas de las variables de importancia

El embolsado mejora la predicción y precisión a expensas de la interpretabilidad. Es decir, interpretar un gran número de árboles embolsados conlleva una mayor dificultad que la interpretación de uno solo; dicho esto, del valor de cada predictor se obtiene una visión generalizada, y para ello se emplea el índice de Gini (para árboles de clasificación) o el RSS (para árboles de regresión). Es un predictor importante cuando el valor del predictor de importancia es grande. De manera similar, con la clasificación de árboles embolsados, sumamos la cantidad total del índice de Gini (menor índice) mediante las divisiones que se realizan sobre un predictor dado.

Las gráficas de variables de importancia se construyen para los bosques aleatorios del mismo modo que para los modelos potenciados por gradiente.

Figura 6

Error de clasificación según número de árboles



Nota. Ejemplo de variable de importancia usando el índice de Gini. Tomado de "*An introduction to statistical learning: with applications in R"*, de Gareth James et al., 2021.

La construcción de la medida de importancia también se realiza para las muestras de los árboles de decisión por cada variable, aparentemente se podría asegurar que mide la fuerza de predicción de cada variable del estudio.

Es decir, cuando se cultiva el árbol b, las muestras fluyen a través de dicho árbol y se puede observar la exactitud de la predicción.

3.1.7. Comparación de modelos - Curva ROC

La curva de ROC se utiliza para visualizar el equilibrio entre la tasa de falsas alarmas y la de aciertos de los clasificadores, mostrando los resultados para problemas de decisión binario de aprendizaje automático. Es decir, se utiliza frecuentemente para evaluar y comparar algoritmos.

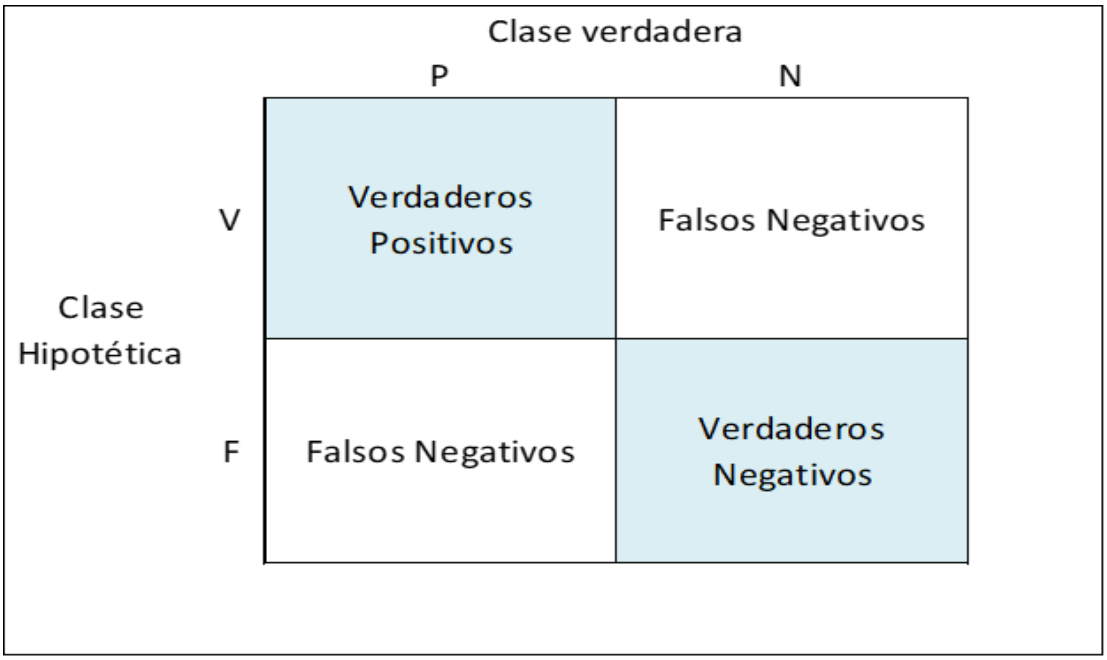
3.1.8. Tabla de clasificación

Esta tabla muestra frecuentemente valores observados y estimados, generando así la probabilidad de sensibilidad, especificidad, precisión, falsos positivos y falsos negativos.

a) Sensibilidad: Conocida como presión positiva. Es la probabilidad de clasificar de forma correcta el modelo seleccionado.
b) Especificidad: Conocida como presión negativa. Es la probabilidad de clasificar correctamente las categorías que no son de interés.
c) Precisión: Probabilidad de que el modelo seleccionado logra clasificar correctamente de manera global.
d) Falso positivo: Proporción de casos negativos que fueron clasificados de manera incorrecta como positivos.
e) Falso negativo: Proporción de casos positivos que fueron clasificados de forma incorrecta como negativos.

Figura 7

Matriz de confusión



Nota. Elaboración propia.

La tasa de precisión positiva se estima como:

$$Tasa\ de\ predicción\ positiva\ = \frac{Verdaderos\ positivos}{Verdaderos\ positivos\ +\ Falsos\ positivos}$$

La tasa de precisión negativa se estima como:

