UNIVERSIDAD NACIONAL DE SAN ANTONIO ABAD DEL CUSCO

ESCUELA DE POSGRADO

MAESTRÍA EN ESTADÍSTICA

TESIS

ALGORITMOS DE MACHINE LEARNING PARA PREDECIR LA ANEMIA EN NIÑOS DE 6 A 35 MESES DE EDAD EN CUSCO

PARA OPTAR AL GRADO ACADÉMICO DE MAESTRO EN ESTADÍSTICA

AUTOR
Br. EULIOS MAMANI BARRIOS

ASESOR
Mg. EDGAR CENTENO HUAMANÍ
CÓDIGO ORCID: 0000-0002-7736-1299

CUSCO - PERÚ

2023

INFORME DE ORIGINALIDAD

(Aprobado por Resolución Nro.CU-303-2020-UNSAAC)

El que suscribe, **Asesor** del trabajo de investigación/tesis titulada:.....
Algoritmos de Machine Learning para predecir la anemia en niños de 6 a 35 meses de edad en Cusco
.....

presentado por: Eulios Mamani Barrios con DNI Nro.: 43634217 presentado por: ..... con DNI Nro.: ..... para optar el título profesional/grado académico de Maestro en Estadística
.....

Informo que el trabajo de investigación ha sido sometido a revisión por 2 veces, mediante el Software Antiplagio, conforme al Art. 6° del ***Reglamento para Uso de Sistema Antiplagio de la UNSAAC*** y de la evaluación de originalidad se tiene un porcentaje de 9 **%.**

Evaluación y acciones del reporte de coincidencia para trabajos de investigación conducentes a grado académico o título profesional, tesis

Porcentaje	Evaluación y Acciones	Marque con una (X)
Del 1 al 10%	No se considera plagio.	X
Del 11 al 30 %	Devolver al usuario para las correcciones.	
Mayor a 31%	El responsable de la revisión del documento emite un informe al inmediato jerárquico, quien a su vez eleva el informe a la autoridad académica para que tome las acciones correspondientes. Sin perjuicio de las sanciones administrativas que correspondan de acuerdo a Ley.	

Por tanto, en mi condición de asesor, firmo el presente informe en señal de conformidad y **adjunto** la primera página del reporte del Sistema Antiplagio.

Cusco, 05 de febrero de 2024

.....
Firma

Post firma... Edgar Centeno Huamani

Nro. de DNI... 31032950

ORCID del Asesor... 0000-0002-7736-1299

Se adjunta:

1. Reporte generado por el Sistema Antiplagio.
2. Enlace del Reporte Generado por el Sistema Antiplagio: **oid:** 27259:323744451

NOMBRE DEL TRABAJO
tesis de maestria.pdf

AUTOR
Eulios Mamani Barrios

RECUENTO DE PALABRAS
14013 Words

RECUENTO DE CARACTERES
78136 Characters

RECUENTO DE PÁGINAS
74 Pages

TAMAÑO DEL ARCHIVO
2.7MB

FECHA DE ENTREGA
Feb 4, 2024 2:06 PM GMT-5

FECHA DEL INFORME
Feb 4, 2024 2:07 PM GMT-5

9% de similitud general

El total combinado de todas las coincidencias, incluidas las fuentes superpuestas, para cada base d

- 9% Base de datos de Internet
- 0% Base de datos de publicaciones
- Base de datos de Crossref
- Base de datos de contenido publicado de Crossr
- 5% Base de datos de trabajos entregados

Excluir del Reporte de Similitud

- Material bibliográfico
- Material citado
- Material citado
- Coincidencia baja (menos de 12 palabras)
- Bloques de texto excluidos manualmente

Dedicatoria

A mi madre Juana María por el apoyo incondicional

A mis hermanos Olger, Elvis y Abel por la compañía en el proceso de mis estudios de la Maestría

A mis amigos por escucharme cuando les comentaba sobre mis estudios de la Maestría

Agradecimiento

A la Escuela de Posgrado de la Universidad Nacional de San Antonio Abad del Cusco por los conocimientos impartidos por medio de la Maestría en Estadística y la paciencia por los pagos retrasados.

A los Docentes de los diferentes cursos impartidos, donde tuve la oportunidad de debatir y compartir mis conocimientos en Estadística.

Índice General

Dedicatoria i

Agradecimiento ii

Índice General iii

Índice de tablas v

Índice de figuras vi

Resumen vii

Abstract viii

Presentación ix

Introducción x

I. PLANTEAMIENTO DEL PROBLEMA 1

1.1. Situación problemática 1

1.2. Formulación del problema 2

1.3. Justificación de la investigación 2

1.4. Objetivos de la investigación 3

II. MARCO TEÓRICO CONCEPTUAL 4

2.1. Bases teóricas 4

2.2. Marco conceptual 25

2.3. Antecedentes empíricos de la investigación 25

III. HIPÓTESIS Y VARIABLES 28

3.1. Hipótesis general 28

3.2. Operacionalización de variables 28

IV. METODOLOGÍA 31

4.1. Ámbito de estudio: localización política y geográfica 31

4.2. Tipo y nivel de investigación 31

4.3. Población de estudio 31

4.4. Tamaño de muestra 31

4.5. Técnicas de recolección de información 32

4.6. Técnicas de análisis de los datos 32

V. Resultados y Discusión 35

5.1. Análisis exploratorio de datos 35

5.2. Comparación de modelos 44

5.3. Discusión de resultados 45

CONCLUSIONES 47

RECOMENDACIONES 48

BIBLIOGRAFÍA 49

ANEXOS 51

Índice de tablas

Tabla 1. Concentración de hemoglobina y niveles de anemia en Niños, Adolescentes, Mujeres gestantes y Puérperas 4

Tabla 2. Operacionalización de variables..... 29

Tabla 3. Métricas de validación de los algoritmos 45

Índice de figuras

Figura 1. Aprendizaje supervisado y no supervisado 6

Figura 2. Árbol de decisión para establecer recomendación de la cirugía ocular 10

Figura 3. Neurona biológica y artificial 12

Figura 4. Red neuronal artificial formada por cuatro capas de neuronas 13

Figura 5. Hiperplano (w, b) equidistante a dos clases, margen geométrico (γ) y vectores soporte (puntos rayados) 17

Figura 6. Topología de un clasificador Naive Bayes 19

Figura 7. Fases del KDD (Knowledge Discovery in Databases)..... 22

Figura 8. Matriz de confusión 24

Figura 9. Frecuencia de diferentes diagnósticos de anemia en niños..... 35

Figura 10. Frecuencia del tipo de servicio higiénico del hogar del niño con diferente diagnóstico de anemia..... 36

Figura 11. Frecuencia del índice de riqueza del niño con diferente diagnóstico de anemia 37

Figura 12. Frecuencia del material predominante del piso de la vivienda del niño con diferente diagnóstico de anemia..... 38

Figura 13. Frecuencia de miembros que reciben ayuda de Wawa wasi y/o Cuna mas en la vivienda de niños con diferente diagnóstico de anemia..... 38

Figura 14. Distribución de los niveles de hemoglobina de niños con diferente diagnóstico de anemia 39

Figura 15. Porcentaje de diferentes diagnósticos de anemia según la altitud de residencia de niños 40

Figura 16. Porcentaje de diferentes diagnósticos de anemia según la talla de niños 40
Figura 17. Correlaciones entre las variables de estudio 41
Figura 18. Distribuciones y dispersión de puntos entre las variables de estudio 42
Figura 19. Variables importantes 43
Figura 20. Precisión de los algoritmos de Machine Learning 44

Resumen

El objetivo de la presente investigación es determinar el algoritmo de Machine Learning más eficiente para predecir la anemia en niñas y niños de 6 a 35 meses de edad a partir de factores sociodemográficos, biológicos, etc. en Cusco. Se utilizo la información de la Encuesta Demográfica y de Salud Familiar (ENDES) de los años 2019 y 2020. La metodología aplicada fue el Knowledge Discovery in Databases (KDD) para el desarrollo de los modelos de Machine Learning. Al evaluar la precisión, los algoritmos K-Nearest Neighbor (KNN), Support Vector Machine (SVM) y Redes Neuronales consiguen más predicciones correctas de acorde a las métricas de Accuracy, Sensibilidad y Especificidad. Finalmente se determinó que el mejor algoritmo fue el Support Vector Machine (SVM) con una Sensibilidad del 100%.

Palabras clave: Machine Learning, Algoritmo, Anemia, Predicción, Support Vector Machine.

Abstract

The objective of this research is to determine the most efficient Machine Learning algorithm to predict anemia in girls and boys from 6 to 35 months of age based on sociodemographic, biological, etc. factors. in progress. Information from the Demographic and Family Health Survey (ENDES) for the years 2019 and 2020 was used. The methodology applied was Knowledge Discovery in Databases (KDD) for the development of Machine Learning models. When evaluating accuracy, the K-Nearest Neighbor (KNN), Support Vector Machine (SVM) and Neural Networks algorithms achieve more correct predictions according to the Accuracy, Sensitivity and Specificity metrics. Finally, it was determined that the best algorithm was the Support Vector Machine (SVM) with a Sensitivity of 100%.

Keywords: Machine Learning, Algorithm, Anemia, Prediction, Support Vector Machine.

Presentación

Señores:

Director(a) de la Escuela de Posgrado

Docentes miembros del jurado.

En cumplimiento al reglamento general de la Escuela de Posgrado con mención en Estadística de la Universidad Nacional de San Antonio Abad del Cusco y con el propósito de optar al grado académico de Maestro en Estadística, presento el trabajo de investigación titulado “ALGORITMOS DE MACHINE LEARNING PARA PREDECIR LA ANEMIA EN NIÑOS DE 6 A 35 MESES DE EDAD EN CUSCO”.

La investigación ofrece un instrumento alternativo que permite predecir la anemia en niños de 6 a 35 meses de edad en Cusco, que servirá para mejorar la toma de decisiones en cuanto a la prevención, procedimientos diagnósticos o tratamientos eficaces.

Introducción

En el Perú, un problema recurrente en salud pública es la anemia. Los departamentos con porcentajes de anemia mayores al 50% en el año 2019 son: Puno, Ucayali, Loreto, Huancavelica, Madre de Dios, Junín y Cusco. En el departamento de Cusco, según la Encuesta Demográfica y de Salud Familiar (ENDES) ejecutada por el Instituto Nacional de Estadística e informática (INEI), en los años 2017, 2018 y 2019 se registraron 55.3%, 54.2% y 57.4% de prevalencia de anemia respectivamente en niñas y niños de 6 a 35 meses de edad. En el año 2020 en un contexto de COVID-19, se estimó la prevalencia de anemia entre 50% y 57.2% en el departamento de Cusco.

Las Tecnologías de Información y Comunicaciones (TICs) tienen una utilidad importante en la administración y vigilancia de prioridades sanitarias. En los últimos años se desarrollaron diferentes aplicaciones de Machine Learning (máquinas de aprendizaje) una rama de la Inteligencia Artificial (IA) para el fortalecimiento de la lucha contra la anemia.

Ante la ausencia de precisión en el diagnóstico de la anemia en niños y niñas por diversos factores, es necesario determinar un algoritmo de Machine Learning para realizar predicciones precisas de la anemia en niñas y niños de 6 a 35 meses de edad en Cusco, de manera que, el personal de salud pueda mejorar la toma de decisiones en cuanto a la prevención, procedimientos diagnostico o tratamientos eficaces.

I. PLANTEAMIENTO DEL PROBLEMA

1.1. Situación problemática

La anemia es un problema de salud pública en el Perú. De acuerdo con las estadísticas de la Encuesta Demográfica y de Salud Familiar (ENDES) realizada por el Instituto Nacional de Estadística en Informática (INEI), en el 2019 se estimó que, el 40.1% de niños y niñas de 6 a 35 meses presentan anemia, porcentaje menor al 43.5% reportado el año 2018. En el ámbito rural (49%) se registró una cifra mayor de casos de anemia en comparación al ámbito urbano (36.7%). En las regiones de Cusco y Puno se incrementaron en 3.2% y 2.2% respectivamente en el año 2019.

Por otra parte, en los últimos años el incremento del uso de teléfonos móviles y el internet han facilitado nuevas maneras de comunicación extensa y de interacción con los ciudadanos. En este ámbito, las aplicaciones de las Tecnologías de Información y Comunicaciones (TICs), asumen una utilidad importante para mejorar la administración y vigilancia de prioridades sanitarias. Espinoza, Henríquez y Villanueva (2019) resumen diferentes aplicaciones, para el fortalecimiento de la lucha contra la anemia en nuestro país, en donde mencionan el uso de mensajes de texto a través de teléfonos móviles y desarrollo de aplicativos para teléfonos inteligentes.

En el campo de la ciencia médica, predecir enfermedades como la probabilidad de presentar anemia a partir de factores sociodemográficos, biológicos, etc. es de suma importancia, para una adecuada toma de decisiones futuras referente a los procesos de asistencia en salud y formular políticas públicas en salud en beneficio de la población de Cusco.

Por lo anterior, consideramos necesario determinar un algoritmo de Machine Learning supervisado (redes neuronales artificiales, arboles de decisión, bosques aleatorios, naives bayes, regresión logística, máquina de vectores de soporte o análisis discriminante lineal) para realizar predicciones precisas de la probabilidad de presentar anemia en niños de 6 a 35 meses de edad asociados a factores frecuentes, utilizando datos de la Encuesta Demográfica y de Salud Familiar (ENDES) correspondiente a los años 2019 y 2020.

1.2. Formulación del problema

a. Problema general.

¿Cuál es el algoritmo de Machine Learning más eficiente para predecir la anemia en niños de 6 a 35 meses de edad en Cusco?

b. Problemas específicos.

¿Cuáles son las variables más importantes que influyen en la presencia anemia en niños de 6 a 35 meses de edad en Cusco?

¿Será posible predecir la anemia en niños de 6 a 35 meses de edad en Cusco mediante los diferentes algoritmos de Machine Learning?

1.3. Justificación de la investigación

La ausencia de precisión en el diagnóstico de la anemia en diversas ocasiones por procedimientos inadecuados y/o por desconocimiento de factores relacionados a esta enfermedad, pueden ser perjudiciales para los pacientes. De manera que, para el personal de salud poder predecir la anemia en los niños de 6 a 35 meses de edad en Cusco a partir de factores de riesgo comunes mediante un algoritmo de Machine Learning, le servirá para mejorar la toma de decisiones en cuanto a la prevención,

procedimientos diagnósticos o tratamientos eficaces, también, para formular políticas públicas en salud, así mismo, para la planificación de recursos necesarios.

1.4. Objetivos de la investigación

a. Objetivo general.

Determinar el algoritmo de Machine Learning más eficiente para predecir la anemia en niños de 6 a 35 meses de edad en Cusco.

b. Objetivos específicos.

Identificar las variables más importantes que influyen en la presencia anemia en niños de 6 a 35 meses de edad en Cusco.

Predecir la anemia en niños de 6 a 35 meses de edad en Cusco mediante los diferentes algoritmos de Machine Learning.

II. MARCO TEÓRICO CONCEPTUAL

2.1. Bases teóricas

Anemia

Es un trastorno en el cual el número de glóbulos rojos o eritrocitos circulantes en la sangre se ha reducido y es insuficiente para satisfacer las necesidades del organismo. En términos de salud pública, la anemia es define como una concentración de hemoglobina por debajo de dos desviaciones estándar del promedio según género, edad a nivel del mar.

Tabla 1

Concentración de hemoglobina y niveles de anemia en Niños, Adolescentes, Mujeres gestantes y Puérperas

Población	Con anemia según niveles de Hemoglobina (g/dL)			Sin anemia según niveles de Hemoglobina
Niños prematuros				
1ra semana de vida		≤ 13.0		> 13.0
2da a 4ta semana de vida		≤ 10.0		> 10.0
5ta a 8va semana de vida		≤ 8.0		> 8.0
Niños nacidos a término				
Menor de 2 meses		< 13.5		13.5 - 18.5
Niños de 2 a 6 meses cumplidos		< 9.5		9.5 - 13.5
	Severa	Moderada	Leve	
Niños de 6 meses a 5 años cumplidos	< 7.0	7.0 - 9.9	10.0 - 10.9	≥ 11.0
Niños de 5 a 11 años de edad	< 8.0	8.0 - 10.9	11.0 - 11.4	≥ 11.5
Adolescentes				
Adolescentes varones y mujeres de 12 a 14 años de edad	< 8.0	8.0 - 10.9	11.0 - 11.9	≥ 12.0
Varones de 15 años a más	< 8.0	8.0 - 10.9	11.0 - 12.9	≥ 13.0
Mujeres no gestantes de 15 años a más	< 8.0	8.0 - 10.9	11.0 - 11.9	≥ 12.0
Mujeres gestantes y puérperas				
Mujer gestante de 15 años a más	< 7.0	7.0 - 9.9	10.0 - 10.9	≥ 11.0
Mujer puérpera	< 8.0	8.0 - 10.9	11.0 - 11.9	≥ 12.0

Nota. De “Valores normales de Concentración de hemoglobina y niveles de anemia en Niños, Adolescentes, Mujeres gestantes y Puérperas (hasta 1000 msnm)” por Ministerio de Salud del Perú, 2017 (http://bvs.minsa.gob.pe/local/MINSA/4190.pdf).

