UNS
ESCUELA DE
POSGRADO

PROGRAMA DOCTORAL

"Aplicación del Algoritmo "Random Forest" para un modelo de clasificación sobre la tenencia de anemia de niños del Perú "

Tesis para obtener el Grado Académico de Doctor en Estadística Matemática

Autor:

Mg. Céspedes Panduro, Bernardo

Asesor:

Dr. Aguilar Marin, Pablo
DNI : 18071385
Codigo ORCID: 0000-0001-6096-4010

Nuevo Chimbote - PERÚ
2022

UNS
ESCUELA DE
POSGRADO

CONSTANCIA DE ASESORAMIENTO

Yo, **Dr. PABLO AGUILAR MARÍN**, mediante la presente certifico mi asesoramiento de la Tesis Doctoral titulada: **APLICACIÓN DEL ALGORITMO "RANDOM FOREST" PARA UN MODELO DE CLASIFICACIÓN SOBRE LA TENENCIA DE ANEMIA DE NIÑOS DEL PERÚ**, elaborada por el magister **BERNARDO CESPEDES PANDURO** para obtener el Grado Académico de Doctor en **ESTADÍSTICA MATEMÁTICA** en la Escuela de Posgrado de la Universidad Nacional del Santa.

Nuevo Chimbote, 15 Abril del 2022

.....

Dr. Pablo Aguilar Marín
ASESOR
DNI: 18071385
Código ORCID: 0000-0001-6096-4010

UNS
ESCUELA DE
POSGRADO

CONFORMIDAD DEL JURADO EVALUADOR

APLICACIÓN DEL ALGORITMO "RANDOM FOREST" PARA UN MODELO DE CLASIFICACIÓN SOBRE LA TENENCIA DE ANEMIA DE NIÑOS DEL PERÚ

Revisado y Aprobado por el Jurado Evaluador:

.....
Dr. LUIS ALBERTO RUBIO JACOBO
PRESIDENTE
DNI: 18069833
Código ORCID: 0000-0001-5060-9998

.....
Dr. ALFONSO TESEN ARROYO
SECRETARIO
DNI: 17578166
Código ORCID: 0000-0002-1088-7093

.....
Dr. PABLO AGUILAR MARIN
VOCAL
DNI: 18071385
Código ORCID: 0000-0001-6096-4010

UNS
ESCUELA DE POSGRADO

ACTA DE EVALUACIÓN DE SUSTENTACIÓN VIRTUAL DE TESIS

A los once días del mes de marzo del año 2022, siendo las **21:00 horas**, a través de la plataforma de videoconferencia *https://meet.google.com/bqg-rcgm-uuv*, se reunieron los miembros del Jurado Evaluador designados mediante Resolución Directoral N° 559-2021-EPG-UNS de fecha 14 de diciembre de 2021, conformado por los docentes: Dr. Luis Alberto Rubio Jacobo (Presidente), Dr. Alfonso Tesen Arroyo (Secretario) y Dr. Pablo Aguilar Marín (Vocal), con la finalidad de evaluar la sustentación virtual de la tesis titulada: ***APLICACIÓN DEL ALGORITMO "RANDOM FOREST" PARA UN MODELO DE CLASIFICACIÓN SOBRE LA TENENCIA DE ANEMIA DE NIÑOS DEL PERÚ*****,** presentado por el tesista **Bernardo Céspedes Panduro**, egresado del programa de **Doctorado en Estadística Matemática**.

Sustentación autorizada mediante Resolución Directoral N° 022-2022-EPG-UNS de fecha 04 de marzo de 2022.

El Presidente del jurado autorizó el inicio del acto académico; producido y concluido el acto de sustentación de tesis, los miembros del jurado procedieron a la evaluación respectiva, haciendo una serie de preguntas y recomendaciones al tesista, quien dio respuestas a las interrogantes y observaciones.

El jurado después de deliberar sobre aspectos relacionados con el trabajo, contenido y sustentación del mismo y con las sugerencias pertinentes, declara la sustentación como **APROBADO**, asignándole la calificación de **17.**

Siendo las **22:48** horas del mismo día se da por finalizado el acto académico, firmando la presente acta en señal de conformidad.

Dr. Luis Alberto Rubio Jacobo
Presidente

Dr. Alfonso Tesen Arroyo
Secretario

Dr. Pablo Aguilar Marín
Vocal

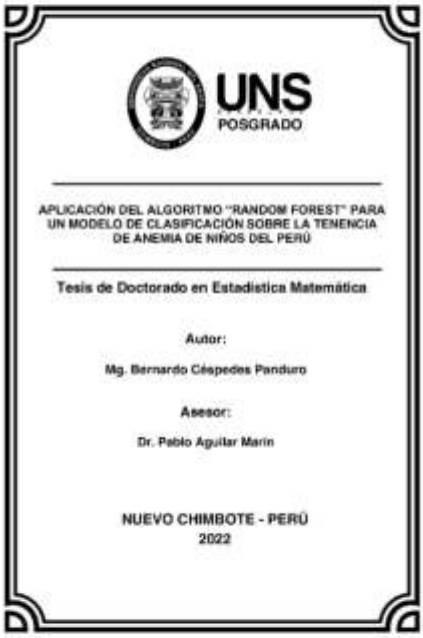
turnitin

Recibo digital

Este recibo confirma quesu trabajo ha sido recibido por Turnitin. A continuación podrá ver la información del recibo con respecto a su entrega.