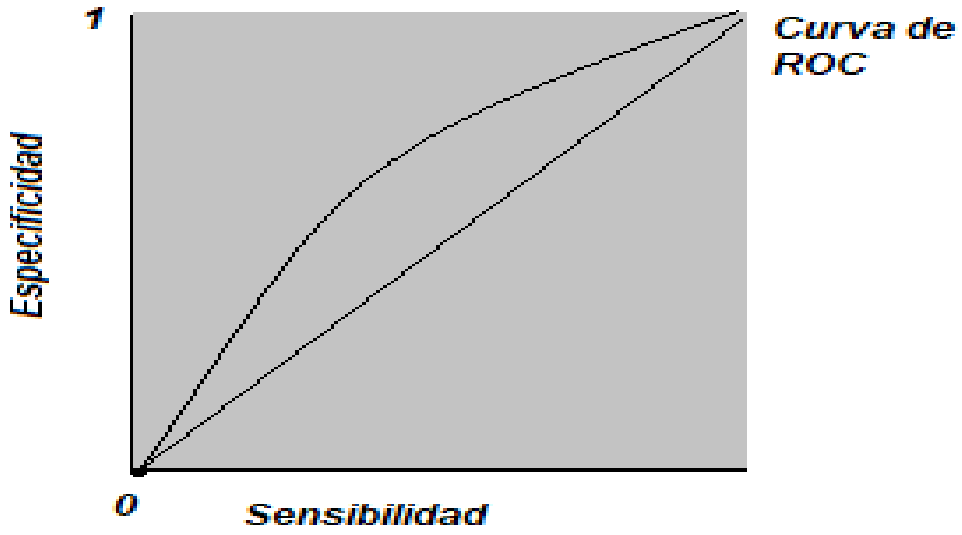
$$Tasa\ de\ predicción\ negativa\ = \frac{Verdaderos\ negativos}{Verdaderos\ negativos\ +\ Falsos\ negativos}$$

3.1.9. Curva ROC

Representación gráfica que se encarga de diagnosticar la capacidad discriminativa para determinar el rendimiento de una clasificación, siendo la división de la tasa de verdaderos positivos y la de falsos negativos.

Figura 8

Curva ROC



Nota. Elaboración propia.

En la curva de ROC, mientras más lejano esté la diagonal principal con área 0.5, mejor es el modelo diagnosticado. La curva de ROC es ideal cuando es igual a 1.

3.1.10. Anemia

Según la OMS (2004), se evidencia a la anemia como un indicador que presenta una nutrición pobre y una salud desfavorable. Se dice que una persona cuenta con anemia si presenta

carencia de hierro, y al presentar alta concentración de hemoglobina es fácil de determinar. En casos más severos, la anemia resulta en anemia ferropénica.

En conclusión, se logra afirmar que la escasez de hierro genera, principalmente, la anemia; sin embargo, no es aceptable la afirmación para entornos donde la causa es más compleja. En cifras porcentuales, se logró evidenciar lo siguiente:

- Niños y niñas, y mujeres que padecían anemia: 40 y 50%
- Niños y niñas, y mujeres que padecían anemia ferropénica: 50%
- Niños y niñas, (2 a 5 años) que padecían anemia ferropénica: 80%

La OMS (s.f.) considera a la anemia como una enfermedad en donde la hemoglobina (proteína esencial para el transporte oxígeno) presenta una concentración baja o un número bajo de glóbulos rojos en relación con lo normal. Además, disminuye “la capacidad de la sangre para transportar oxígeno a los tejidos del organismo, lo que puede causar síntomas como agotamiento, debilidad, mareos y dificultad para respirar, entre otros”.

Según la OMS (s.f.), la anemia se debe a diversos factores:

- La inflamación, las enfermedades crónicas, los problemas ginecológicos y obstétricos, las infecciones (paludismo, infecciones parasitarias, tuberculosis, infección por VIH) y los trastornos hereditarios de los glóbulos rojos son algunos de los factores que pueden provocar deficiencias nutricionales, al igual que una dieta deficiente o una nutrición inadecuada.

- Comúnmente la insuficiencia de hierro es la causa nutricional; mientras que la falta de consumo de folato, vitamina B12 y vitamina A contribuyen de forma significativa.
- Con mayor frecuencia, es causada por la pérdida de sangre durante la menstruación (mujeres adolescentes en edad reproductiva).

Así mismo, las cifras de la OMS afirman que el 20% de los niños de entre 6 a 59 meses padecen de anemia.

3.1.11. Antecedentes internacionales

En el año 2015, el trabajo titulado "Estudio transversal: desnutrición, anemia y su relación con factores asociados en niños de 6 a 59 meses, Cuenca 2015", en Azuay (Ecuador), en niños de 6 a 59 meses, tuvo por finalidad evaluar la prevalencia de desnutrición, anemia y sus factores asociados (parto prematuro, bajo peso y baja estatura). Los datos recolectados se dieron por medio de interrogatorios, observación directa (antropometría fetal) y las historias clínicas (edad gestacional).

Es así que, por consiguiente, con los resultados se calculó el chi-cuadrado y la razón de prevalencia, obteniendo los siguientes resultados:

- De un total de 737 participantes (niños y niñas), el 47.6% fueron niñas y el 52.4%, niños; logrando concluir que el 5% de la población tiene desnutrición.
- Prevalencia de la anemia en niños y niñas: 2.4% con bajo peso y 10.8% con baja estatura.

- Existe una significancia de la asociación entre la anemia, bajo peso y baja estatura en la etapa neonatal.

De acuerdo con la investigación del año 2019 sobre "Factores asociados a la anemia en niños ecuatorianos entre 1 a 4 años", se buscó determinar los factores relacionados con la anemia en niños y niñas que asisten al Centro de Desarrollo Infantil "Los Pitufos de El Valle", en Cuenca. La investigación tuvo como muestra 52 casos (casos clínicos diagnosticados con anemia) y 52 controles (casos clínicos diagnosticados sin anemia); la población participante de la investigación estuvo conformada por niños y niñas (pacientes del Centro de Desarrollo Infantil "Los Pitufos"). Para el análisis de información se revisó los casos clínicos de los pacientes (hemoglobina en la sangre, los suplementos vitamínicos, el peso y la estatura; en cuanto al peso al nacer y la edad gestacional). El estadístico que se utilizó para analizar los resultados de la investigación fue el método de asociaciones de la razón de momios, $\chi 2$ (prueba de chi-cuadrado) y regresión logística. Por lo tanto, los factores identificados como asociados con la anemia incluyen la residencia rural, las deficiencias de micronutrientes, el bajo peso al nacer y el parto prematuro.