Inteligencia artificial (IA)

“La rama de la ciencia de la computación que se ocupa de la automatización de la conducta inteligente” (Luger y Stubblefield, 1993).

“Un campo de estudio que se enfoca en la explicación y emulación de la conducta inteligente en función de procesos computacionales” (Schalkoff, 1990).

El campo de la inteligencia artificial (IA) abarca diversas áreas de estudio, siendo las más comunes y reconocidas las siguientes:

- Investigación de soluciones
- Desarrollo de sistemas expertos
- Procesamiento del lenguaje natural
- Robótica
- Aprendizaje de maquinas
- Lógica
- Incertidumbre y “lógica difusa”

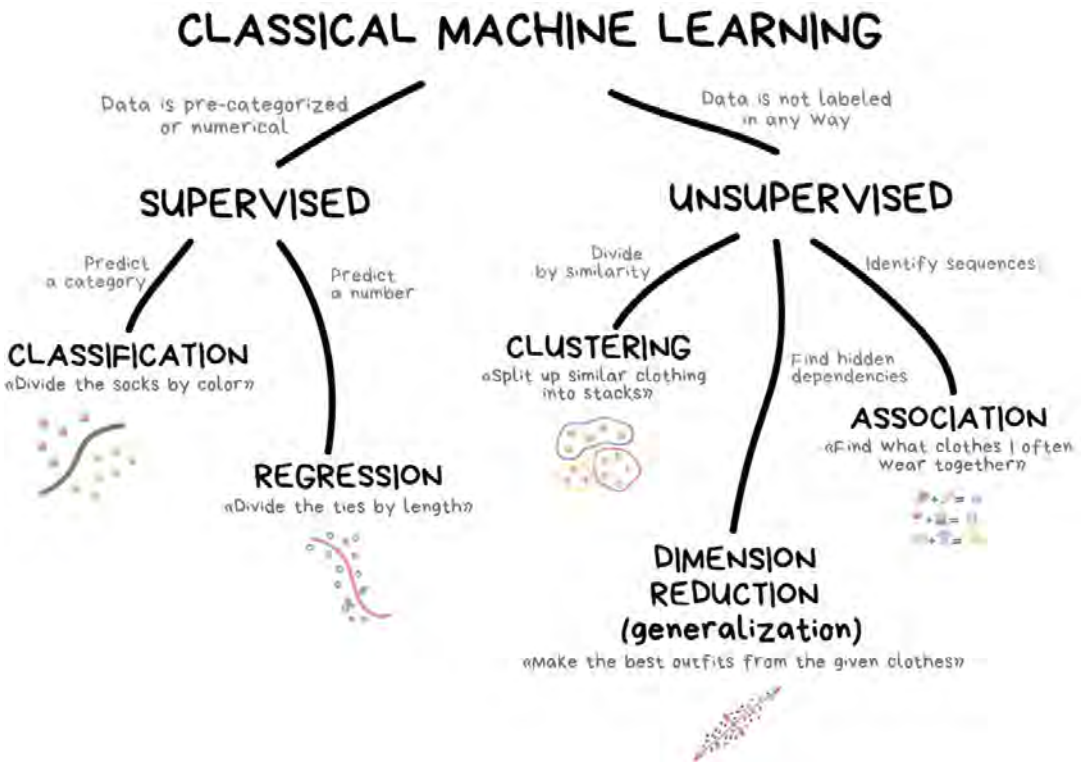
Aprendizaje de máquinas (Machine Learning)

El aprendizaje automático o aprendizaje automatizado o aprendizaje de máquinas (del inglés, Machine Learning). Lewis (2017) afirma que es una colección de algoritmos que generan información a partir de los datos. Esa información podría ser utilizada por humanos u otras máquinas para tomar una decisión.

El aprendizaje automático clásico a menudo se divide en dos categorías: aprendizaje supervisado y no supervisados.

Figura 1

Aprendizaje supervisado y no supervisado



Nota. De Machine Learning para todos, de forma simple e com exemplos!, por P. C. Tebaldi, 2019, DataGeeks (https://www.datageeks.com.br/machine-learning/).

- **Aprendizaje supervisado**

  La máquina tiene un supervisor o un maestro que le proporciona todas las respuestas, como si es un gato en la imagen o un perro. El maestro ya ha dividido (etiquetado) los datos en perros y gatos, y la máquina está utilizando estos ejemplos para aprender. Claramente, la máquina aprenderá más rápido con un maestro, por lo que se usa más comúnmente en tareas de la vida real. Hay dos tipos de tareas de este tipo: clasificación (predicción de la categoría de un objeto) y regresión (predicción de un punto especifico en un eje numérico).

Clasificación: siempre necesitas un maestro. Los datos deben etiquetarse con características para que la máquina pueda asignar las clases en función de ellas. Todo se puede clasificar: usuarios según intereses, artículos según idioma y tema (que es importante para los motores de búsqueda), música según género (listas de reproducción de Spotify) e incluso sus correos electrónicos. Algunos algoritmos populares: Naive Bayes, árbol de decisión, regresión logística, k-vecinos más cercanos y máquina de vectores de soporte.

Regresión: es básicamente una clasificación en la que pronosticamos un número en lugar en una categoría. Algunos ejemplos son el precio del coche por kilometraje, el tráfico por hora del día, volumen de demanda por crecimiento de la empresa, etc. la regresión es perfecta cuando algo depende del tiempo. Los algoritmos populares son las regresiones lineales y polinómicas.

- **Aprendizaje no supervisado**

  La tecnología no supervisada se inventó un poco más tarde, en los años 90. Se usa con menos frecuencia, pero a veces simplemente no tenemos otra opción.

  Agrupación: es una clasificación sin clases predefinidas. Los algoritmos de agrupación que intenta encontrar objectos similares (por algunas características) y fusionarlos en un grupo. Aquellos que tienen muchas características similares se unen en una clase. Con algunos algoritmos, incluso puedes especificar el número exacto de clústeres que deseas.

  Reducción de dimensionalidad: estos métodos eran utilizados por científicos de datos incondicionalmente, que tenían que encontrar “algo

interesante” entre enorme cantidad de números. Cuando los gráficos de Excel no ayudaron, obligaron a las maquinas a buscar patrones. Así es como obtuvieron los métodos de Reducción de Dimensiones.

Aprendizaje de reglas de asociación: incluye todos los métodos para analizar carritos de compras, automatizar estrategias de marketing y otras tareas relacionadas con eventos. Cuando tengas una secuencia de algo y quieres encontrar patrones en ella, los algoritmos como Apriori, Euclat, FP-crecimiento son adecuados para analizar estos patrones.

Algoritmos de Machine Learning

Regresión logística

Aunque el término regresión logística se utiliza comúnmente, se le conoce también como discriminación logística, ya que su principal aplicación es la resolución de problemas de clasificación. Este enfoque se utiliza para estimar probabilidades en relación a variables categóricas de respuesta, convirtiendo a la regresión logística en uno de los modelos lineales generalizados más comunes. Este modelo se emplea para modelar la probabilidad de eventos como la compra de un producto entre varias opciones, la probabilidad de fraude o la probabilidad de morosidad. En estos casos, la variable de respuesta puede tener dos o más posibilidades, cada una con su respectiva probabilidad, y la suma de estas probabilidades siempre equivale a uno.

Sea y_i una variable respuesta binaria:

$$y_i = \begin{cases} 1, & Prob_i(1) = p_i, \\ 0, & Prob_i(0) = 1 - p_i. \end{cases}$$

La variable de respuesta sigue pues una distribución binomial de parámetros 1 y p_i, donde p_i es probabilidad de que un individuo responda con 1 puede

variar de un individuo a otro. Observamos que la parte sistemática del modelo es precisamente esta probabilidad:

$$\mu_i = E[y_i] = 1\ x\ p_i + 0\ x\ (1 - p_i) = p_i$$

La función de unión r^{-1} debe transformar el predictor lineal en un valor dentro del rango [0,1]. La función más comúnmente utilizada, aunque no la única, es la función logística:

$$p_i = \frac{1}{1 + e^{-\beta' x_i}}$$

Esta dependencia también se escribe como:

$$log\frac{p_i}{1 - p_i} = \beta' x_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

Esto es, el logit, definido como $log(p_i/1 - p_i)$, es igual al predictor lineal clásico.

Arboles de decisión

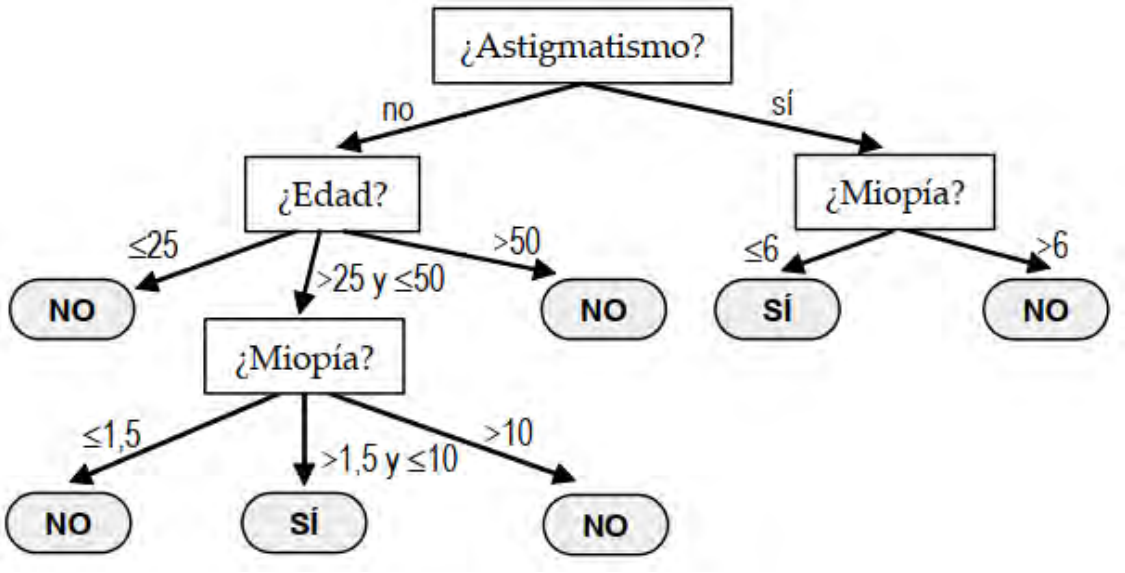
Entre todos los métodos de aprendizaje, los sistemas basados en árboles de decisión son posiblemente los más simples de utilizar y comprender. Un árbol de decisión consiste en un conjunto de condiciones organizadas de manera jerárquica, de modo que la decisión final se puede determinar siguiendo las condiciones que se cumplen desde la raíz del árbol hasta alguna de sus hojas. Los árboles de decisión han sido utilizados durante siglos y son especialmente adecuados para expresar procedimientos en campos como la medicina, el derecho, los negocios, la estrategia, las matemáticas, la lógica, entre otros.

Por ejemplo, consideremos un hospital público que realiza cirugías refractivas (LASIK) para personas miopes que lo solicitan. Obviamente, no todas las personas son candidatas para esta cirugía, y algunos casos

pueden ser excluidos en una primera etapa para evitar riesgos o efectos secundarios potenciales. Aunque la decisión de realizar o no la cirugía requiere una evaluación minuciosa por parte del servicio de oftalmología del hospital, existen ciertas condiciones claras que pueden determinar si una persona es, en principio, adecuada para un estudio detallado (incluyendo la medición de la tensión ocular y la paquimetría) y, en última instancia, para la cirugía. En la figura 2 se presenta un diagrama de un árbol de decisión que se utiliza para evaluar las solicitudes.

Figura 2

Árbol de decisión para establecer recomendación de la cirugía ocular



Nota. De Introducción a la Minería de Datos, por J. H. Orallo, J. R. Quintana, y C. F. Ramírez, 2004, Pearson Educación S.A.

La figura 2 muestra que aplicar un árbol de decisión a un nuevo paciente con el fin de determinar si se debe recomendar una operación es un proceso sencillo. Simplemente se hacen preguntas y se siguen las respuestas hasta llegar a una de las hojas del árbol, etiquetadas como "no" o "sí". Este árbol

funciona como un "clasificador", es decir, clasifica a un nuevo individuo en una de dos posibles categorías: "no" o "sí".

Redes neuronales artificiales

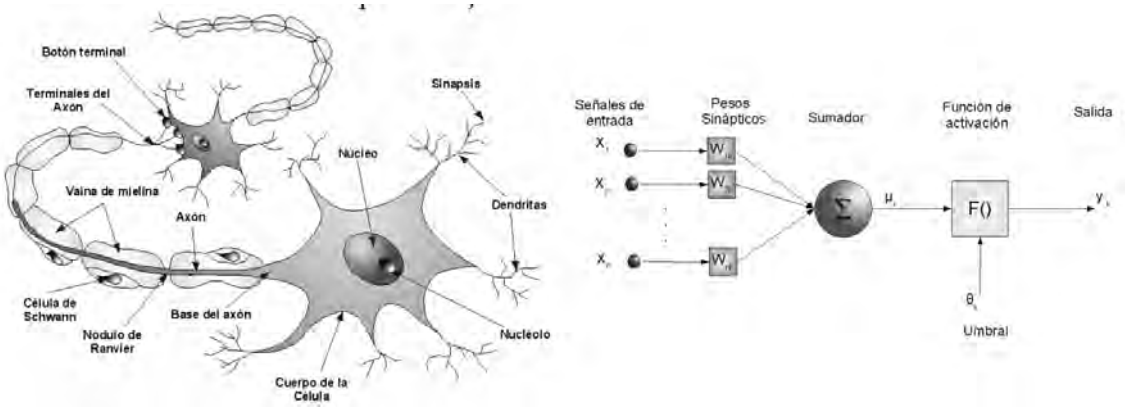
Las redes neuronales artificiales (RNA) son un método de aprendizaje diseñado originalmente para emular los procesadores de información biológicos. Estas redes parten del supuesto de que la capacidad humana para procesar información se debe a la naturaleza biológica de nuestro cerebro. Por lo tanto, para imitar esta característica, debemos estudiar y basarnos en el uso de soportes artificiales similares a los presentes en nuestro cerebro.

Neuronas biológicas y artificiales

En la figura 3, se muestra una neurona simple en la parte izquierda. Esta neurona recibe información a través de las sinapsis en sus dendritas, que representan la conexión entre un axón de otra neurona y una dendrita de la neurona representada en la figura. Una transmisión electroquímica ocurre en la sinapsis, permitiendo que la información se transmita de una neurona a otra. La información viaja a lo largo de las dendritas hasta llegar al cuerpo de la célula, donde se realiza la suma de los impulsos eléctricos recibidos, seguida de la aplicación de una función de activación. La neurona se activa si el resultado supera un cierto límite o umbral. En ese caso, envía una señal en forma de una onda de ionización a lo largo de su axón para comunicarse con otras neuronas, permitiendo la transferencia de información en la red. Es importante notar que las sinapsis tienen diferentes rendimientos que cambian con el tiempo de vida de la neurona.

Figura 3

Neurona biológica y artificial



Nota. De Introducción a la Minería de Datos, por J. H. Orallo, J. R. Quintana, y C. F. Ramírez, 2004, Pearson Educación S.A.

En la parte derecha de la figura 3 se representa una neurona artificial, que incluye una entrada adicional externa llamada "polarización" o "bias," denotada como θ_k. Esta entrada ajusta el umbral de excitación de la neurona, aumentándolo o disminuyéndolo dependiendo de si su valor es positivo o negativo.

Las entradas se representan mediante el vector de entrada, x, y el rendimiento de las sinapsis se modela a través de un vector de pesos, w. Por lo tanto, el valor de salida de esta neurona se determina mediante una función:

$$y = f\left(\sum_i w_i\, x_i\right) = f(w.x) = f(w^T x)$$

donde f es la función de activación.

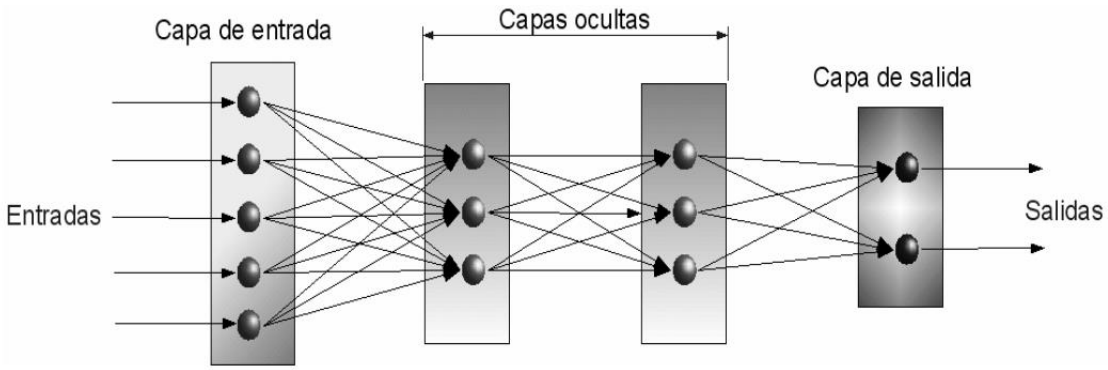
Cuando se tiene una red de neuronas, las salidas de unas se conectan a las entradas de otras. Si el peso entre dos neuronas conectadas es positivo,

provoca un efecto de excitación, mientras que, si es negativo, se produce un efecto de inhibición.

En consecuencia, una única neurona se considera una unidad de procesamiento muy básica, y el potencial de las redes neuronales artificiales radica en la capacidad que ofrecen al emplear muchas de estas unidades sencillas y resistentes en funcionamiento simultáneo. Normalmente imaginamos las neuronas actuando conjuntamente en capas como se muestra en la siguiente figura:

Figura 4

Red neuronal artificial formada por cuatro capas de neuronas



Nota. De Introducción a la Minería de Datos, por J. H. Orallo, J. R. Quintana, y C. F. Ramírez, 2004, Pearson Educación S.A.

En esta ilustración se puede observar un grupo de entradas (representado por el vector de entrada, x) ingresando a la red desde el extremo izquierdo y avanzando a través de la red hasta que la activación llega a la capa de salida. Las capas intermedias se denominan capas ocultas, ya que no son visibles desde fuera de la red.

Hay dos formas de trabajo de una Red Neuronal Artificial:

- Método de transferencia de la activación: esto ocurre cuando la activación se propaga a lo largo de toda la red. Este método de operación

o aplicación está asociado con la operación de propagación hacia adelante.

- Modo de aprendizaje: Este enfoque de aprendizaje se basa en la organización natural de la red a partir de la transferencia de la activación más reciente.

Aprendizaje supervisado en Red Neuronal Artificial

En este tipo de aprendizaje, se suministra a la red un conjunto de datos de entrada junto con la respuesta correcta deseada. Los datos de entrada se propagan a través de la red hasta que la activación llega a las neuronas de la capa de salida. Luego, se compara la respuesta calculada por la red con la respuesta deseada, es decir, el valor objetivo o "blanco" (en inglés, target). Luego, se ajustan los pesos de la red para asegurar que, en futuras ocasiones en las que se presente un patrón de entrada similar, la red sea más propensa a producir una respuesta correcta. Este tipo de aprendizaje resulta particularmente útil en tareas de regresión y clasificación.

Perceptrón multicapa

Tanto el Perceptrón simple y el modelo Adaline son métodos potentes de aprendizaje, aunque hay algunas situaciones en las que no dan lugar buenos resultados. Estos casos se caracterizan por ser no linealmente separables. Hoy en día es posible mostrar que muchos conjuntos de datos que no son linealmente separables pueden ser modelados mediante el empleo del Perceptrón Multicapa (Multilayer Perceptron, MLP), es decir una red neuronal en forma de cascada, que tiene una o más capas ocultas, como las vista en la figura 4.

Algoritmo de retropropagación

Debido a su importancia, se muestra el algoritmo de forma detalla en seguida:

1. Inicializar los pesos a valores pequeños aleatorios.
2. Escoger un patrón de entrada, X^p, y presentarlo a la capa de entrada.
3. Propagar la activación hacia adelante a través de los pesos hasta que la activación alcance las neuronas de la capa de salida.
4. Calcular los valores de "δ" para las capas de salida $\delta_j^P = \left(t_j^P - o_j^P\right) f'\left(Act_j^P\right)$ usando los valores de los blancos deseados para el patrón de entrada seleccionado.
5. Calcular los valores de "δ" para la capa oculta usando $\delta_i^P = \sum_{j=1}^{N} \delta_j^P w_{ji} f'(Act_i^P)$.
6. Actualizar los pesos de acuerdo con: $\Delta_P w_{ij} = \gamma \delta_i^P o_j^P$.
7. Repetir del paso 2 al 6 para todos los patrones de entrada.

Aquí, $f'(Act_j^p)$ representa la derivada de la función de activación.

Para definir completamente el algoritmo anterior, es necesario especificar diversos factores que influyen en él, como la forma de proporcionar ejemplos (ya sea en forma de lotes o en línea), la inicialización de los pesos iniciales, las funciones de activación y el valor del coeficiente de aprendizaje γ, entre otros.

Máquinas de vectores de soporte

Las Máquinas de Vectores de Soporte, también conocidas como SVM por sus siglas en inglés (Support Vector Machines), son clasificadores lineales que generan separadores lineales o hiperplanos en espacios de características de alta dimensionalidad, logrando esto mediante funciones

núcleo. Estos modelos tienen una inclinación inductiva especial que implica maximizar el margen.

En primer lugar, es importante recordar que cualquier hiperplano en un espacio D-dimensional, $\mathbb{R}^D$, puede representarse como $h(x) = \langle w, x \rangle + b$, donde $w \in \mathbb{R}^D$ es el vector ortogonal al hiperplano, $b \in \mathbb{R}$ y $\langle ., . \rangle$ denota el producto escalar convencional en $\mathbb{R}^D$. En el contexto de la clasificación binaria, la regla de clasificación se expresa como: $f(x) = signo(h(x))$, donde la función signo se define como:

$$signo(x) = \begin{cases} +1, & si\ x \geq 0 \\ -1, & si\ x < 0 \end{cases}$$

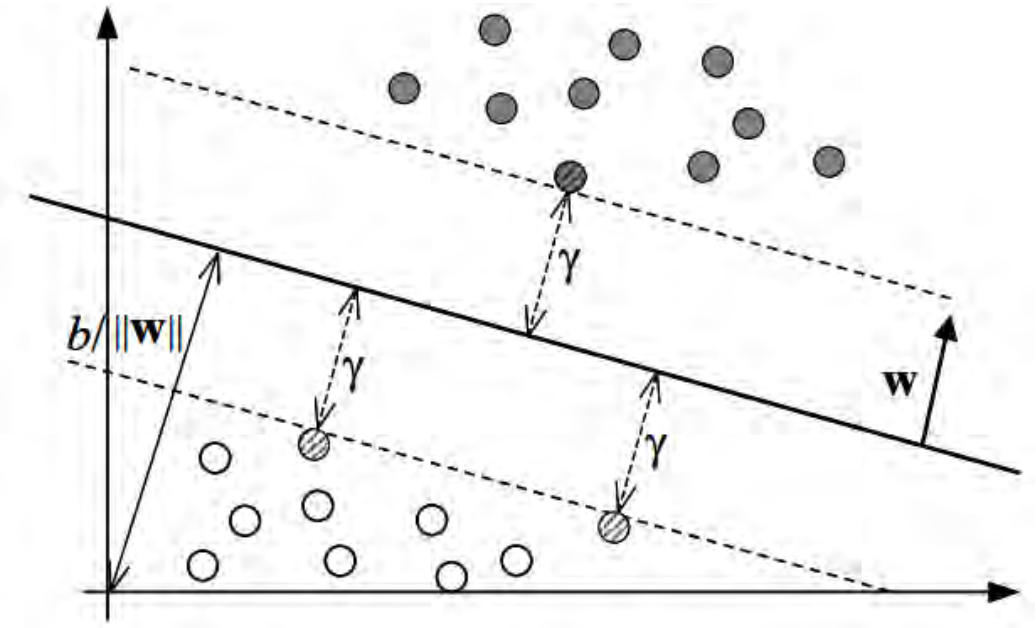
En el ámbito de la clasificación, los vectores $x \in \mathbb{R}^D$ son representaciones vectoriales con una componente real para cada atributo, y el vector w suele ser llamado el vector de pesos, indicando la importancia o contribución de cada atributo en la regla de clasificación. Por último, b se conoce como sesgo (bias) y establece el umbral de decisión. Cuando se trata de un conjunto de datos binario linealmente separable, existen varios algoritmos incrementales para construir hiperplanos (w, b) que logren clasificarlos con precisión, como el Perceptrón, Widrow-Hoff, Winnow, Exponentiated-Gradient y Sleeping Experts. Aunque todos estos algoritmos convergen hacia una solución, las particularidades de cada uno pueden llevar a soluciones ligeramente diferentes debido a la existencia de múltiples hiperplanos separadores posibles.

En el caso de datos linealmente separables, surge la pregunta de cuál es el hiperplano separador "óptimo" en términos de generalización. Las SVM de margen máximo buscan seleccionar el hiperplano que se encuentra a igual distancia de los puntos más cercanos de ambas clases. En otras palabras,

este hiperplano maximiza la distancia mínima (margen geométrico) entre el conjunto de datos y el hiperplano. Intuitivamente, se sitúa en una posición neutral con respecto a las clases más numerosas y solo considera los puntos en las fronteras de la región de decisión, conocidos como vectores soporte.

Figura 5

Hiperplano (w, b) *equidistante a dos clases, margen geométrico* (γ) *y vectores soporte (puntos rayados)*



Nota. De Introducción a la Minería de Datos, por J. H. Orallo, J. R. Quintana, y C. F. Ramírez, 2004, Pearson Educación S.A.

En la figura 5 se presenta geométricamente este hiperplano equidistante (o de margen máximo) para el caso bidimensional.

Desde una perspectiva algorítmica, el aprendizaje de las SVM es un problema de optimización con restricciones que se resuelve mediante técnicas de programación cuadrática (QP). La convexidad garantiza una solución única, lo que constituye una ventaja sobre los modelos tradicionales de redes neuronales. Además, las implementaciones actuales permiten una

eficiencia razonable al abordar problemas reales con miles de datos y atributos.

Para el aprendizaje de separadores no lineales, las SVM utilizan una transformación no lineal del espacio de atributos de entrada en un espacio de características de mayor dimensionalidad, donde es posible lograr una separación lineal. Esto se logra mediante el uso de funciones núcleo (kernel functions), que calculan el producto escalar de dos vectores en el espacio de características. Este enfoque permite trabajar de manera eficiente en el espacio de características sin necesidad de realizar explícitamente las transformaciones de los datos de entrenamiento. Es importante destacar que el uso de funciones núcleo no se limita exclusivamente a las SVM, ya que muchos otros algoritmos pueden "kernelizarse" para aprender funciones no lineales, incluyendo el perceptrón, los discriminantes de Fisher, el análisis de componentes principales, y otros.

Naive Bayes

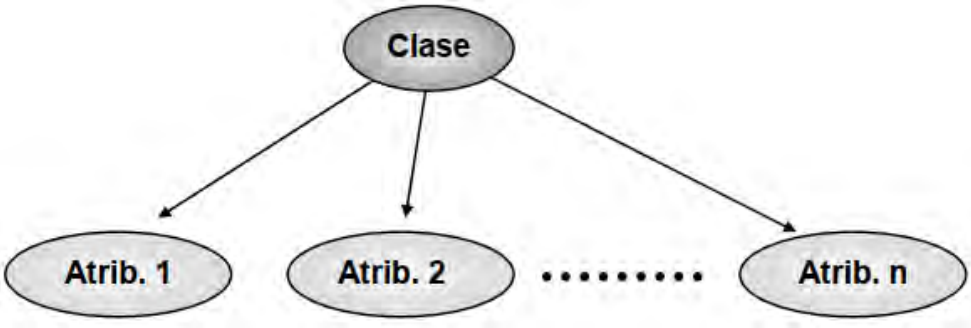
Indudablemente, este es el enfoque de clasificación más sencillo que se utiliza en las redes bayesianas. En este escenario, la estructura de la red permanece constante y solo se requiere aprender los parámetros (las probabilidades). El fundamento fundamental del clasificador Naive Bayes radica en la suposición de que todos los atributos son independientes dado el valor de la variable clase. Aunque esta suposición es bastante fuerte y poco realista en la mayoría de los casos, el clasificador Naive Bayes es ampliamente utilizado. Además, diversos estudios indican que sus resultados son competitivos con otras técnicas, como redes neuronales y árboles de decisión, en muchos problemas, e incluso los superan en algunos

casos. Un ejemplo de un problema en el que el clasificador Naive Bayes demuestra ser altamente eficaz es la detección de correo no deseado o spam.

La suposición de independencia que el clasificador NB asume da como resultado un modelo gráfico probabilístico en el cual hay un único nodo raíz (la clase), y todos los atributos son nodos hojas que tienen a la variable clase como único padre. Visualmente, esta estructura se representa de la siguiente manera:

Figura 6

Topología de un clasificador Naive Bayes



Nota. De Introducción a la Minería de Datos, por J. H. Orallo, J. R. Quintana, y C. F. Ramírez, 2004, Pearson Educación S.A.

Selección de predictores

Cuando se lleva a cabo el entrenamiento de un modelo, resulta fundamental incorporar como variables predictoras únicamente aquellas que guardan una verdadera relación con la variable de respuesta, ya que son estas las que aportan información valiosa al modelo. La inclusión de un exceso de variables suele llevar consigo una disminución en la capacidad predictiva del modelo al enfrentarse a nuevos datos (sobreajuste). Algunos algoritmos de Machine Learning, como los bosques aleatorios, lasso y boosting,

incorporan estrategias propias para la selección de variables predictoras, lo que los hace modelos altamente versátiles. Aparte de estos, existen dos categorías principales de métodos para reducir el número de variables predictoras antes de ajustar el modelo: los métodos wrapper y los métodos de filtrado.

- **Métodos wrapper**

  Los métodos de wrapper examinan diversos modelos creados mediante la inclusión o exclusión de predictores, con el objetivo de encontrar la combinación más efectiva que maximice la capacidad del modelo. Estos métodos se pueden concebir como algoritmos de búsqueda que consideran los predictores disponibles como variables de entrada y emplean una métrica del modelo, como su error de predicción, como el objetivo de optimización.

- **Métodos de filtrado**

  Los enfoques de filtrado en la evaluación de modelos analizan la pertinencia de los predictores fuera del modelo, seleccionando solo aquellos que cumplen con criterios específicos. Este método implica examinar la relación de cada predictor con la variable respuesta. Por ejemplo, en situaciones de clasificación con predictores continuos, se puede aplicar un análisis de varianza (ANOVA) a cada predictor para identificar aquellos que exhiben variaciones en función de la variable respuesta. Los predictores con valores de p inferiores a un límite predeterminado y ubicados entre los n mejores se incorporan al modelo. Para mitigar el riesgo de sobreajuste debido a la influencia excesiva de

los datos de entrenamiento, se recomienda repetir el proceso mediante validación cruzada o bootstrapping.

Minería de datos

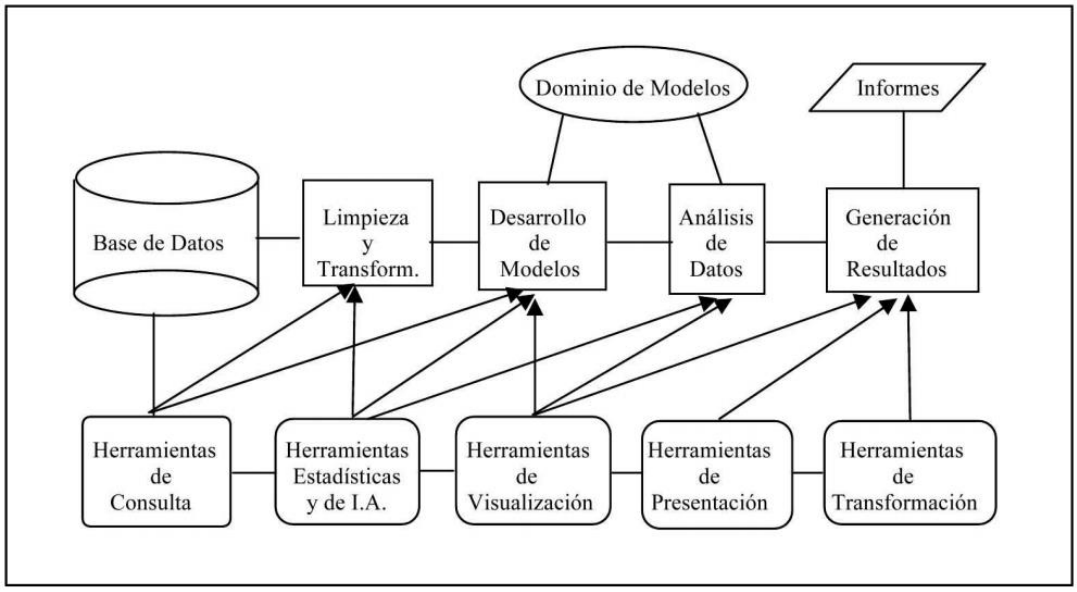
"La minería de datos puede definirse inicialmente como un proceso de descubrimiento de nuevas y significativas relaciones, patrones y tendencias al examinar grandes cantidades de datos" (Pérez lópez y Santín González, 2007, p.1).

El proceso de extracción del conocimiento

La minería de datos es solo una parte del proceso de obtener conocimiento a partir de datos según el enfoque KDD (Knowledge Discovery in Databases). Este procedimiento incluye múltiples etapas, como la preparación de datos (que implica seleccionar, limpiar y transformar los datos), la exploración y auditoría de los mismos, la minería de datos propiamente dicha (donde se desarrollan modelos y se analizan los datos), la evaluación de resultados, la difusión y la utilización de modelos. Además, este proceso de obtención de conocimiento hace uso de diversas técnicas de diferentes campos, como árboles de decisión, regresión lineal, redes neuronales artificiales, técnicas bayesianas y máquinas de soporte vectorial, y aborda una variedad de problemas, como clasificación, categorización, estimación/regresión y agrupamiento. La figura 7 muestra las etapas del KDD (Knowledge Discovery in Databases).