En el 2019, se realizó la investigación "Machine Learning Algorithms To Predict The Childhood Anemia In Bangladesh" con el interés de conocer la probabilidad de anemia usando la predicción del estado de la enfermedad (clave para la formulación de políticas comunitarias y de servicios de salud) y la revisión de la planificación de recursos. Consideramos el aprendizaje automático (ML) para realizar la predicción de la anemia de niños y niñas (menor de 5 años) usando factores de riesgo. Se tomó información secundaria de la encuesta nacional representativa Bangladesh Demografía y Salud Encuesta (BDHS) realizada en el 2011. En este estudio, fue seleccionado una muestra de 2013 niños. Los algoritmos utilizados de ML para predecir el estado

de la anemia son: análisis discriminante lineal (LDA), clasificación y árboles de regresión (CART), k-vecinos más cercanos (k-NN), máquinas de vectores de soporte (SVM), bosque aleatorio (RF) y regresión logística (LR). Por lo tanto, se realizó una evaluación mediante el Random Forest (RF) y se obtuvieron los siguientes resultados: precisión (68.53%), sensibilidad (70.73%), especificidad (66.41%) y el área bajo la curva (0.6857). En el otro lado, el algoritmo LR clásico generó una clasificación del 62.75%, sensibilidad de 63.41%, especificidad de 62.11% con un AUC de 0.6276. El investigador identificó que el k-NN genera menor precisión. Concluyendo que los métodos ML son técnicas de regresión clásicas efectivas que generan buena predicción. Es así que para esta investigación el modelo de factores de riesgo y características sociodemográficas ayuda a predecir si el paciente padece de anemia.

Comparamos varios modelos de predicción de aprendizaje automático para predecir si un paciente tiene anemia dados los factores de riesgo. Entre los modelos considerados, el Random Forest realizó la mejor precisión de clasificación para predecir la anemia en la población de Bangladesh. Este estudio destaca no solo la utilidad de los algoritmos de ML, sino también la importancia de utilizar características sociodemográficas y de salud comunes para predecir la enfermedad estado. Además, nuestros hallazgos serían útiles para la identificación de los niños y niñas en riesgo de anemia en el futuro, brindando a los formuladores de políticas y proveedores de atención médica una herramienta para implementar intervenciones necesarias y mejorar las prácticas de atención. Así, un modelo construido sobre el riesgo común de factores ayudaría en la prevención y control de la anemia infantil.

3.1.12. Antecedentes nacionales

En el 2019, el estudio “Análisis del modelo multicausal sobre el nivel de la anemia en niños de 6 a 35 meses en Perú” identificó la prevalencia de los niveles de anemia y sus factores asociados en niños y niñas menores de 3 años donde se usó el modelo multicausal utilizando la data del INEI, correspondiente a la ENDES (2019). Al finalizar el proceso de estudio (“bondad de ajuste” y “modelo de regresión ordinal”) se identificaron los factores de riesgo: factores subyacentes (“edad de la mujer”, “la diarrea en las últimas dos semanas como factor inmediato”, “control prenatal”, “edad del niño”, “fuente de agua potable” y “anemia”) y factores protectores (“quintil de riqueza superior” y “amamantamiento por alguna vez”).

Nakandakari y Carreño (2023), en su investigación denominada “Factores asociados a la anemia en niños menores de cinco años de un distrito de la Libertad, Huaraz, Ancash”, tuvieron por objetivo determinar qué factores se asocian a la anemia, identificando los siguientes: “sexo masculino”, “edad: mayor a 1 año”, “el pertenecer a un caserío diferente a Cajamarquilla” y “no contar con servicios básicos completos”. Cabe precisar que en el estudio participaron 110 niños con historia clínica, realizando un comparativo de niños con anemia (hemoglobina < 11 gr/ dl) y sin anemia (hemoglobina > 11 gr/ dl).

IV. METODOLOGÍA

4.1. Diseño de investigación y tipo de estudio

El diseño es transversal, no experimental, de alcance descriptivo y explicativo. Con enfoque cuantitativo, del tipo aplicado, ya que es la aplicación de conocimientos teóricos a determinados contextos sociales y las consecuencias prácticas de que ella se deriva (Hernández Sampieri y Mendoza, 2018).

4.2.Población y muestra

Según la ENDES (2022) la población lo conformaron todas las mujeres de doce (12) a cuarenta y nueve (49) años, así como niños y niñas menores de cinco (5) años. El tipo de muestreo que se aplicó es bietápico y probabilístico, independiente y estratificado, con inferencia departamental y área (urbana y rural).

El siguiente cuadro resume la muestra seleccionada.

Cuadro 2

Resumen de la muestra

Área sede	Área urbana	Área rural
43 distritos de Lima Metropolitana y capitales del Perú.		
18 820	9 230	12 600

Nota. Muestra anual conformada por 36 650 viviendas.

4.3.Método para recolección de datos

El presente TSP[1] usó una fuente secundaria de la data, correspondiente a la ENDES (2022). Las bases de datos son de acceso abierto y pueden consultarse en la web del INEI.

Se realizó la revisión de la ficha técnica de la ENDES (2022); los datos recolectados fueron mediante la "entrevista directa" (presencial), los cuales fueron realizados por personas debidamente capacitadas para la recopilación de datos, quienes visitaron las viviendas seleccionadas para aplicar el cuestionario.

Encuestado:

- Jefa(e) del hogar, esposa(o) o una persona de 18 años a más.

4.4. Instrumento de medición

La información recopilada de las viviendas seleccionadas se obtuvo mediante un dispositivo móvil, el dispositivo fue una Tablet (ficha técnica ENDES, 2022).