Figura 7

Fases del KDD (Knowledge Discovery in Databases)



Nota. De Introducción a la Minería de Datos, por J. H. Orallo, J. R. Quintana, y C. F. Ramírez, 2004, Pearson Educación S.A.

El proceso de KDD comienza con la recopilación e integración de datos de diversas fuentes, tanto internas como externas, con el objetivo de obtener información valiosa en un dominio específico de la organización. Esto implica generalmente utilizar bases de datos y otras fuentes de información, como demografía, páginas amarillas, perfiles por zonas, uso de Internet y datos adquiridos de otras organizaciones. La disponibilidad de grandes cantidades de información en esta etapa puede requerir el uso de técnicas de muestreo para seleccionar datos.

La siguiente fase de KDD incluye la exploración, limpieza y transformación de datos. En esta etapa, se eliminan datos incorrectos o irrelevantes, utilizando herramientas de consulta y estadísticas. La exploración de datos implica el uso de técnicas de análisis exploratorio, como histogramas y diagramas de caja, para detectar datos atípicos o faltantes. La presencia de

datos anómalos o faltantes puede llevar al uso de algoritmos resistentes a estos problemas, filtrado de información, imputación de valores faltantes y la transformación de datos continuos en discretos.

La siguiente fase del proceso implica la minería de datos real, que se realiza a través del desarrollo de modelos predictivos y descriptivos y el análisis de datos. La elección de la técnica de minería de datos depende del tipo de conocimiento que se busca extraer.

Posteriormente, se requiere una fase para seleccionar y validar los modelos utilizando criterios de evaluación de hipótesis. La implementación o interpretación del modelo puede ser necesaria en esta etapa, y se utilizan herramientas estadísticas y de visualización.

Finalmente, la última etapa del proceso implica la difusión y el uso del conocimiento obtenido a través de las técnicas de minería de datos, lo que generalmente se traduce en la generación de resultados. Los modelos pueden tener múltiples usuarios y, por lo tanto, necesitar difusión, lo que puede requerir que se expresen de manera comprensible para su distribución dentro de la organización. Se utilizan herramientas de visualización, presentación y transformación de datos en esta fase.

Matriz de confusión

La matriz de confusión presenta en una tabla una visión grafica de los errores cometidos por el modelo de clasificación. Se trata de un modelo grafico para visualizar el nivel de aciertos de un modelo de predicción. También es conocido en la literatura como tabla de contingencia o matriz de errores.

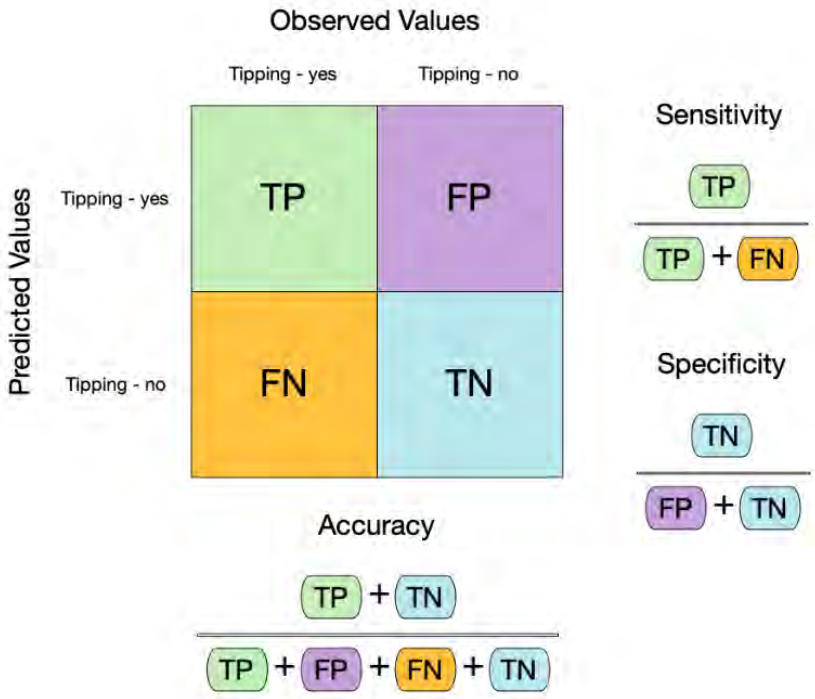
En resumen, la matriz de confusión refleja la cantidad de aciertos y errores en la clasificación. Los parámetros que nos proporciona son los siguientes:

- Verdadero positivo (TP): representa la cantidad de aciertos al clasificar en la clase positiva (P).
- Verdadero negativo (TN): indica la cantidad de aciertos al clasificar en la clase negativa (N).
- Falso negativo (FN): refleja la cantidad de errores al clasificar como negativa una instancia que pertenece a la clase positiva.
- Falso positivo (FP): muestra la cantidad de errores al clasificar como positiva una instancia que pertenece a la clase negativa.

En la figura 8 nos muestra el procedimiento para obtener los diferentes parámetros de la matriz de confusión.

Figura 8

Matriz de confusión



Nota. Adaptado de Machine learning with tidymodels, por LatinR, 2023 (https://workshops.tidymodels.org/).

2.2. Marco conceptual

Anemia. Es una situación en la que la sangre no cuenta con la cantidad adecuada de glóbulos rojos o cuando la concentración de hemoglobina se encuentra por debajo de los niveles de referencia para la edad, el género y la altitud.

Hemoglobina. Es un conjunto de proteínas que contiene hierro y se produce en los glóbulos rojos del organismo humano. Su insuficiencia sugiere, en esencia, una carencia de hierro.

El Programa Nacional Cuna Más. Es un programa social que tiene como objetivo mejorar el desarrollo infantil de niñas y niños menores de 36 meses de edad, en localidades en situación de pobreza y pobreza extrema.

Perceptrón simple. Su estructura tiene varios nodos o neuronas de entrada y uno o más de salida, no tiene capa oculta o intermedia.

Adaline. Una red neuronal artificial con topología idéntica al perceptrón simple, la diferencia entre esta red y el Perceptrón es la presencia o no de un umbral.

2.3. Antecedentes empíricos de la investigación

Antecedentes Internacionales

Rahman Khan, Chowdhury, Islam y Raheem (2019) en su artículo refieren que la utilidad de conocer la probabilidad de anemia en niños menores a cinco años dados factores de riesgo comunes a través de algoritmos de aprendizaje es clave para la formulación de políticas de servicio de salud y de la comunidad. Los datos se extrajeron de la Encuesta Demográfica y de Salud de Bangladesh (BDHS) realizada en 2011. Se encontró que el algoritmo de bosques aleatorios (random forest) alcanzó la mejor precisión

de clasificación, que llegó al 68.53%. Además, demostró una sensibilidad del 70.73%, una especificidad del 66.41% y un AUC de 0.6857. Como resultado, se puede concluir que cuando se trata de predecir la anemia, los métodos de Machine Learning son una opción que debe ser considerada.

En el artículo de **Sow, Mukhtar, Ahmad y Suguri (2019)**, titulado: "Assessing the relative importance of social determinants of health in malaria and anemia classification based on Machine Learning techniques", mencionan que en el aprendizaje supervisado para predecir el resultado de una enfermedad especifica, los factores sociales decisivos de la salud han sido muy poco explorados y prometen ser predictores significativos de problemas de salud pública como la malaria y anemia entre los niños. En este artículo se consideró estudiar el poder de contribución en las predicciones de la malaria y anemia de estas determinantes sociales aplicando los algoritmos de redes neuronales artificiales, K nearest neighbors (KNN), Random forests (RF), Support vector machine (SVM) y Naive bayes (NB) para clasificar ambas enfermedades. De todos ellos, las redes neuronales artificiales consiguieron los mejores resultados con un 94.74% y un 84.17% de precisión en la predicción de malaria y anemia respectivamente, resultados que fueron consistentes y que reflejan la importancia de los factores no médicos en la predicción de enfermedades.

Jaiswal, Srivastava y Siddiqui (2019) en su estudio destacan que, en la ciencia médica la predicción de enfermedades en el momento apropiado es el problema central de los profesionales para la prevención y el programa de tratamiento eficaz. En su estudio se utilizó algoritmos de aprendizaje supervisado como Naive Bayes, Bosque aleatorio y Árbol de decisión para

la predicción de la anemia utilizando datos referentes a la CBC (complete blood count) recopilados de centros patológicos. Los resultados manifestaron que la técnica Naive-Bayes fue superior en precisión en comparación a C4.5 y Bosque aleatorio.

En su investigación **Abdullah y Al-Asmari (2016)** indican que la anemia es una de las enfermedades hematológicas más habituales. En su estudio se centran en cinco tipos más comunes de anemia y a través de algoritmos de clasificación especifica el tipo de anemia para los pacientes anémicos. Los datos fueron construidos a partir de los resultados de la prueba de hemogramas completo (CBC) de los pacientes. Se utilizo la herramienta de minería de datos WEKA (Waikato Environment for Knowledge Analysis), luego de varias pruebas demostraron que el algoritmo de árbol de decisiones J48 mostro la mejor clasificación posible de tipos de anemia.

Antecedentes Nacionales

El objetivo principal de la investigación de **Canaza Espezua (2021)** se centró en la creación de un modelo predictivo para evaluar el riesgo asociado a la anemia en niños menores de 5 años en la Microred Yauri, ubicada en la provincia de Espinar, en Cusco. Los datos se obtuvieron del Sistema de Información del Estado Nutricional (SIEN) correspondiente a la provincia de Espinar. Se utilizo un modelo de regresión logística, cuya tasa de clasificación de verdaderos negativos (no anémicos predichos como no anémicos) fue del 93.8% y la clasificación correcta de los niños anémicos (verdaderos positivos) fue de 96.2%, exponiendo el modelo una gran capacidad predictiva.

III. HIPÓTESIS Y VARIABLES

3.1. Hipótesis general

El algoritmo de Machine Learning más eficiente para predecir la anemia en niños de 6 a 35 meses de edad en Cusco es el de bosques aleatorios.

3.2. Operacionalización de variables

Por la cantidad de variables encontradas relacionadas a los niños de 6 a 35 meses de edad de la Encuesta Demográfica y de Salud Familiar (ENDES), la mayoría de las variables para nuestra investigación fueron consideradas según los antecedentes de investigación y de acuerdo al contexto del estudio.

Tabla 2

Operacionalización de variables

Variables	Definición conceptual	Indicadores	Valor final	Tipo de variable
Variable endógena				
Anemia	Anemia de cada niño(a) menor de 6 años	Concentración de Hemoglobina	Con anemia, Sin anemia	Categórica nominal
Variable exógena				
Número de personas en el hogar	Es el número total de miembros del hogar	Número de personas en el hogar	1, 2, 3, …	Numérica discreta
Número de habitaciones	Permite conocer la cantidad de habitaciones que se usan en el hogar para dormir	Número de habitaciones	1, 2, 3, …	Numérica discreta
Números de niños menor a 5 años	Permite conocer el número de niños menores de 5 años	Números de niños menor a 5 años	1, 2, 3, …	Numérica discreta
Altitud de residencia	Altitud de residencia de los niños	Altitud de residencia	m.s.n.m.	Numérica continua
Nivel de Hemoglobina	Determinar el nivel de hemoglobina (G/DL) de cada niño(a) menor de 6 años de edad	Nivel de Hemoglobina	g/dl	Numérica continua
Edad del niño(a)	Permite conocer la edad en meses de los niños menores de 6 años	Edad del niño(a)	6m, 7m, …, 35m	Numérica discreta
Talla del niño(a)	Permite conocer la medición antropométrica (talla) de cada niño(a) menor de 6 años de edad	Talla del niño(a)	cm.	Numérica continua
Peso del niño(a)	Permite conocer la medición antropométrica (peso) de cada niño(a) menor de 6 años de edad	Peso del niño(a)	kg.	Numérica continua
Sexo del niño(a)	Permite conocer el sexo de cada niño(a) menor de 6 años de edad	Sexo del niño(a)	Mujer, Hombre	Categórica nominal
Número de orden de nacimiento	Es el número de orden al nacer de cada niño(a) menor de 6 años	Número de orden de nacimiento	1, 2, 3, …	Numérica discreta
Edad de la madre	Permite conocer la edad de cada mujer elegible de 12 a 49 años	Edad de la madre	años	Numérica discreta
Talla de la madre	Permite conocer la medición antropométrica (peso) de cada mujer de 12 a 49 años de edad	Talla de la madre	cm.	Numérica continua
Peso de la madre	Permite conocer la medición antropométrica (peso) de cada mujer de 12 a 49 años de edad	Peso de la madre	kg.	Numérica continua
Índice de riqueza	Índice que nos permite conocer la riqueza con que cuenta el hogar	Índice de riqueza	Los más pobres, Pobre, Medio, Rico, Más rico	Categórica ordinal
Área de residencia	Refiere al área geográfica de residencia donde se encuentra la vivienda	Área de residencia	Urbano, Rural	Categórica nominal
Electricidad en el hogar	Permite conocer si el hogar del informante se alumbra con un generador de electricidad	Electricidad en el hogar	Si, No	Categórica nominal
Refrigeradora o congeladora en el hogar	El propósito de la variable es saber si el hogar posee refrigeradora	Refrigeradora o congeladora en el hogar	Si, No	Categórica nominal

Variables	Definición conceptual	Indicadores	Valor final	Tipo de variable
El agua que se bebe o toma es hervida	Refiere al método utilizado por la persona entrevistada para hacer bebible el agua que consumen	El agua que se bebe o toma es hervida	Si. No	Categórica nominal
Televisión por cable en el hogar	Permite conocer si el hogar del informante tiene televisión por cable	Televisión por cable en el hogar	Si, No	Categórica nominal
Internet en casa	Permite conocer si el hogar del informante tiene acceso a Internet en casa	Internet en casa	Si, No	Categórica nominal
Afiliado al seguro de salud de EsSalud	Permite conocer si la persona que está incluida en la lista de miembros del hogar, también está afiliado en el seguro de salud de EsSalud	Afiliado al seguro de salud de EsSalud	Si, No	Categórica nominal
Afiliado al Seguro Integral de Salud (SIS)	Permite conocer si la persona que está incluida en la lista de miembros del hogar, también está afiliado o inscrito al Seguro Integral de Salud (SIS)	Afiliado al Seguro Integral de Salud (SIS)	Si, No	Categórica nominal
Lugar de residencia	Refiere al lugar de residencia donde se encuentra la vivienda entrevistada	Lugar de residencia	Capital gran ciudad o pequeña ciudad, Pueblo, Campo	Categórica ordinal
Fuente principal de abastecimiento de agua para tomar o beber	Permite evaluar la calidad y limpieza del agua que utilizan en el hogar para tomar o beber	Fuente principal de abastecimiento de agua para tomar o beber	Dentro de la vivienda, Fuera de la vivienda pero dentro del edificio, Otros	Categórica nominal
Tipo de servicio higiénico en el hogar	Refiere al tipo de servicio higiénico que tienen en el hogar del entrevistado	Tipo de servicio higiénico en el hogar	Dentro de la vivienda, Pozo séptico, Fuera de la vivienda pero dentro del edificio, Otros	Categórica nominal
Material predominante del piso de la vivienda	Conocer el material con que fue construido la mayor parte del piso de la vivienda de la entrevistada	Material predominante del piso de la vivienda	Cemento/ladrillo, Tierra/arena, Madera, Losetas terrazos similares, Otros	Categórica nominal
Material predominante de las paredes exteriores de la vivienda	Conocer el material predominante con el que fue construido la mayor parte de las paredes de la vivienda de la entrevistada	Material predominante de las paredes exteriores de la vivienda	Adobe o tapia, Ladrillo o bloques de cemento, Tablones/Madera, Otros	Categórica nominal
Material predominante del techo de la vivienda	Conocer el material con el que fue construido la mayor parte del techo de la vivienda de la entrevistada	Material predominante del techo de la vivienda	Plancha de calamina fibra de cemento o similares, Tejas, Concreto armado, Otros	Categórica nominal
Combustible más frecuente en el hogar para cocinar	Conocer el tipo de combustible que utilizan para cocinar	Combustible más frecuente en el hogar para cocinar	GLP, Leña, Otros	Categórica nominal
Nivel educativo de la madre	Conocer el nivel de estudios más alto aprobado por la madre	Nivel educativo de la madre	Sin educación, Primaria, Secundaria, Superior	Categórica ordinal
Diarrea en el niño(a)	Permite conocer si el niño de la entrevistada tuvo diarrea en las últimas 2 semanas	Diarrea en el niño(a)	Si, No	Categórica nominal
Recibe ayuda de Wawa wasi/Cuna más	Determinar si algún miembro del hogar recibió ayuda de Wawa wasi / Cuna más	Recibe ayuda de Wawa wasi/Cuna más	Si, No	Categórica nominal

IV. METODOLOGÍA

4.1. Ámbito de estudio: localización política y geográfica

El ámbito del estudio se instala en la región de Cusco, situado en sureste del país con una superficie de 71986 km^2 con un total poblacional de 1,205,527 habitantes, la capital está a 3399 m.s.n.m. La geografía a lo extenso de sus 13 provincias es variada ocupado de montañas, ríos y lagunas con climas tropicales, templado cálido y húmedo.