[1] TSP: trabajo de suficiencia

4.5.Variables

Cuadro 3

Resumen de variables sociodemográficas para la aplicación del Random Forest

	Indicador/ Variable	**Categoría variable**	**Tipo de variable**
Variables respuesta	Indicador de anemia	Con y sin anemia	Cualitativa
Variable independiente	Edad	De 0 a 56 meses	Cuantitativo
	Sexo	Hombre, Mujer	Cualitativa
	Área	Urbana, Rural	Cualitativa
	Espacio de residencia	Capital, gran ciudad, Pequeña ciudad, Pueblo, Campo	Cualitativa
	Departamento	Los 25 departamentos	Cualitativa
	Etnicidad	Idioma o legua materno que aprendió en la niñez. (quecha, aimara, idioma extranjero, castellano)	Cualitativa
	Altitud del conglomerado		Cualitativa
	Nivel educativo de la madre		Cualitativa
	Peso		Cuantitativo
	Talla		Cuantitativo
	Índice de riqueza		Cualitativo

4.6.Procesamiento de la base de datos

Las bases disponibles de la ENDES (2022) están en formato SPSS (*sav), en diferentes módulos. Los módulos usados para el análisis de información fueron:

El software libre que se utilizó para el análisis de información fue "R-Studio", y se procedió a seguir los siguientes pasos para el análisis:

- Revisión y limpieza, e imputación de datos
- Unión de tablas [formato SPS(*sav)]
- Entrenamiento de datos y selección de variable
- Análisis descriptivo de la información (gráficos y tablas)
- Aplicación del método estadístico técnica estadística. ("Random Forest")

La librería usada:

Random Forest: El paquete tiene la capacidad de generar modelos predictivos como el algoritmo de bosque aleatorio de Breiman (que se basa en el código Fortran de Breiman y Cutler) para la clasificación y regresión. Es utilizado también sin supervisión para la evaluación de proximidades entre puntos de datos.

Rocr: Permite evaluar y visualizar el desempeño de clasificadores de puntuación. Los métodos estándar para investigar las compensaciones entre medidas de desempeño específicas están disponibles dentro de un marco uniforme, incluidos gráficos de características operativas del receptor (ROC), gráficos de precisión/recuperación, gráficos de elevación y las curvas de costos (Tobias, Oliver y Beerenwinkel, 2022).

Encargado de realizar las predicciones numéricas junto con las correspondientes etiquetas de clase verdadera, recopilando opcionalmente predicciones y etiquetas para varias ejecuciones de validación cruzada.

V. RESULTADOS

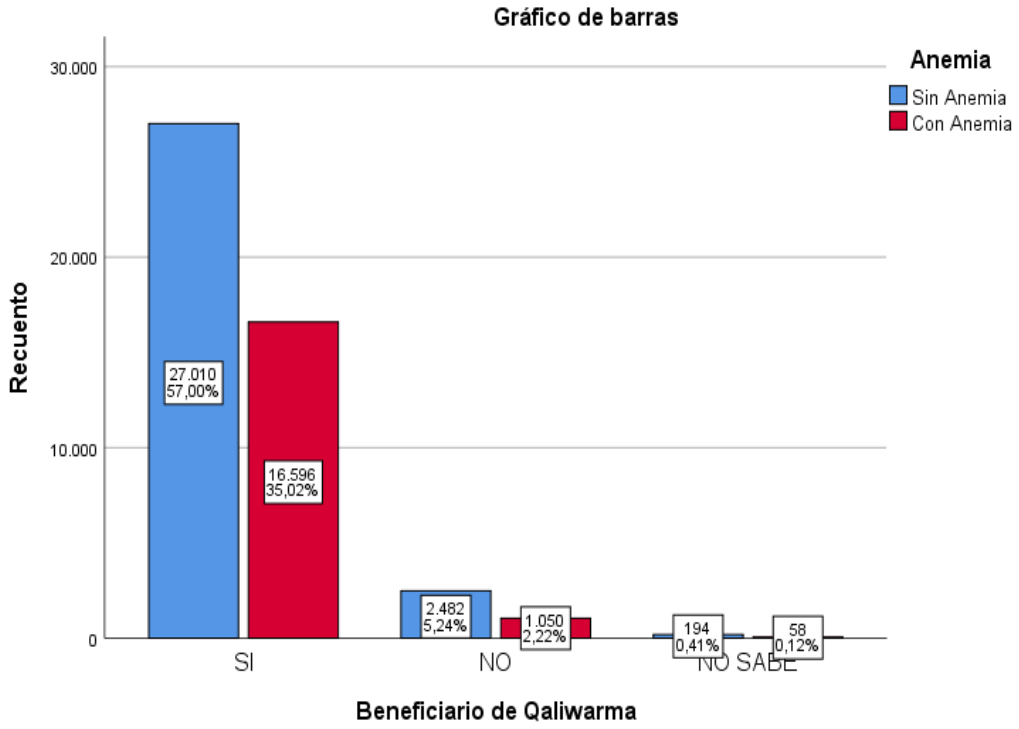
5.1.Análisis descriptivo de factores asociados a la anemia en el Perú

Se analizaron 47 390 niños de 0 a 59 meses de edad, con un total de 29 686 niños sin anemia y 17 704 niños con anemia.

El 35% de beneficiarios de Qali Warma entrevistados tienen anemia, y el 2.22% de niños que no son beneficiarios también tienen anemia.

Figura 9

Beneficios del Programa de Alimentación según indicador de anemia en el Perú, 2022