4.2. Tipo y nivel de investigación

Tipo de investigación: es de tipo retrospectivo porque los datos de la Encuesta Demográfica y de Salud Familiar (ENDES) relacionados a los años 2019 y 2020 fueron obtenidos de una base de datos pública que se encuentra en la web del Instituto Nacional de Estadística en Informática (INEI).

Nivel de investigación: es de nivel predictivo porque se predice la probabilidad de ocurrencia de la anemia en niños de 6 a 35 meses de edad en Cusco.

4.3. Población de estudio

La población estuvo constituida por niños y niñas de 6 a 35 meses de edad de Cusco de los años 2019 y 2020.

4.4. Tamaño de muestra

La muestra fue no probabilística por conveniencia. La muestra fue conformada por 505 niños y niñas de 6 a 35 meses de edad de Cusco de los años 2019 y 2020.

4.5. Técnicas de recolección de información

La documentación o información en formato electrónico de la Encuesta Demográfica y de Salud Familiar (ENDES) correspondiente a los años 2019 y 2020 se obtuvo de la página web del Instituto Nacional de Estadística en Informática (https://proyectos.inei.gob.pe/microdatos/). Los archivos digitales descargados están en un formato SAV relacionado al SPSS (acrónimo en inglés de Statistical Package for the Social Sciences).

4.6. Técnicas de análisis de los datos

Para conseguir nuestro objetivo se siguió la metodología KDD (Knowledge Discovery in Databases) y para el procesamiento de los datos se utilizó el lenguaje de programación R.

Procedimiento de análisis de datos

- **Etapa de selección:** las variables priorizadas están relacionados principalmente a los niños y sus respectivas madres, los módulos que contienen estas variables son: Características del hogar, Características de la vivienda, Inmunización y Salud, Mortalidad Materna – Violencia Familiar, Peso y Talla – Anemia y Programas Sociales de la Encuesta Demográfica y de Salud Familiar (ENDES).
- **Etapa de preprocesamiento:** se eliminan las variables que contengan una significativa cantidad de valores ausentes, si una variable posee varianza cero, esta genera ruido en el modelo, por esta razón es conveniente excluir a la variable y si algunas de las categorías de una variable cualitativa posee muy pocas observaciones en comparación a las otras categorías, puede generar errores en la validación de modelos,

una alternativa es eliminar estas observaciones dependiendo del número de categorías de la variable.

- **Etapa de transformación:** referente a las correlaciones, si dos variables numéricas tienen una elevada correlación, no es conveniente incluir a las dos variables en el modelo, una opción sería combinar dichas variables en una sola variable mediante un análisis de componentes principales (Principal Component Analysis, PCA) y normalización de datos, consiste en restar y dividir los datos con el promedio y desviación estándar respectivamente, de modo que, la escala como la varianza de las variables numéricas no puedan influir significativamente en los modelos.
- **Etapa de minería de datos:** la variable de interés (variable dependiente) a predecir es categórica, por lo tanto, los algoritmos de Machine Learning utilizados son: la regresión logística, k-nearest neighbor (KNN), naive bayes, análisis discriminante lineal (LDA), árbol de clasificación, bosques aleatorios, gradient boosting, support vector machine (SVM) y redes neuronales. Para conseguir un modelo idóneo que represente los patrones presentes en los datos, se considera lo siguiente: ajustar (entrenar) mediante algoritmos de Machine Learning en un conjunto de datos de entrenamiento, para validar el modelo se necesita predecir nuevas observaciones para poder verificar el error en que inciden, las estrategias de validación que destacan es el conjunto de test, Bootstrap y validación cruzada, respecto a la optimización de hiperparámetros, se tienen uno o varios en los diferentes algoritmos de Machine Learning, para aproximarnos a un valor óptimo de un hiperparámetro tenemos que

recurrir a estrategias de validación, una vez seleccionado el modelo optimo, se puede ejecutar para predecir nuevas observaciones.

- **Etapa de interpretación:** interpretación de los patrones encontrados.

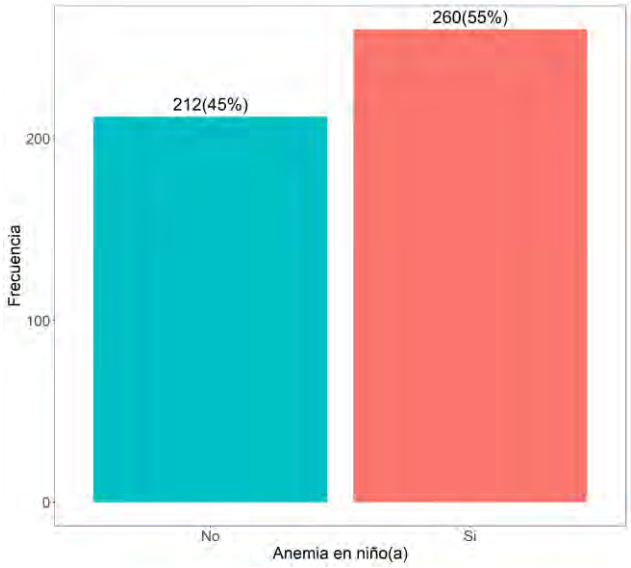
V. Resultados y Discusión

5.1. Análisis exploratorio de datos

En la etapa de procesamiento de datos, se recategorizo algunas variables categóricas como Lugar de residencia, Fuente principal de abastecimiento de agua para tomar o beber, Tipo de servicio higiénico en el hogar, Material predominante del piso de la vivienda, Material predominante de las paredes exteriores de la vivienda, Material predominante del techo de la vivienda, etc., esto para no generar errores en la validación de modelos. El procedimiento se puede verificar en el anexo 1.

Figura 9

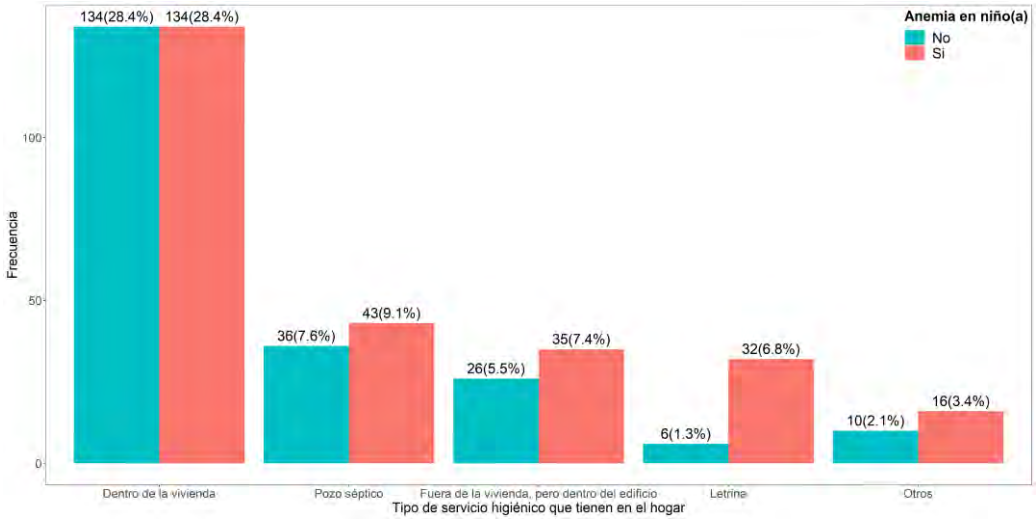
Frecuencia de diferentes diagnósticos de anemia en niños



En la figura 9 se observa la distribución de la variable anemia (variable respuesta) en niños de 6 a 35 meses de edad, 55% de niños tienen anemia de un total de 472 niños.

Figura 10

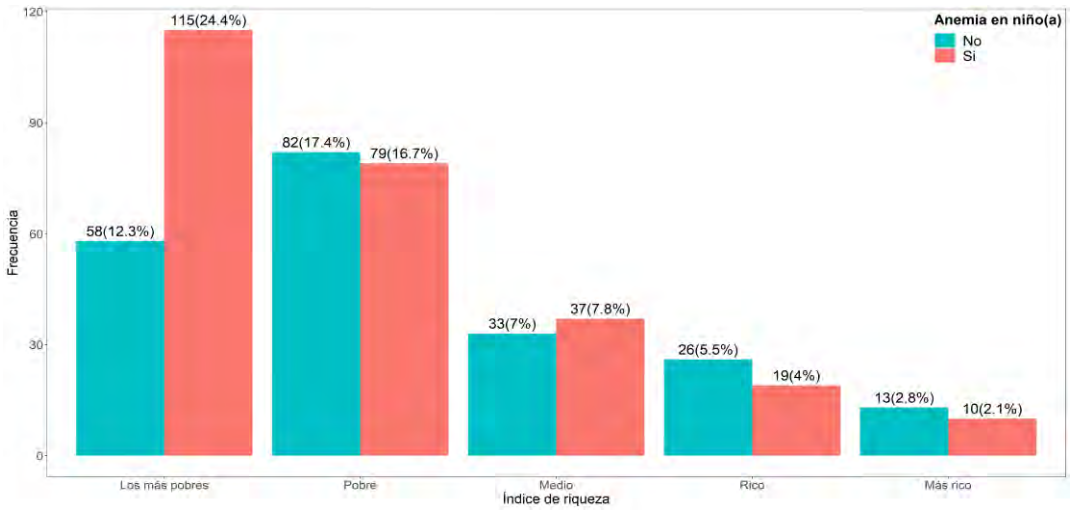
Frecuencia del tipo de servicio higiénico del hogar del niño con diferente diagnóstico de anemia



Referente a la figura 10, con porcentajes mayores es notorio que no hay diferencia entre niños con y sin anemia cuando el tipo de servicio higiénico está dentro de la vivienda de los niños, también no se observa una diferencia importante si el tipo servicio higiénico es un pozo séptico, en las demás categorías de tipo de servicio higiénico se tiene resultados similares.

Figura 11

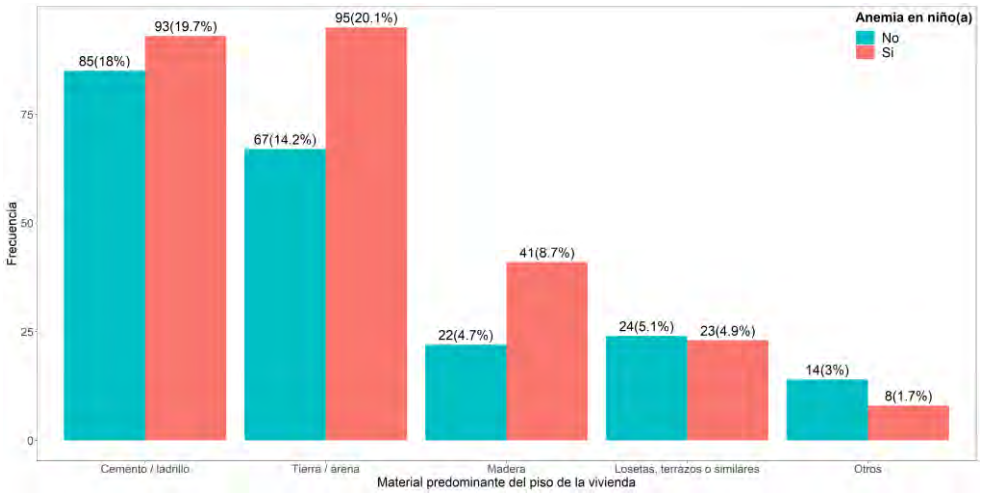
Frecuencia del índice de riqueza del niño con diferente diagnóstico de anemia



En relación a la figura 11, en la categoría de índice de riqueza de más pobres, verificamos que hay una diferencia notable entre niños con anemia (24.4%) y niños sin anemia (12.3%), en las otras categorías de índice de riqueza (Pobre, Medio, Rico y Más rico) de los niños también existen diferencias. Finalmente, los p-valores son menores a $\alpha = 0.05$ de las pruebas de Chi-cuadrado (p = 0.00326) y de Fisher (p = 0.002927), se rechaza H_0 (el índice de riqueza y la presencia de anemia en niños son independientes) y concluimos que hay alguna relación entre las variables de anemia en niños y el índice de riqueza.

Figura 12

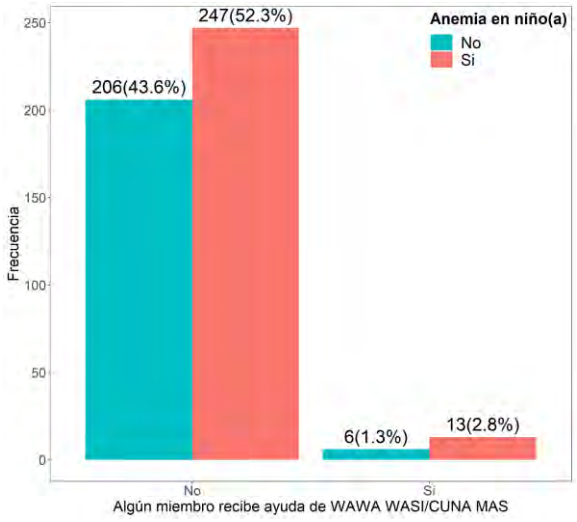
Frecuencia del material predominante del piso de la vivienda del niño con diferente diagnóstico de anemia



En la figura 12 apreciamos un porcentaje mayor de niños con anemia en las categorías de cemento/ladrillo, tierra/arena y madera sobre el material predominante del piso de la vivienda, lo que nos indica que no necesariamente tener un material del piso adecuado en la vivienda está relacionado en la disminución de la anemia en niños.

Figura 13

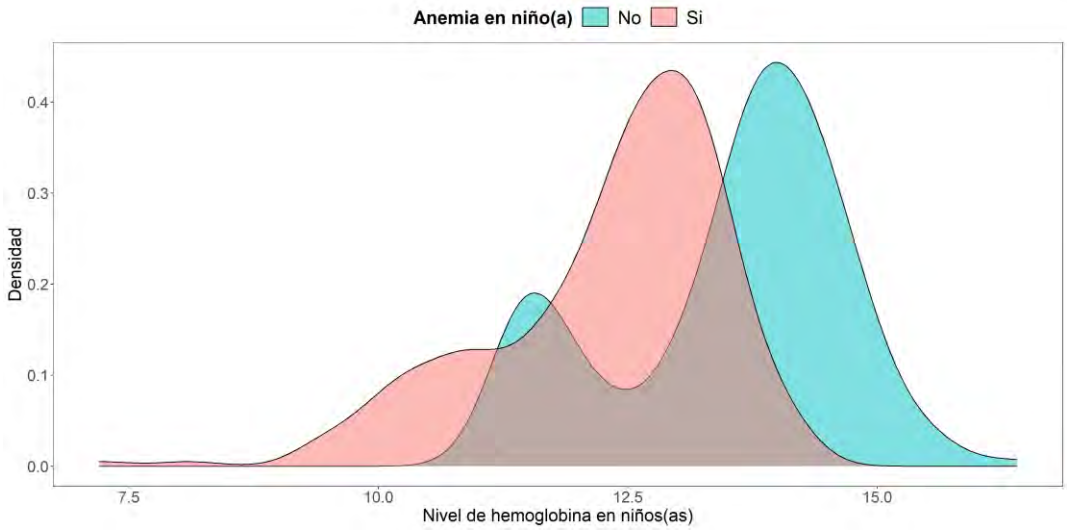
Frecuencia de miembros que reciben ayuda de Wawa wasi y/o Cuna más en la vivienda de niños con diferente diagnóstico de anemia



En la figura 13 es evidente con un porcentaje muy elevado de niños sin anemia (52.3%) y niños con anemia (43.6%), los miembros de la vivienda del niño no reciben ayuda de Wawa wasi y/o Cuna más.

Figura 14

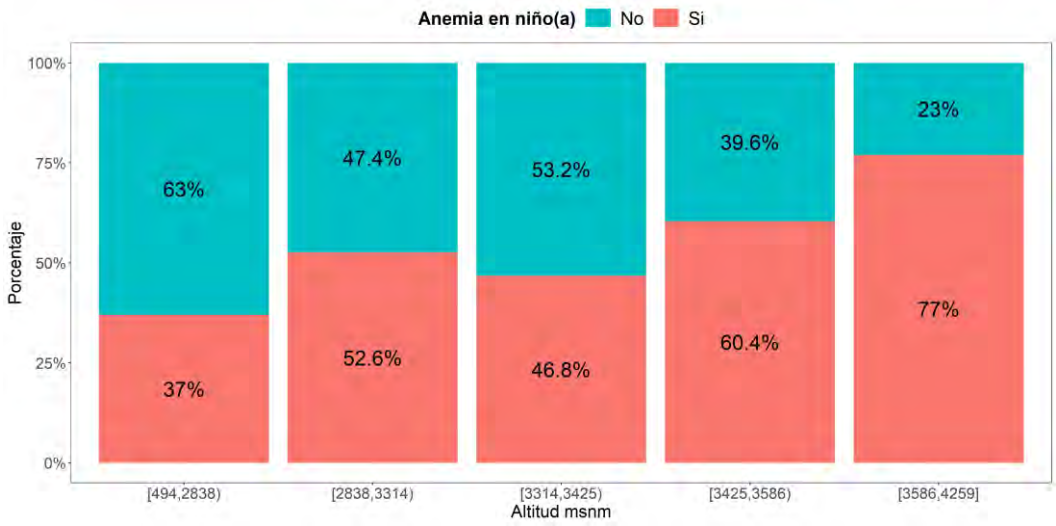
Distribución de los niveles de hemoglobina de niños con diferente diagnóstico de anemia



En la distribución de los niveles de hemoglobina relacionado a niños con anemia, hay valores mayores a 11 g/dL, lo que significa que la medida no es precisa para clasificar si un niño(a) presenta anemia o no, debido a que estas medidas generalmente varían con la altitud de residencia de los niños (ver figura 14).

Figura 15

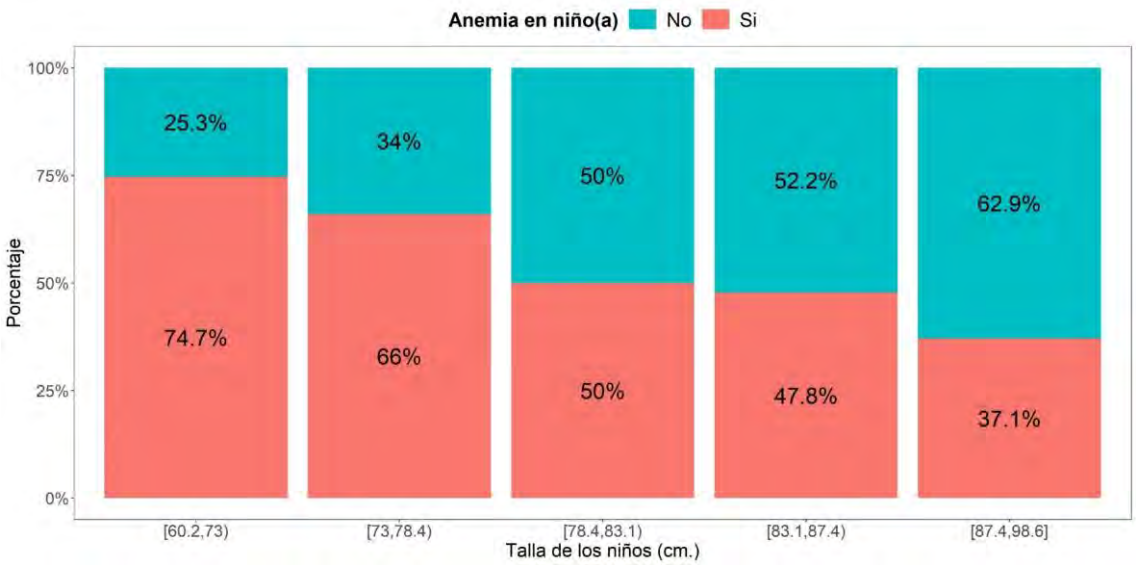
Porcentaje de diferentes diagnósticos de anemia según la altitud de residencia de niños



De acuerdo a la figura 15, verificamos que a mayor altitud de residencia del niño(a) mayor son los casos de anemia en niños de 6 a 35 meses de edad de la región de Cusco.

Figura 16

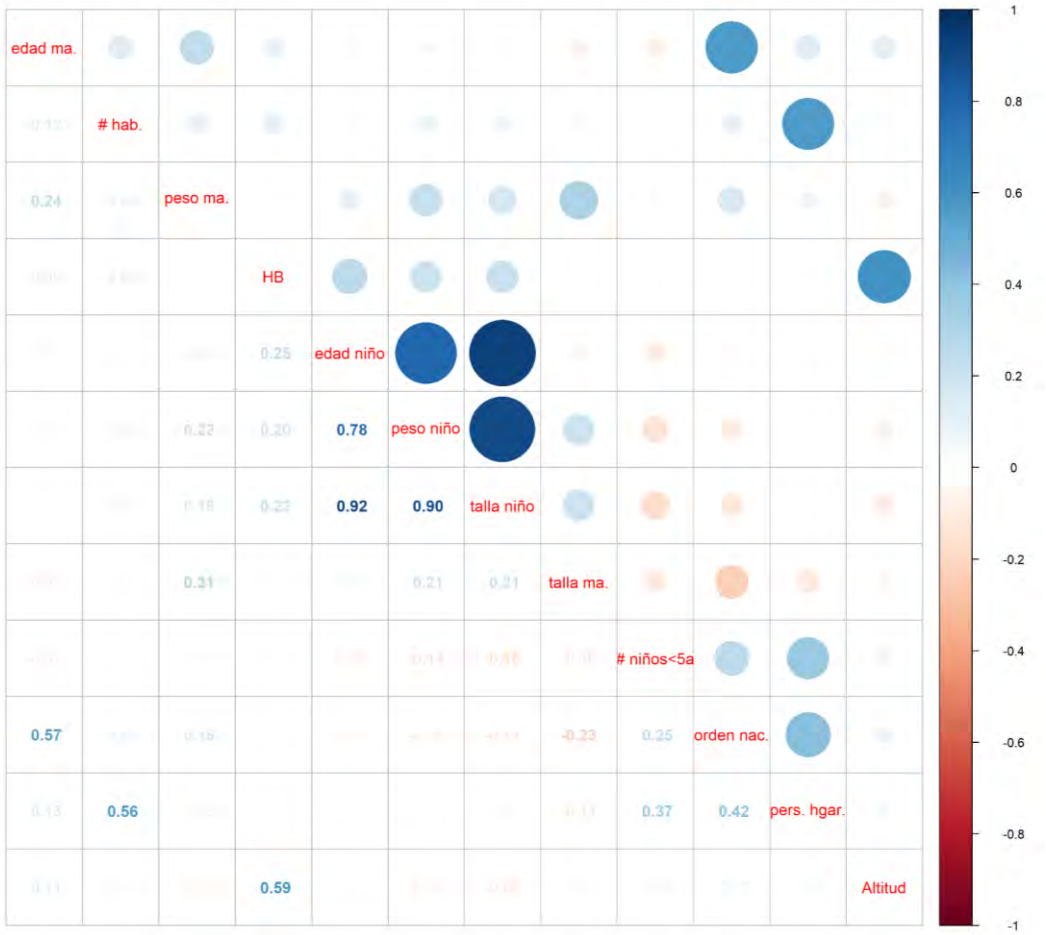
Porcentaje de diferentes diagnósticos de anemia según la talla de niños



Respecto a la relación entre la talla y la presencia de anemia en los niños, se aprecia que a mayor talla del niño es menor el porcentaje de la presencia de anemia (ver figura 16).

Figura 17

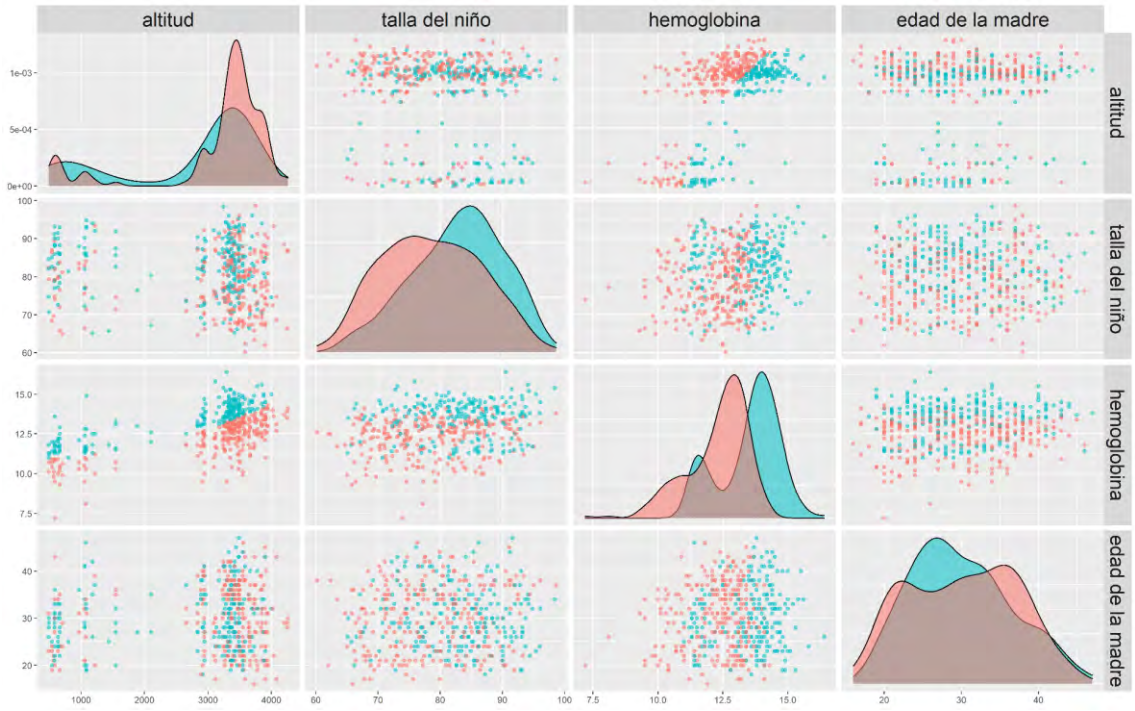
Correlaciones entre las variables de estudio



Las correlaciones positivas son evidentes entre edad, peso y talla del niño(a), las mismas variables referentes a las madres tienen correlaciones muy bajas, se observa una correlación (0.59) moderada entre el nivel de hemoglobina y la altitud. Se decidió eliminar las variables edad y peso del niño(a) para no generar información redundante a los modelos de Machine Learning (ver figura 17).

Figura 18

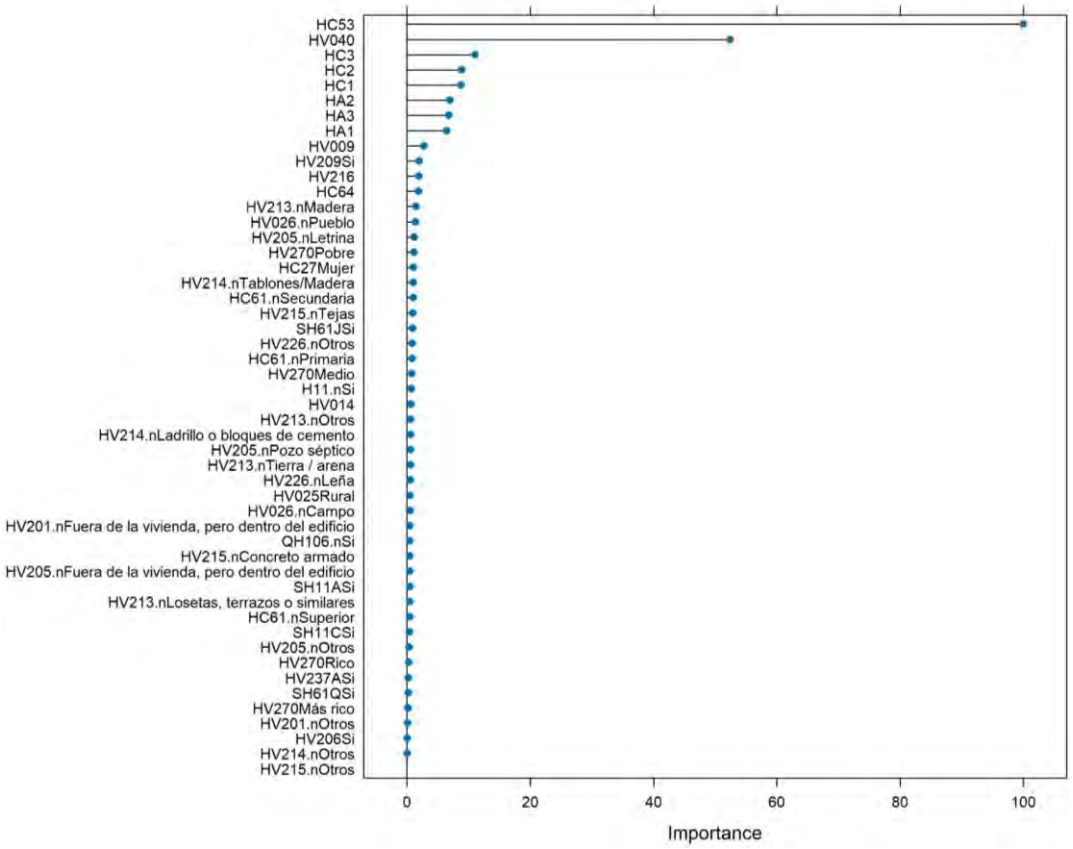
Distribuciones y dispersión de puntos entre las variables de estudio



En la figura 18 es notable que la relación entre la variable hemoglobina y altitud (talla) de los niños mediante un gráfico de dispersión de puntos clasifica adecuadamente a los niños con y sin anemia, las variables talla del niño, edad de la madre en relación con la variable hemoglobina son pésimas clasificando la variable dependiente (anemia).

Figura 19

Variables importantes



Durante la aplicación del método de exclusión recursiva de variables utilizando bosques aleatorios (random forest) se determinó que el nivel de hemoglobina (HC53) y la altitud de residencia (HV040) del niño destacan como los predictores más influyentes (ver figura 19).

De acuerdo con los análisis anteriores se determinó utilizar las variables nivel de hemoglobina y altitud de residencia para el ajuste de los modelos de Machine Learning.

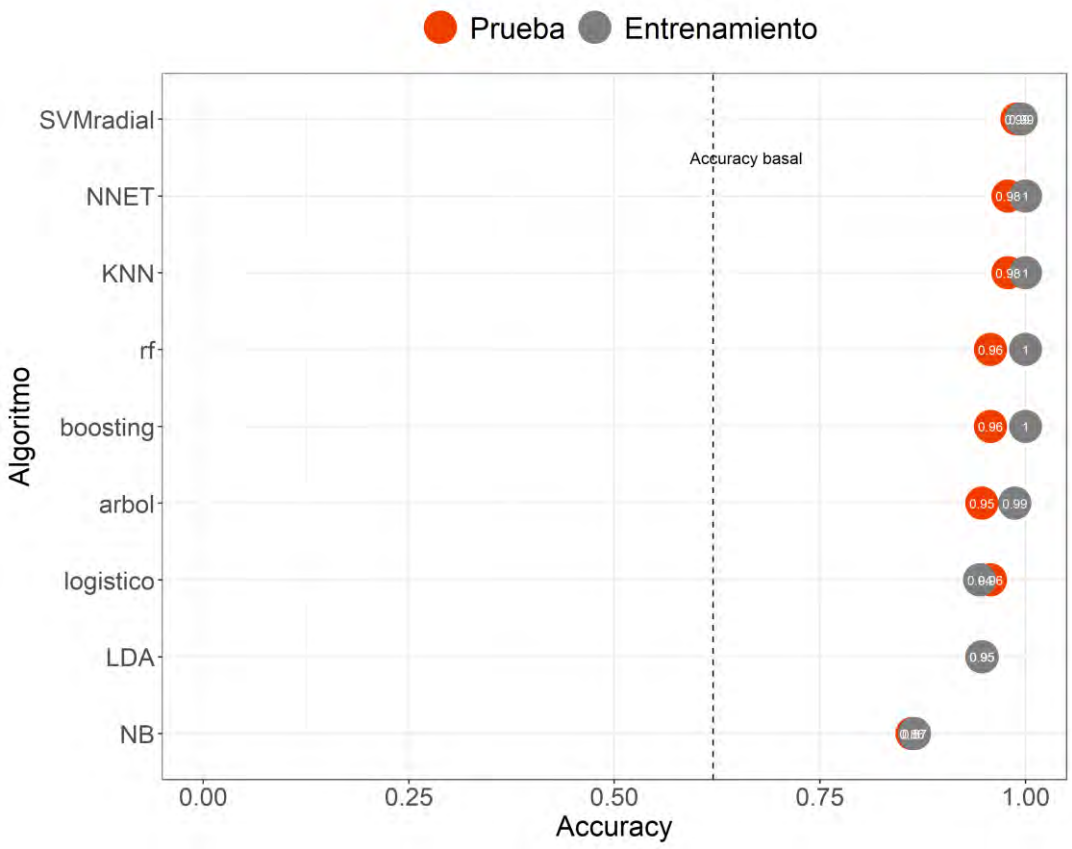
La identificación asignada a las variables según la base de datos de la Encuesta Demográfica y de Salud Familiar (ENDES) se encuentra en el anexo 2.

5.2. Comparación de modelos

Previo a determinar los modelos de Machine Learning adecuados para la clasificación de la anemia en niños y niñas de 6 a 35 meses de edad, se estableció el conjunto de entrenamiento en 378 registros y el conjunto de prueba en 94 registros.

Figura 20

Precisión de los algoritmos de Machine Learning



En la figura 20 los algoritmos a excepción de la regresión logística, consiguen más predicciones correctas en el conjunto de entrenamiento (training) que en el conjunto de prueba (test). El algoritmo Support Vector Machine (SVM) Radial consigue una predicción general (accuracy) más elevado en el conjunto de prueba.

Tabla 3

Métricas de validación de los algoritmos

Algoritmo	Sensibilidad	Especificidad	Balanced Accuracy
Regresión logística	0.9444444	0.9750000	0.9597222
KNN	**0.9807692**	**0.9761905**	**0.9784799**
Naive bayes	0.8421053	0.8918919	0.8669986
LDA	0.9433962	0.9512195	0.9473079
Árbol de clasificación	0.9433962	0.9512195	0.9473079
Bosques aleatorios	0.9615385	0.9523810	0.9569597
Gradient boosting	0.9615385	0.9523810	0.9569597
SVM	**1.0000000**	**0.9767442**	**0.9883721**
Redes neuronales	**0.9807692**	**0.9761905**	**0.9784799**

En la tabla 2, los algoritmos K-Nearest Neighbor (KNN), Support Vector Machine (SVM) y Redes neuronales obtuvieron la mayor cantidad de predicciones correctas en el conjunto de prueba. El modelo basado en SVM es el que en su totalidad predijo correctamente a los niños(as) con anemia (Sensibilidad = 1), se seleccionó el modelo SVM por su capacidad predictiva.