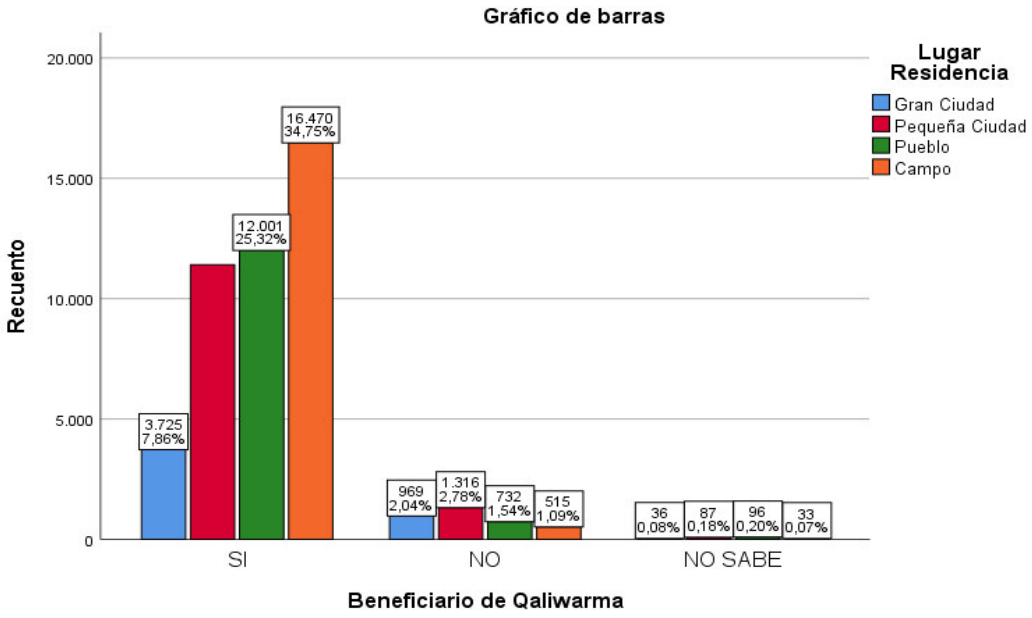


Nota. ENDES (2022).

Gran parte de la población de beneficiarios del programa Qali Warma viven en el campo (34.8%), pueblo (25.3%) y pequeña ciudad.

Figura 10

Beneficios del Programa Nacional de Alimentación según el lugar de residencia, 2022

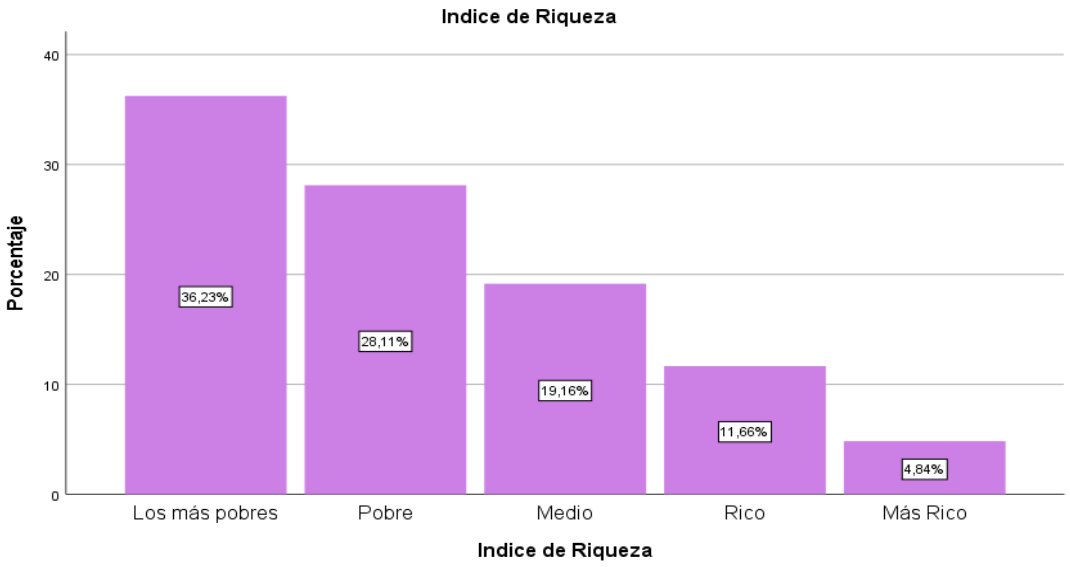


Nota. ENDES (2022).

El índice de riqueza de los entrevistados, más del 50% están en la entrevistados tienen categoría entre índice de riqueza medio, pobre y lo más pobre, destacando el índice más pobre en 36%.

Figura 11

Índice de riqueza de los entrevistados en el Perú, 2022



Nota. ENDES (2022).

En el área urbana viven el 64.1% de niños y niñas registrados de 6 a 59 meses y en el área rural viven el 35.9%.

Cuadro 4

Área de residencia de niños y niñas de 6 a 59 meses en el Perú, 2022

	Frecuencia	Porcentaje
Urbano	30,372	64,1
Rural	17,018	35,9
Total	47,390	100,0

Nota. ENDES (2022).

El 35.9% de niños y niñas de 6 a 59 meses registrados viven en el campo y el 10% de los registrados viven en la capital y gran ciudad.

Cuadro 5

Lugar de residencia de niños y niñas de 6 a 59 meses en el Perú, 2022

	Frecuencia	Porcentaje
Capital, Gran Ciudad	4,730	10.0
Pequeña Ciudad	12,813	27.0
Pueblo	12,829	27.1
Campo	17,018	35.9
Total	47,390	100.0

Nota. ENDES- 2022.

5.2.Porcentaje de sensibilidad y especificidad con el modelo Random Forest

Este modelo determinó una sensibilidad del 46.2% y una especificidad del 64% con una AUC del 57%, logrando mantener un óptimo resultado para la investigación.

Cuadro 6

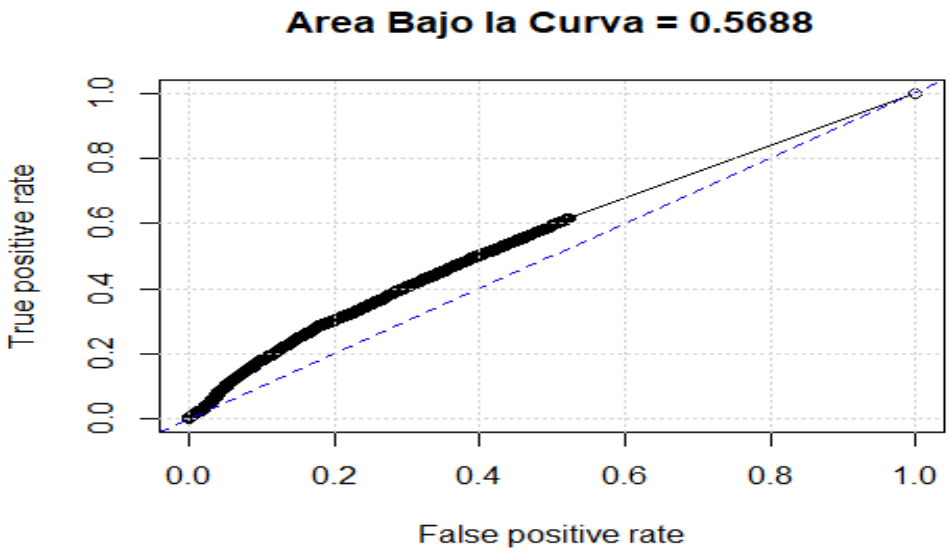
Indicador del modelo Random Forest

Indicador	
Especificidad	46.2%
Sensibilidad	64.0%
AUC	56.9%

Nota. Elaboración propia.

Figura 12

Área bajo la curva (AUC)



Nota. Elaboración propia.

5.3.Importancia de variables

Cuadro 7

Importancia de variables en el modelo Random Forest

Nacionalidad	18.7
Sexo	20.2
Benef_QaliWarma	26.9
Lugar_residencia	46.0
Nivel_est_apo	48.9
Parentesco_jefe_Hogar	50.6
Indice_riqueza	64.4

Nota. Elaboración propia.

5.4.Porcentaje de sensibilidad y especificidad con el modelo Random Forest por área

En los resultados ara el área urbana, se determinó una sensibilidad del 63.3% y una especificidad del 42.3% con una AUC del 56.9%. En el área rural, se determinó una sensibilidad del 63.2% y una especificidad del 44.9% con una AUC del 56.6%.

Cuadro 8

Importancia de variables por área según en el modelo Random Forest

Indicador	Urbana	Rural
Sensibilidad	63.3	63.2
Especificidad	42.8	44.9
AUC	56.9	56.6

Nota. Elaboración propia

Cuadro 9

Importancia de variables en el modelo Random Forest

	meandecreaseaccuracy	
	área urbana	área rural
Indice_riqueza	45.34538	46.21404
Parentesco_jefe_Hogar	31.69221	31.98712
Nivel_est_apo	31.45087	32.01572
Lugar_residencia	22.30887	21.19701
Benef_QaliWarma	15.36229	14.46824
Nacionalidad	13.35503	12.36485
Sexo	10.09839	10.34273

area_urbano	18.61403	-
area_rural	-	18.15834

Nota. Elaboración propia

VI. CONCLUSIONES

La aplicación del modelo machine learning “Random Forest” permitió determinar, en niños y niñas de 6 a 59 meses, los factores que se asocian a la anemia. Se encontró un AUC del 57%, con un indicador de sensibilidad del 46% y una especificidad del 64%. Es decir, el modelo genera una buena predicción dando como resultado su eficiencia para la identificación de dichos factores.

Dentro de la investigación, se identificó los factores influyentes en la anemia, los cuales son: el índice de riqueza, el parentesco con el jefe del hogar, máximo nivel de estudio del apoderado, el lugar de residencia, el beneficiario del PNAEQW.

Los indicadores determinaron lo siguiente: en el área urbana una sensibilidad del 63.3% y una especificidad del 42.3% con una AUC del 56.9%. En el área rural, una sensibilidad del 63.2% y una especificidad del 44.9% con una AUC del 56.6%. Es así que, al calcular los factores de importancia de los resultados obtenidos, estos no varían según el área, logrando obtener los mismos factores por área. Entre los factores más importantes se encuentran los índices de riqueza, el parentesco con el jefe del hogar, máximo nivel de estudios del apoderado, el lugar donde viven.

VII. RECOMENDACIONES

El modelo machine learning “Random Forest” es un buen modelo predictivo para la determinación de los factores asociados a la anemia en niños de 6 a 59 meses, para efectos de aplicación práctica es para identificar el mejor modelo predictivo se recomienda balancear los datos, comparar con distintos modelos y evaluar el AUC potente, así como no olvidar usan datos de test y entrenamiento. Los modelos a comparar son los modelos “bayes”, “redes neuronales”; y este va permitir evaluar y generar mejores resultados para la toma de decisiones. Entrenar el modelo respecto al balanceo de datos, permitirá generar una falsa alarma en la predicción del modelo y reducir el sesgo en la estimación., así como analizar y evaluar la tasa de predicción positiva y negativa.

VIII. REFERENCIAS

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. y Rubin, D. B. (2014). *Bayesian Data Analysis* [Análisis de datos bayesianos] (3° ed.). CRC Press.

Gobierno del Perú. (s.f.). *Información institucional*. https://www.gob.pe/institucion/qaliwarma/institucional

Hernández-Sampieri, R. y Mendoza Torres, C. P. (2018). *Metodología de la Investigación. Las rutas cuantitativa, cualitativa y mixta*. Editorial McGraw Hill.

Instituto Nacional de Estadística e Informática. (2023). *Perú: Encuesta demográfica y de salud familiar ENDES nacional y departamental*. https://www.inei.gob.pe/media/MenuRecursivo/publicaciones_digitales/Est/Lib1898/libro.pdf

Instituto Nacional de Estadística e Informática. (s.f.). *Microdatos*. Recuperado el 12 de julio de 2023. https://proyectos.inei.gob.pe/microdatos/

James, G., et al. (2021). *An Introduction to Statistical Learning with Applications in R* [Una introducción al aprendizaje estadístico con aplicaciones en R] (2° ed.). Springer.

Lasso Lazo, S. R., et al. (2016). Estudio transversal: Desnutrición, anemia y su relación con factores asociados en niños de 6 a 59 meses, Cuenca 2015. *Revista Médica del Hospital José Carrasco Arteaga,* 8(7), 231-237.