5.3. Discusión de resultados

Del análisis exploratorio de datos, las variables hemoglobina de los niños, altitud de residencia de los niños, la talla de los niños, y el índice de riqueza relacionado a los niños están muy asociados con la variable anemia en los niños. Sin embargo, al analizar estas variables en conjunto con los algoritmos de Machine Learning, se observó que las variables hemoglobina y la altitud de residencia de los niños tuvieron un aporte muy importante a comparación de las otras variables en los modelos establecidos, debido a que los datos observados de hemoglobina son ajustados dependiendo de la altitud de la residencia de los niños. Sow, Mukhtar, Ahmad y Suguri (2019) precisan que la variable que tiene más relevancia en la predicción de la anemia, es el indice de riqueza porque indica el nivel de vida y en

consecuencia los niños tienen acceso a alimentos y nutrición ricos en hierro. Jaiswal, Srivastava y Siddiqui (2019) y Abdullah y Al-Asmari (2016) consideradon necesario e importante la variable hemoglobina para la prediccion de la anemia. Por ultimo Canaza Espezua (2021) tambien cosidera significativa la variable niveles de hemoglobina para predecir el riesgo asociado a la anemia.

En los antecedentes de investigacion, de los algoritmos con las optimas metricas para predecir la anemia en niños, dos concluyeron que el algoritmo de bosques aleatorios (random forest) tuvieron mejor precisión de clasificación pero con una sensibilidad menor a comparación del algoritmo Support Vector Machine (SVM) en la presente investigación a causa de que la variable objetivo (variable dependiente) solo tiene dos categorías a predecir, sin embargo las otras investigaciones expuestas tienen más de dos categorías en la variable anemia, y definitivamente las variables biológicas como la Hemoglobina contribuyen significativamente en la predicción en los modelos de Machine Learning. Finalmente, en el resto de investigaciones mencionadas, concluyeron que los algoritmos de redes neuronales, naive bayes y regresión logista fueron superiores en precisión.

CONCLUSIONES

El algoritmo de Machine Learning seleccionado para esta investigación es el Support Vector Machine (SVM), es el algoritmo que tiene una alta capacidad para predecir correctamente la presencia de anemia de niños de 6 a 35 meses de edad en Cusco, lo que indica que el algoritmo es eficiente.

Las variables de mayor importancia que influyen en la presencia de la anemia en niños de 6 a 35 meses de edad en Cusco son el nivel de hemoglobina y la altitud de residencia del niño(a).

Los algoritmos de Machine Learning que se aplicaron en el conjunto de entrenamiento (train) y el conjunto de prueba (test), obtuvieron indicadores de exactitud muy buenos en ambos conjuntos, de manera que, cada uno de los algoritmos de Machine Learning predijo con alta precisión la proporción de niños con y sin anemia. Por consiguiente, se demostró que si es posible predecir la anemia en niños de 6 a 35 meses de edad en Cusco mediante diferentes algoritmos de Machine Learning.

RECOMENDACIONES

Considerar los datos de algunas de las regiones del Perú como Puno, Apurímac, Madre de Dios, Huancavelica, Pasco, Ucayali o Loreto que tienen altos porcentajes de anemia para el ajuste de los algoritmos de Machine Learning.

Se recomienda realizar la combinación de predicciones (Ensemble models) de los modelos de Machine Learning de K-Nearest Neighbor (KNN), Support Vector Machine (SVM) y Redes neuronales, por ejemplo, con el objetivo de optimizar las predicciones finales.

Poner en producción (utilizar el modelo) el algoritmo de Machine Learning seleccionado mediante una aplicación web para que el personal de salud interactúe de una forma sencilla y pueda evaluar la anemia en niños nuevos.

BIBLIOGRAFÍA

Abdullah, M., & Al-Asmari, S. (Noviembre de 2016). Anemia types prediction based on data mining classification algorithms. *ResearchGate*.

Canaza Espezua, G. (2021). *Modelo predictivo de riesgo asociado a la anemia en niños menores de 5 años en la microred yauri provincia de espinar - cusco, 2019.* Universidad Nacional del Altiplano, Escuela Profesional de Estadística e Informática, Puno.

Jaiswal, M., Srivastava, A., & Siddiqui, T. J. (Enero de 2019). Machine Learning Algorithms for Anemia disease Prediction. *ResearchGate*. doi:10.1007/978-981-13-2685-1_44

Lewis, N. D. (2017). *Machine Learning Made Easy with R: An Intuitive Step by Step Blueprint for Beginners.* CreateSpace Independent Publishing Platform.

Ministerio de Salud del Perú. (2017). *Norma Técnica - Manejo Terapéutico Y Preventivo De La Anemia En Niños, Adolescentes, Mujeres Gestantes Y Puérperas.* Lima. Obtenido de http://bvs.minsa.gob.pe/local/MINSA/4190.pdf

Pérez lópez, C., & Santín González, D. (2007). *Minería De Datos Técnicas y Herramientas.* Madrid: PARANINFO.

R Core Team. (2021). R: A Language and Environment for Statistical Computing. Viena, Austria. Obtenido de https://www.R-project.org/

Rahman Khan, J., Chowdhury, S., Islam, H., & Raheem, E. (2019). Machine Learning Algorithms To Predict The Childhood Anemia In Bangladesh. *Journal of Data Science*, págs. 195-218. doi:10.6339/JDS.201901_17(1).0009

Sow, B., Mukhtar, H., Ahmad, H. F., & Suguri, H. (2019). Assessing the relative importance of social determinants of health in malaria and anemia classification based on machine learning techniques. *Informatick for Health and Social Care*. doi:10.1080/17538157.2019.1582056

ANEXOS

Anexo 01. Script en el lenguaje de programación R

Paquetes

```
library(tidyverse)
library(haven)
library(rpart)
library(rpart.plot)
library(randomForest)
library(caret)
library(recipes)
library(xgboost)
library(doParallel)
```

Importar datos

```
## Caracteristicas del hogar
RECH0 <- as_factor(read_sav("RECH0.SAV"))
RECH0

### Cobertura de los seguros de salud
RECH4 <- as_factor(read_sav("RECH4.SAV"))
RECH4

## Caracteristicas de la vivienda
RECH23 <- as_factor(read_sav("RECH23.SAV"))
RECH23

## Peso y talla - Anemia
### Antropometria/Anemia - Niños menores de 6 años
RECH6 <- as_factor(read_sav("RECH6.SAV"))
RECH6

### Antropometria/Anemia - Mujeres de 12 a 49 años
RECH5 <- as_factor(read_sav("RECH5.SAV"))
RECH5

## Inmunización y salud
REC43 <- as_factor(read_sav("REC43.SAV"))
REC43

REC95 <- as_factor(read_sav("REC95.SAV"))
REC95

## Mortalidad materna - Violencia familiar
REC84DV <- as_factor(read_sav("REC84DV.SAV"))
REC84DV

## Programas sociales
PS_WAWAWASI <- as_factor(read_sav("PS_WAWAWASI.SAV"))
PS_WAWAWASI

PS_VL <- as_factor(read_sav("PS_VL.SAV"))
PS_VL

PS_HOGAR <- as_factor(read_sav("Programas sociales x hogar.SAV"))
PS_HOGAR
```

Procesar datos

```
## Extraer unicamente numeros de HHID
RECH0$HHID <- str_extract(RECH0$HHID, "[0-9]+")
RECH4$HHID <- str_extract(RECH4$HHID, "[0-9]+")
RECH23$HHID <- str_extract(RECH23$HHID, "[0-9]+")
```

```r
RECH6$HHID <- str_extract(RECH6$HHID, "[0-9]+")
RECH5$HHID <- str_extract(RECH5$HHID, "[0-9]+")
PS_WAWAWASI$HHID <- str_extract(PS_WAWAWASI$HHID, "[0-9]+")
PS_VL$HHID <- str_extract(PS_VL$HHID, "[0-9]+")
PS_HOGAR$HHID <- str_extract(PS_HOGAR$HHID, "[0-9]+")
RECH5$HC60 <- factor(RECH5$HA0)
RECH4$HC0 <- RECH4$IDXH4
REC95$HIDX <- REC95$IDX95

## Unir datos
datos.1 <- left_join(RECH0, RECH23, by = "HHID")

datos.2 <- left_join(RECH6, RECH5, by = c("HHID", "HC60"))

datos.3 <- left_join(datos.2, RECH4, by = c("HHID", "HC0"))

datos.4 <- left_join(REC43, REC95, by = c("CASEID", "HIDX"))

datos.5 <- left_join(datos.4, REC84DV, by = c("CASEID"))

datos.6 <- left_join(PS_VL, PS_WAWAWASI, by = c("HHID", "HVIDX"))

datos.3 <- datos.3 %>%
  group_by(HHID, HC60) %>%
  arrange(desc(HC60)) %>%
  mutate(HIDX = row_number()) %>%
  ungroup()

datos.5$HHID <- str_sub(datos.5$CASEID, 1, (str_length(datos.5$CASEID)-3))
datos.5$HHID <- str_extract(datos.5$HHID, "[0-9]+")
datos.5$HC60 <- str_sub(datos.5$CASEID, -2, -1)
datos.5$HC60 <- str_extract(datos.5$HC60, "[0-9]+")

datos.7 <- left_join(datos.3, datos.5, by = c("HHID", "HC60", "HIDX"))

datos.6$HC0 <- datos.6$HVIDX

datos.8 <- left_join(datos.7, datos.6, by = c("HHID", "HC0"))

datos.9 <- left_join(datos.8, PS_HOGAR, by = "HHID")

datos.2019 <- right_join(datos.1, datos.9, by = "HHID")

## Juntar datos de los años 2019 y 2020
datos <- rbind(datos.2019, datos.2020)
```

```r
## Datos finales
datos <- datos %>%
  select(-c(H12Z, S465DB_C, D104, PS107_1A, PS107_1M, PS102_1A, PS102_1M, QH101)) %>%
  na.omit() %>%
  mutate(
    #--- RECH0
    HV026.n = case_when(HV026 %in% c("Pueblo", "Campo") ~ as.character(HV026),
                        TRUE ~ "Capital, gran ciudad o pequeña ciudad"),
    HV026.n = factor(HV026.n,
                     levels = c("Capital, gran ciudad o pequeña ciudad",
                                "Pueblo", "Campo")),
    #--- RECH23
    HV201.n = case_when(HV201 %in% c("Dentro de la vivienda",
                                     "Fuera de la vivienda, pero dentro del edificio"
)
```

```
                        ~ as.character(HV201),
                        TRUE ~ "Otros"),
    HV201.n = factor(HV201.n),
    HV205.n = fct_collapse(HV205,
                           Letrina = c("Latrina (pozo ciego o negro)",
                                       "Letrina mejorada ventilada",
                                       "Latrina mejorada colgada/flotante")),
    HV205.n = case_when(HV205.n %in% c("Dentro de la vivienda",
                                       "Pozo séptico",
                                       "Fuera de la vivienda, pero dentro del edifici
o",
                                       "Letrina")
                        ~ as.character(HV205.n),
                        TRUE ~ "Otros"),
    HV205.n = factor(HV205.n,
                     levels = c("Dentro de la vivienda",
                                "Pozo séptico",
                                "Fuera de la vivienda, pero dentro del edificio",
                                "Letrina",
                                "Otros")),
    HV213.n = fct_collapse(HV213,
                           Otros = c("Parquet o madera pulida",
                                     "Láminas asfálticas, vinílicos o similares",
                                     "Otro")),
    HV213.n = factor(HV213.n,
                     levels = c("Cemento / ladrillo",
                                "Tierra / arena",
                                "Madera",
                                "Losetas, terrazos o similares",
                                "Otros")),
    HV214.n = case_when(HV214 %in% c("Adobe o tapia",
                                     "Ladrillo o bloques de cemento",
                                     "Tablones/Madera")
                        ~ as.character(HV214),
                        TRUE ~ "Otros"),
    HV214.n = factor(HV214.n,
                     levels = c("Adobe o tapia",
                                "Ladrillo o bloques de cemento",
                                "Tablones/Madera",
                                "Otros")),
    HV215.n = case_when(HV215 %in% c("Plancha de calamina, fibra de cemento o similar
es",
                                     "Tejas",
                                     "Concreto armado")
                        ~ as.character(HV215),
                        TRUE ~ "Otros"),
    HV215.n = factor(HV215.n,
                     levels = c("Plancha de calamina, fibra de cemento o similares",
                                "Tejas",
                                "Concreto armado",
                                "Otros")),
    HV226.n = case_when(HV226 %in% c("GLP", "Leña") ~ as.character(HV226),
                        TRUE ~ "Otros"),
    HV226.n = factor(HV226.n),
    #--- RECH6
    HC2 = HC2/10,
    HC3 = HC3/10,
    HC53 = HC53/10,
    anemia = factor(case_when(HC57 == "Sin anemia" ~ "No",
                              TRUE ~ "Si")),
    HC61.n = case_when(HC61 %in% c("Primaria", "Secundaria", "Superior") ~ as.charact
er(HC61),
                       TRUE ~ "Sin educación"),
    HC61.n = factor(HC61.n,
                    levels = c("Sin educación",
                               "Primaria",
```

```
                                "Secundaria",
                                "Superior")),
    #--- RECH5
    HA2 = HA2/10,
    HA3 = HA3/10,
    #--- REC43
    H11.n = case_when(H11 %in% c("Sí, las últimas 24 horas", "Sí, las últimas dos sem
anas")
                      ~ "Si",
                      TRUE ~ "No"),
    H11.n = factor(H11.n),
    #--- Programas sociales x hogar
    QH106.n = case_when(QH106 %in% c("No", "No Sabe/ No Recuerda", "8") ~ "No",
                        TRUE ~ "Si"),
    QH106.n = factor(QH106.n)
  ) %>%
  select(-c(HV024, HV026, HV201, HV205, HV213, HV214, HV215, HV226, HC56, HC61,
            HC57, HA50, H11, QH106))

## Eliminar correlaciones significativas
datos <- datos %>% select(-c(HC1, HC2))
```

```
## Conjunto de entrenamiento y prueba
set.seed(2022)
train <- createDataPartition(y = datos$anemia, p = 0.8, list = FALSE, times = 1)
datos.train <- datos[train, ]
datos.test  <- datos[-train, ]

## Crear el objeto recipe
objeto.recipe <- recipe(formula = anemia ~ ., data = datos)

## Normalización de variables numéricas
objeto.recipe <- objeto.recipe %>% step_center(all_numeric())
objeto.recipe <- objeto.recipe %>% step_scale(all_numeric())

## Binarización de variables categoricas
objeto.recipe <- objeto.recipe %>% step_dummy(all_nominal(), -all_outcomes())

## Se entrena el objeto recipe
trained.recipe <- prep(objeto.recipe, training = datos.train)

## Se aplican las transformaciones al conjunto de entrenamiento y de prueba
datos.train.prep <- bake(trained.recipe, new_data = datos.train)
datos.test.prep <- bake(trained.recipe, new_data = datos.test)
```

Modelos

```
particiones  <- 10
repeticiones <- 5

## Modelo de Regresión logística múltiple
hiperparametros <- data.frame(parameter = "none")

set.seed(123)
seeds <- vector(mode = "list", length = (particiones * repeticiones) + 1)
for (i in 1:(particiones * repeticiones)) {
  seeds[[i]] <- sample.int(1000, nrow(hiperparametros))
}
seeds[[(particiones * repeticiones) + 1]] <- sample.int(1000, 1)

control.train <- trainControl(method = "repeatedcv", number = particiones,
                              repeats = repeticiones, seeds = seeds,
```

```
                              returnResamp = "final", verboseIter = FALSE,
                              allowParallel = TRUE)

set.seed(2022)
modelo.logistico <- train(anemia ~ HC53 + HV040,
                          data = datos.train.prep,
                          method = "glm",
                          tuneGrid = hiperparametros,
                          metric = "Accuracy",
                          trControl = control.train,
                          family = "binomial")

modelo.logistico

summary(modelo.logistico$finalModel)

pred.logistico <- predict(modelo.logistico, datos.test.prep)
confusionMatrix(datos.test.prep$anemia, pred.logistico, positive = "Si")

## K Nearest Neighbor
hiperparametros <- data.frame(k = c(1, 2, 5, 10, 15))

set.seed(123)
seeds <- vector(mode = "list", length = (particiones * repeticiones) + 1)
for (i in 1:(particiones * repeticiones)) {
  seeds[[i]] <- sample.int(1000, nrow(hiperparametros))
}
seeds[[(particiones * repeticiones) + 1]] <- sample.int(1000, 1)

control_train <- trainControl(method = "repeatedcv", number = particiones,
                              repeats = repeticiones, seeds = seeds,
                              returnResamp = "final", verboseIter = FALSE,
                              allowParallel = TRUE)

set.seed(2022)
modelo.knn <- train(anemia ~ HC53 + HV040,
                    data = datos.train.prep,
                    method = "knn",
                    tuneGrid = hiperparametros,
                    metric = "Accuracy",
                    trControl = control_train)
modelo.knn

pred.knn <- predict(modelo.knn, datos.test.prep)
confusionMatrix(datos.test.prep$anemia, pred.knn, positive = "Si")

## Naive Bayes
hiperparametros <- data.frame(usekernel = FALSE, fL = 0 , adjust = 0)

set.seed(123)
seeds <- vector(mode = "list", length = (particiones * repeticiones) + 1)
for (i in 1:(particiones * repeticiones)) {
  seeds[[i]] <- sample.int(1000, nrow(hiperparametros))
}
seeds[[(particiones * repeticiones) + 1]] <- sample.int(1000, 1)

control_train <- trainControl(method = "repeatedcv", number = particiones,
                              repeats = repeticiones, seeds = seeds,
                              returnResamp = "final", verboseIter = FALSE,
                              allowParallel = TRUE)

set.seed(2022)
modelo.nb <- train(anemia ~ HC53 + HV040,
```

```
                  data = datos.train.prep,
                  method = "nb",
                  tuneGrid = hiperparametros,
                  metric = "Accuracy",
                  trControl = control_train)

modelo.nb

pred.nb <- predict(modelo.nb, datos.test.prep)
confusionMatrix(datos.test.prep$anemia, pred.nb, positive = "Si")

## Análisis discriminante lineal
hiperparametros <- data.frame(parameter = "none")

set.seed(123)
seeds <- vector(mode = "list", length = (particiones * repeticiones) + 1)
for (i in 1:(particiones * repeticiones)) {
  seeds[[i]] <- sample.int(1000, nrow(hiperparametros))
}
seeds[[(particiones * repeticiones) + 1]] <- sample.int(1000, 1)

control_train <- trainControl(method = "repeatedcv", number = particiones,
                              repeats = repeticiones, seeds = seeds,
                              returnResamp = "final", verboseIter = FALSE,
                              allowParallel = TRUE)

set.seed(2022)
modelo.lda <- train(anemia ~ HC53 + HV040,
                    data = datos.train.prep,
                    method = "lda",
                    tuneGrid = hiperparametros,
                    metric = "Accuracy",
                    trControl = control_train)

modelo.lda

pred.lda <- predict(modelo.lda, datos.test.prep)
confusionMatrix(datos.test.prep$anemia, pred.lda, positive = "Si")

## Árbol de clasificación
hiperparametros <- data.frame(parameter = "none")

set.seed(123)
seeds <- vector(mode = "list", length = (particiones * repeticiones) + 1)
for (i in 1:(particiones * repeticiones)) {
  seeds[[i]] <- sample.int(1000, nrow(hiperparametros))
}
seeds[[(particiones * repeticiones) + 1]] <- sample.int(1000, 1)

control_train <- trainControl(method = "repeatedcv", number = particiones,
                              repeats = repeticiones, seeds = seeds,
                              returnResamp = "final", verboseIter = FALSE,
                              allowParallel = TRUE)

set.seed(2022)
modelo.c50tree <- train(anemia ~ HC53 + HV040,
                        data = datos.train.prep,
                        method = "C5.0Tree",
                        tuneGrid = hiperparametros,
                        metric = "Accuracy",
                        trControl = control_train)

modelo.c50tree
```

```r
summary(modelo.c50tree$finalModel)

pred.c50tree <- predict(modelo.c50tree, datos.test.prep)
confusionMatrix(datos.test.prep$anemia, pred.c50tree, positive = "Si")

## Random Forest
hiperparametros <- expand.grid(mtry = c(1, 2),
                               min.node.size = c(2, 3, 4, 5, 10, 15, 20, 30),
                               splitrule = "gini")

set.seed(123)
seeds <- vector(mode = "list", length = (particiones * repeticiones) + 1)
for (i in 1:(particiones * repeticiones)) {
  seeds[[i]] <- sample.int(1000, nrow(hiperparametros))
}
seeds[[(particiones * repeticiones) + 1]] <- sample.int(1000, 1)

control_train <- trainControl(method = "repeatedcv", number = particiones,
                              repeats = repeticiones, seeds = seeds,
                              returnResamp = "final", verboseIter = FALSE,
                              allowParallel = TRUE)

set.seed(2022)
modelo.rf <- train(anemia ~ HC53 + HV040,
                   data = datos.train.prep,
                   method = "ranger",
                   tuneGrid = hiperparametros,
                   metric = "Accuracy",
                   trControl = control_train,
                   num.trees = 500)
modelo.rf

pred.rf <- predict(modelo.rf, datos.test.prep)
confusionMatrix(datos.test.prep$anemia, pred.rf,  positive = "Si")

## Gradient Boosting
hiperparametros <- expand.grid(interaction.depth = c(1, 2),
                               n.trees = c(500, 1000, 2000),
                               shrinkage = c(0.001, 0.01, 0.1),
                               n.minobsinnode = c(2, 5, 15))

set.seed(123)
seeds <- vector(mode = "list", length = (particiones * repeticiones) + 1)
for (i in 1:(particiones * repeticiones)) {
  seeds[[i]] <- sample.int(1000, nrow(hiperparametros))
}
seeds[[(particiones * repeticiones) + 1]] <- sample.int(1000, 1)

control_train <- trainControl(method = "repeatedcv", number = particiones,
                              repeats = repeticiones, seeds = seeds,
                              returnResamp = "final", verboseIter = FALSE,
                              allowParallel = TRUE)

set.seed(2022)
modelo.boost <- train(anemia ~ HC53 + HV040,
                      data = datos.train.prep,
                      method = "gbm",
                      tuneGrid = hiperparametros,
                      metric = "Accuracy",
                      trControl = control_train,
                      distribution = "adaboost",
                      verbose = FALSE)
```

```r
modelo.boost

pred.boost <- predict(modelo.boost, datos.test.prep)
confusionMatrix(datos.test.prep$anemia, pred.boost,  positive = "Si")

## Máquinas vectorial de soporte
hiperparametros <- expand.grid(sigma = c(0.001, 0.01, 0.1, 0.5, 1),
                               C = c(1, 20, 50, 100, 200, 500, 700))

set.seed(123)
seeds <- vector(mode = "list", length = (particiones * repeticiones) + 1)
for (i in 1:(particiones * repeticiones)) {
  seeds[[i]] <- sample.int(1000, nrow(hiperparametros))
}
seeds[[(particiones * repeticiones) + 1]] <- sample.int(1000, 1)

control_train <- trainControl(method = "repeatedcv", number = particiones,
                              repeats = repeticiones, seeds = seeds,
                              returnResamp = "final", verboseIter = FALSE,
                              allowParallel = TRUE)

set.seed(2022)
modelo.svmrad <- train(anemia ~ HC53 + HV040,
                       data = datos.train.prep,
                       method = "svmRadial",
                       tuneGrid = hiperparametros,
                       metric = "Accuracy",
                       trControl = control_train)

modelo.svmrad

pred.svmrad <- predict(modelo.svmrad, datos.test.prep)
confusionMatrix(datos.test.prep$anemia, pred.svmrad,  positive = "Si")

## Modelo de Red Neuronal Artificial
hiperparametros <- expand.grid(size = c(2, 4, 6, 8, 10, 20),
                               decay = c(0.0001, 0.1, 0.5))

set.seed(123)
seeds <- vector(mode = "list", length = (particiones * repeticiones) + 1)
for (i in 1:(particiones * repeticiones)) {
  seeds[[i]] <- sample.int(1000, nrow(hiperparametros))
}
seeds[[(particiones * repeticiones) + 1]] <- sample.int(1000, 1)

control.train <- trainControl(method = "repeatedcv", number = particiones,
                              repeats = repeticiones, seeds = seeds,
                              returnResamp = "final", verboseIter = FALSE,
                              allowParallel = TRUE)

set.seed(2022)
modelo.nnet <- train(anemia ~ HC53 + HV040,
                     data = datos.train.prep,
                     method = "nnet",
                     tuneGrid = hiperparametros,
                     metric = "Accuracy",
                     trControl = control.train,
                     MaxNWts = 2000,
                     trace = FALSE)
modelo.nnet
```

```r
pred.nnet <- predict(modelo.nnet, datos.test.prep)
confusionMatrix(datos.test.prep$anemia, pred.nnet,  positive = "Si")
```

Comparación de modelos

```r
## Métricas de validación
modelos <- list(logistico = modelo.logistico,
                KNN = modelo.knn,
                NB = modelo.nb,
                LDA = modelo.lda,
                arbol = modelo.c50tree,
                rf = modelo.rf,
                boosting = modelo.boost,
                SVMradial = modelo.svmrad,
                NNET = modelo.nnet)

resultados_resamples <- resamples(modelos)
resultados_resamples$values

metricas_resamples <- resultados_resamples$values %>%
  gather(key = "modelo", value = "valor", -Resample) %>%
  separate(col = "modelo", into = c("modelo", "metrica"),
           sep = "~", remove = TRUE)

metricas_resamples

metricas_resamples %>%
  group_by(modelo, metrica) %>%
  summarise(total = n())

metricas_resamples %>%
  group_by(modelo, metrica) %>%
  summarise(media = mean(valor)) %>%
  spread(key = metrica, value = media) %>%
  arrange(desc(Accuracy))

## Error de test
predicciones <- extractPrediction(
  models = modelos,
  testX = datos.test.prep %>% select(c(HC53, HV040)),
  testY = datos.test.prep$anemia
)

predicciones %>%
  filter(model == "glm" & dataType == "Test")

metricas_predicciones <- predicciones %>%
  mutate(acierto = ifelse(obs == pred, TRUE, FALSE)) %>%
  group_by(object, dataType) %>%
  summarise(accuracy = mean(acierto))

metricas_predicciones %>%
  spread(key = dataType, value = accuracy) %>%
  arrange(desc(Test))
```

Anexo 02. Variables seleccionadas de la base de datos de la ENDES

Base de datos (código)	Variable	Descripción de la variable
Hogar (RECH0)	HV009	Número de personas en el hogar
	HV014	Números de niños menor a 5 años
	HV025	Área de residencia
	HV026	Lugar de residencia
	HV040	Altitud de residencia
Vivienda (RECH23)	HV201	Fuente principal de abastecimiento de agua para tomar o beber
	HV205	Tipo de servicio higiénico en el hogar
	HV206	Electricidad en el hogar
	HV209	Refrigeradora o congeladora en el hogar
	HV213	Material predominante del piso de la vivienda
	HV214	Material predominante de las paredes exteriores de la vivienda
	HV215	Material predominante del techo de la vivienda
	HV216	Número de habitaciones
	HV226	Combustible más frecuente en el hogar para cocinar
	HV237A	El agua que se bebe o toma es hervida
	SH61J	Televisión por cable en el hogar
	SH61Q	Internet en casa
	HV270	Índice de riqueza
Antropometría/Anemia (RECH6)	HC1	Edad del niño(a)
	HC3	Talla del niño(a)
	HC2	Peso del niño(a)
	HC27	Sexo del niño(a)
	HC53	Nivel de Hemoglobina
	HC61	Nivel educativo de la madre
	HC64	Número de orden de nacimiento
Antropometría/Anemia - Mujeres (RECH5)	HA1	Edad de la madre
	HA3	Talla de la madre
	HA2	Peso de la madre
Programas Sociales x Hogar	QH106	Recibe ayuda de Wawa wasi/Cuna más
Seguros de Salud (RECH4)	SH11A	Afiliado al seguro de salud de EsSalud
	SH11C	Afiliado al Seguro Integral de Salud (SIS)
Inmunización y Salud (REC43)	H11	Diarrea en el niño(a)

Anexo 03. Datos

HV009	HV014	HV025	HV040	HV206	HV209	HV216	HV237A	SH61J	SH61Q	HV270	HC1	HC2	HC3	HC27	HC53	HC64	HA1	HA2	HA3	SH11A	SH11C	HV026.n	HV201.n	HV205.n	HV213.n	HV214.n	HV215.n	HV226.n	anemia	HC61.n	H11.n	QH106.n
6	2	Urbano	3359	Si	Si	2	Si	No	No	Medio	11	10.5	73.3	Hombre	12.7	4	40	67.7	147	No	Si	Capital, gran ci	Dentro de la vi	Dentro de la v	Cemento / lac	Ladrillo o bl	Concreto	GLP	Si	Secundaria	No	No
4	1	Urbano	3359	Si	Si	1	Si	No	No	Pobre	23	11	83	Mujer	13.6	2	34	62.7	153.2	No	Si	Capital, gran ci	Dentro de la vi	Letrina	Cemento / lac	Adobe o tap	Tejas	GLP	No	Secundaria	No	No
6	1	Urbano	3359	Si	No	4	Si	No	No	Pobre	25	11.2	85.4	Mujer	13.3	3	42	84.1	152.4	Si	No	Capital, gran ci	Dentro de la vi	Dentro de la v	Tierra / arena	Adobe o tap	Tejas	GLP	Si	Secundaria	No	No
4	1	Urbano	3285	Si	Si	2	Si	Si	No	Rico	29	14.3	94.4	Hombre	13.8	1	22	58.5	162.3	No	Si	Capital, gran ci	Dentro de la vi	Dentro de la v	Losetas, terra	Ladrillo o bl	Concreto	GLP	No	Secundaria	No	No
4	1	Urbano	3285	Si	No	2	Si	No	No	Pobre	34	13.2	94.2	Mujer	15.2	3	33	76.4	160.4	No	Si	Capital, gran ci	Fuera de la viv	Dentro de la v	Tierra / arena	Adobe o tap	Tejas	GLP	No	Secundaria	No	No
4	1	Urbano	3266	Si	Si	3	Si	Si	Si	Más ric	7	8.5	68.9	Hombre	11.5	2	37	70.9	150.9	No	Si	Capital, gran ci	Dentro de la vi	Dentro de la v	Losetas, terra	Ladrillo o bl	Concreto	GLP	Si	Secundaria	Si	No
5	1	Urbano	3266	Si	No	1	Si	No	No	Los má	15	9.8	77	Mujer	11.5	4	37	60.9	142.2	No	Si	Capital, gran ci	Fuera de la viv	Fuera de la vi	Cemento / lac	Adobe o tap	Tejas	GLP	Si	Secundaria	Si	No
3	1	Urbano	3266	Si	Si	2	Si	No	No	Medio	16	8.6	74.3	Mujer	13.6	1	41	61.1	152.4	Si	No	Capital, gran ci	Dentro de la vi	Dentro de la v	Madera	Adobe o tap	Tejas	GLP	No	Superior	Si	No
5	1	Urbano	3328	Si	Si	2	Si	Si	No	Rico	26	11.4	83	Hombre	14.4	3	36	67.1	150	No	Si	Capital, gran ci	Dentro de la vi	Dentro de la v	Losetas, terra	Ladrillo o bl	Concreto	GLP	No	Secundaria	No	No
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
4	1	Rural	540	Si	No	1	Si	No	No	Pobre	20	11.2	82.6	Hombre	11.4	1	20	51.9	159.4	No	Si	Campo	Dentro de la vi	Dentro de la v	Cemento / lac	Ladrillo o bl	Concreto	GLP	No	Secundaria	No	No
4	1	Rural	540	Si	No	1	Si	No	No	Pobre	23	10.8	85.1	Mujer	12	3	25	64.9	155.6	No	Si	Campo	Dentro de la vi	Fuera de la vi	Cemento / lac	Ladrillo o bl	Plancha de	GLP	No	Secundaria	Si	No
5	3	Rural	540	Si	No	1	Si	No	No	Pobre	6	8.6	68.4	Hombre	10.7	3	23	76	152.1	No	Si	Campo	Fuera de la viv	Fuera de la vi	Cemento / lac	Ladrillo o bl	Concreto	GLP	Si	Secundaria	No	No
6	1	Rural	2890	Si	No	4	Si	No	No	Medio	15	8.2	72.3	Mujer	13	4	35	76.7	140.2	No	Si	Campo	Dentro de la vi	Dentro de la v	Cemento / lac	Ladrillo o bl	Tejas	GLP	No	Primaria	Si	No
4	1	Rural	2890	Si	Si	1	Si	No	No	Medio	17	10	79.1	Mujer	11.3	2	36	63.3	158.2	No	No	Campo	Dentro de la vi	Dentro de la v	Cemento / lac	Ladrillo o bl	Concreto	GLP	Si	Superior	No	No
4	1	Rural	2890	Si	Si	3	Si	No	No	Pobre	7	9.5	69.2	Hombre	13.7	2	38	78.4	157.1	Si	No	Campo	Dentro de la vi	Dentro de la v	Tierra / arena	Adobe o tap	Tejas	GLP	No	Secundaria	Si	No
8	2	Rural	2890	Si	No	2	Si	Si	No	Los má	12	9.5	74.5	Mujer	11.7	1	18	53.7	160.1	No	Si	Campo	Dentro de la vi	Letrina	Tierra / arena	Adobe o tap	Tejas	Leña	Si	Secundaria	No	No
5	1	Rural	2890	Si	No	1	Si	No	No	Los má	26	10.1	81.9	Hombre	13.1	3	34	82.6	150.2	No	Si	Campo	Dentro de la vi	Otros	Tierra / arena	Adobe o tap	Tejas	Leña	No	Secundaria	No	No