Nakandakari, M. D. y Carreño-Escobedo, R. (2023). Factores asociados a la anemia en niños menores de cinco años de un distrito de Huaraz, Ancash. *Revista Médica Herediana*, *34*(1), 20-26. http://www.scielo.org.pe/scielo.php?script=sci_arttext&pid=S1018-130X2023000100020#:~:text=Los%20factores%20asociados%20a%20una,contar%20con%20servicios%20b%C3%A1sicos%20completos

Organización Mundial de la Salud. (s.f.). *Anemia.* Recuperado el 19 de julio de 2023. https://www.who.int/es/health-topics/anaemia#tab=tab_1

Organización Mundial de la Salud. (2004). La anemia como centro de atención (Organización Panamericana de la Salud, Trad.). https://www.unscn.org/layout/modules/resources/files/La_anemia_como_centro_de_atenci%C3%B3n_1.pdf

Sing, T. et al. (2005). ROCR: visualización del rendimiento del clasificador en R [ROCR: visualización del rendimiento del clasificador en R]. *Bioinformatics*, *21*(20), 3940-3941. https://cran.r-project.org/web/packages/ROCR/vignettes/ROCR.html

Tobias, T. et al. (2022). *Visualizing the Performance of Scoring Classifiers* [Visualizar el rendimiento de los clasificadores de puntuación].

Zavaleta, N. y Astete-Robilliard, L. (2017). Efecto de la anemia en el desarrollo infantil: consecuencias a largo plazo. *Revista Peruana de Medicina Experimental y Salud Pública*, 34(4), 716-722.

I. Anexos

a. Anexo 1: Resultados descriptivos

		Anemia				Total
		Sin Anemia		Con Anemia		
		N	%	N	%	N
Sexo	Hombre	14,003	47.2%	8,455	47.8%	22,458
	Mujer	15,683	52.8%	9,249	52.2%	24,932
	Total	29,686	100.0%	17,704	100.0%	47,390
Region	Lima Metropolitana	3,362	11.3%	1,368	7.7%	4,730
	Resto Costa	9,804	33.0%	3,200	18.1%	13,004
	Sierra	8,953	30.2%	6,636	37.5%	15,589
	Selva	7,567	25.5%	6,500	36.7%	14,067
	Total	29,686	100.0%	17,704	100.0%	47,390
Lugar Residencia	Gran Ciudad	3,362	11.3%	1,368	7.7%	4,730
	Pequeña Ciudad	8,465	28.5%	4,348	24.6%	12,813
	Pueblo	8,380	28.2%	4,449	25.1%	12,829
	Campo	9,479	31.9%	7,539	42.6%	17,018
	Total	29,686	100.0%	17,704	100.0%	47,390
Nacionalidad	ARGENTINA	7	0.0%	4	0.0%	11
	BOLIVIANA	15	0.1%	3	0.0%	18
	BRASILERA	6	0.0%	8	0.0%	14
	CHILENA	13	0.0%	3	0.0%	16
	COLOMBIANA	7	0.0%	13	0.1%	20
	ECUATORIANA	5	0.0%	1	0.0%	6
	ESPAÑOLA	2	0.0%	0	0.0%	2
	ESTADOUNIDENSE	0	0.0%	1	0.0%	1
	ITALIANA	0	0.0%	4	0.0%	4
	PERUANA	29,307	98.7%	17,529	99.0%	46,836
	VENEZOLANA	324	1.1%	138	0.8%	462
	Total	29,686	100.0%	17,704	100.0%	47,390
Etnicidad		506	1.7%	221	1.2%	727
	Quechua	4,873	16.4%	3,700	20.9%	8,573
	Castellano	22,864	77.0%	12,259	69.2%	35,123
	Portugues	23	0.1%	37	0.2%	60
	Otra Lengua Extranjera	0	0.0%	4	0.0%	4
	Aimara	651	2.2%	300	1.7%	951
	Ashaninka	179	0.6%	376	2.1%	555
	Awajún/Aguaruna	290	1.0%	281	1.6%	571
	Shipibo/Konibo	119	0.4%	174	1.0%	293
	Shawi/Chayahuita	41	0.1%	132	0.7%	173
	Matsigenka/Machiguenga	0	0.0%	15	0.1%	15
	Achuar	7	0.0%	0	0.0%	7
	Otra lengua nativa u originaria	133	0.4%	205	1.2%	338
	Total	29,686	100.0%	17,704	100.0%	47,390
Indice de Riqueza	Los más pobres	9,045	30.5%	8,125	45.9%	17,170
	Pobre	8,248	27.8%	5,075	28.7%	13,323
	Medio	6,518	22.0%	2,561	14.5%	9,079
	Rico	4,061	13.7%	1,465	8.3%	5,526
	Más Rico	1,814	6.1%	478	2.7%	2,292
	Total	29,686	100.0%	17,704	100.0%	47,390
Beneficiario de Qaliwarma	SI	27,010	91.0%	16,596	93.7%	43,606
	NO	2,482	8.4%	1,050	5.9%	3,532
	NO SABE	194	0.7%	58	0.3%	252
	Total	29,686	100.0%	17,704	100.0%	47,390

Nota. Elaboración propia

b. Anexo 2: Scrip utilizados en Rstudio

```
#..... Ramdom forest ..... #

# Carga el paquete específico del método Random Forest
library(readxl)
library(pacman)
library(dplyr)
library(randomForest)

BD_Anemia <-
read_excel("C:/Users/QUINIA/Downloads/BD_Final_TSP_2023_Karina_Ariza_vf.xlsx")
BD_Anemia <-BD_Anemia[,-1:-5]
BD_Anemia <-BD_Anemia[,-4]
str(BD_Anemia)

#crear variable por àrea

BD_Anemia$area_urbano <-case_when (BD_Anemia$area == "1" ~ 1,
                                   BD_Anemia$area == "2" ~ 0)


BD_Anemia$area_rural <-case_when (BD_Anemia$area == "2" ~ 1,
                                  BD_Anemia$area == "1" ~ 0)

BD_Anemia <- mutate_if(BD_Anemia, is.character, as.factor)
BD_Anemia <- mutate_if(BD_Anemia, is.numeric, as.factor)
str(BD_Anemia)
head(BD_Anemia)

table(BD_Anemia$Edad)


# Selección de una submuestra del 70% de los datos
BD_Anemia$random<-sample(0:1,size = nrow(BD_Anemia),replace = T,prob = c(0.3,0.7))

train<-filter(BD_Anemia,random==1)
test<-filter(BD_Anemia,random==0)

#Eliminamos ya la random
```

```
BD_Anemia$random <- NULL

str(BD_Anemia)
modelo <- randomForest(BD_Anemia$T_anemia ~ BD_Anemia$Indice_riqueza +
BD_Anemia$Benef_QaliWarma + BD_Anemia$Lugar_residencia
                + BD_Anemia$Sexo  + BD_Anemia$Nivel_est_apo +
BD_Anemia$Parentesco_jefe_Hogar + BD_Anemia$Nacionalidad
                , data=train,
                importance=TRUE)

modelo_rural <- randomForest(T_anemia ~ Indice_riqueza + Benef_QaliWarma + Lugar_residencia
                    + Sexo  + Nivel_est_apo + Parentesco_jefe_Hogar + Nacionalidad + area_rural

                    , data=train,
                    importance=TRUE)
modelo_urbano <- randomForest(T_anemia ~ Indice_riqueza + Benef_QaliWarma +
Lugar_residencia
                      + Nivel_est_apo + Parentesco_jefe_Hogar + Nacionalidad + area_urbano
                     , data=train,
                     importance=TRUE)

# Resumen del ajuste del modelo
modelo
varImpPlot(modelo)

pred = predict(modelo, data=train)
pred_urbano = predict(modelo_urbano, data=train)
pred1 = prediction(pred_urbano , data=train)
table(pred, BD_Anemia$T_anemia)
pred=as.factor(pred)

roc = rendimiento (pred, "tpr", "fpr")

imp <- importance(modelo, type=1)
imp

library(ROCR)
library(pROC)
library(ggplot2)
```

```
# PREDICCION
#
predict <- predict(modelo , data=test , type="prob") [,2 ]
predict_urbano <- predict(modelo_urbano , data=test , type="prob") [,2 ]
predict_rural <- predict(modelo_rural , data=test , type="prob") [,2 ]

rocobj <- prediction (predict , BD_Anemia$T_anemia )
rocobj1 <- prediction (predict_urbano  , train$T_anemia )
rocobj2 <- prediction (predict_rural  , train$T_anemia )
perf.rocr <- performance(rocobj,"tpr","fpr") #True y False postivie.rate
perf.rocr1 <- performance(rocobj1,"tpr","fpr")
perf.rocr2 <- performance(rocobj2,"tpr","fpr")

auc <- as.numeric(performance(rocobj ,"auc")@y.values)
auc_urbano <- as.numeric(performance(rocobj1 ,"auc")@y.values)
auc_rural <- as.numeric(performance(rocobj2 ,"auc")@y.values)
plot(perf.rocr,type='o', main = paste('Area Bajo la Curva =',round(auc,4)))
plot(perf.rocr1,type='o', main = paste('Area Bajo la Curva =',round(auc_urbano,4)))
plot(perf.rocr2,type='o', main = paste('Area Bajo la Curva =',round(auc_rural,4)))
abline(a=0,b=1,col="blue",lty=2)
grid()
auc_urbano
auc_rural

table(predict, test$T_anemia)

### Predicciòn por àrea : urbana, rural
library(agricolae)

BD_Anemia_area<-BD_Anemia

str(BD_Anemia_area)

BD_Anemia_area$area_urbano <-case_when (BD_Anemia_area$area == "1" ~ 1,
                          BD_Anemia_area$area == "2" ~ 0)

str(BD_Anemia_area$area_urbano)

BD_Anemia_area$area_rural <-case_when (BD_Anemia_area$area == "2" ~ 1,
                          BD_Anemia_area$area == "1" ~ 0)
```

```
BD_Anemia_area <- mutate_if(BD_Anemia_area, is.numeric, as.factor)
BD_Anemia_area <- mutate_if(BD_Anemia_area, is.character, as.factor)

str(BD_Anemia_area)

BD_Anemia_area$random<-sample(0:1,size = nrow(BD_Anemia_area),replace = T,prob = c(0.3,0.7))
train1<-filter(BD_Anemia_area,random==1)
test1<-filter(BD_Anemia_area,random==0)

#Eliminamos ya la random
BD_Anemia_area$random <- NULL

str(BD_Anemia_area)
head(BD_Anemia_area)
modelo_rural <- randomForest(T_anemia ~ Indice_riqueza + Benef_QaliWarma + Lugar_residencia
                + Sexo  + Nivel_est_apo + Parentesco_jefe_Hogar + Nacionalidad + area_rural
                , data=train1,
                importance=TRUE)
modelo_urbano <- randomForest(T_anemia ~ Indice_riqueza + Benef_QaliWarma +
Lugar_residencia
                 + Sexo  + Nivel_est_apo + Parentesco_jefe_Hogar + Nacionalidad + area_urbano
                 , data=train1,
                 importance=TRUE)
# Resumen del ajuste del modelo
modelo_urbano
modelo_rural
varImpPlot(modelo_rural)
varImpPlot(modelo_urbano)

pred_urbano = predict(modelo_urbano, data=train1)
pred_rural = predict(modelo_rural, data=train1)
table(pred_rural, BD_Anemia_area$T_anemia)
pred=as.factor(pred_urbano)

pred
imp_urbano <- importance(modelo_urbano, type=1); imp_urbano
imp_rural <- importance(modelo_rural, type=1); imp_rural

summary(modelo_urbano)
```

```
library(ROCR)
library(pROC)
library(ggplot2)

# PREDICCION
#
predict_urbano <- predict(modelo_urbano , data=test1[,-16 ] , type="prob") [,2 ]
predict_rural <- predict(modelo_rural , data=test1 , type="prob") [,2 ]

rocobj1 <- prediction (predict_urbano , BD_Anemia_area$T_anemia )
rocobj2 <- prediction (predict_rural , BD_Anemia_area$T_anemia )
perf.rocr1 <- performance(rocobj1,"tpr","fpr") #True y False postivie.rate

auc <- as.numeric(performance(rocobj ,"auc")@y.values)
plot(perf.rocr,type='o', main = paste('Area Bajo la Curva =',round(auc,4)))
abline(a=0,b=1,col="blue",lty=2)
grid()
auc
```