La primera página de tus entregas se muestra abajo.

Autor de la entrega:	Bernardo CESPEDES PANDURO
Título del ejercicio:	DOCTORADO ESTADÍSTICA MATEMÁTICA
Título de la entrega:	APLICACIÓN DEL ALGORITMO "RANDOM FOREST" PARA UN ...
Nombre del archivo:	TESIS_UNS_-_BERNARDO_CESPEDES_PANDURO_F.doc
Tamaño del archivo:	4.19M
Total páginas:	153
Total de palabras:	34,092
Total de caracteres:	191,596
Fecha de entrega:	15-abr.-2022 12:56a. m. (UTC-0500)
Identificador de la entre...	1217996132



UNS
POSGRADO

APLICACIÓN DEL ALGORITMO "RANDOM FOREST" PARA UN MODELO DE CLASIFICACIÓN SOBRE LA TENENCIA DE ANEMIA DE NIÑOS DEL PERÚ

Tesis de Doctorado en Estadística Matemática

Autor:

Mg. Bernardo Céspedes Panduro

Asesor:

Dr. Pablo Aguilar Marín

NUEVO CHIMBOTE - PERÚ
2022

DEDICATORIA

A Dios, por permitirme seguir con vida.

A mis padres y hermanos, por su apoyo constante, por ser quien soy en la vida.

A mis hijos, Giulio, Luana, Joshua y Piero, quienes son mi motor y motivo.

A mi esposa Rocío, por ser mi compañera de vida.

AGRADECIMIENTO

A mi asesor de tesis Dr. Pablo Aguilar Marín por su invaluable apoyo en la culminación del presente trabajo de tesis.

A los docentes de la Escuela de Postgrado de la Universidad Nacional del Santa por compartir sus conocimientos y experiencias académicas.

A mis compañeros y amigos que siempre me inculcaron crecer profesionalmente para el servicio de la sociedad.

ÍNDICE GENERAL

LISTA DE TABLAS

LISTA DE FIGURAS

RESUMEN

En este trabajo de investigación se ha elaborado y aplicado el algoritmo “random forest” para un modelo de clasificación, con la finalidad de predecir la tenencia de anemia en niños de 6 a 35 meses de edad nacidos en todo el Perú, utilizando la base de datos recolectada a través de la Encuesta Demográfica y de Salud Familiar (ENDES) por el Instituto Nacional de Estadística e Informática (INEI), durante los años 2015 al 2019, conformada por 57410 registros de encuestados. Se seleccionaron 33 variables independientes de todas las que recoge la ENDES. Se plantearon seis procedimientos alternativos utilizando una combinación de los criterios de balanceo de datos y reajuste de parámetros para la predicción de anemia, obteniéndose valores de los indicadores, Área Bajo la Curva (AUC), nivel de especificidad y nivel de sensibilidad para cada uno de ellos. De los seis procedimientos, el que mejor predijo la tenencia de anemia con valores para los indicadores de especificidad (63,6%) y sensibilidad (65,9%) más similares fue el que utiliza datos balanceados con un reajuste de los parámetros, reduciendo la cantidad de arboles y con selección de variables. Las 5 variables independientes más importantes para este modelo en la tenencia de anemia son: variables relacionadas con el niño (edad del niño, en meses), variables sociodemográficas (altitud del conglomerado, en metros), variables del cuidado materno e infantil (número de visitas prenatales por embarazo, meses de embarazo del primer control prenatal y talla de la madre en centímetros).

Palabras clave: tenencia de anemia, “random forest”, balanceo de datos, indicador área bajo la curva, sensibilidad, especificidad, modelo de clasificación.

ABSTRACT

In this research work, the "random forest" algorithm has been developed and applied for a classification model, with the aim of predicting the presence of anemia in children from 6 to 35 months of age born throughout Peru, using the basis of data collected through the Demographic and Family Health Survey (ENDES) by the National Institute of Statistics and Informatics (INEI), during the years 2015 to 2019, made up of 57,410 records of respondents. 33 independent variables were selected from all those included in the ENDES. Six alternative procedures were proposed using a combination of data balancing criteria and parameter readjustment for anemia prediction, obtaining values of the indicators, Area Under the Curve (AUC), level of specificity and level of sensitivity for each one of them. Of the six procedures, the one that best predicted the presence of anemia with values for the indicators of specificity (63.6%) and sensitivity (65.9%) more similar was the one that uses balanced data with a readjustment of the parameters, reducing the number of trees and with variable selection. The 5 most important independent variables for this model in the presence of anemia are: variables related to the child (age of the child, in months), sociodemographic variables (altitude of the cluster, in meters), variables of maternal and child care (number of prenatal visits by pregnancy, months of pregnancy of the first prenatal control and height of the mother in centimeters).

Keywords: tenure of anemia, random forest, data balancing, indicator area under the curve, sensitivity, specificity, classification model.

INTRODUCCIÓN

La anemia se define como una reducción anormal de la concentración de hemoglobina en sangre como consecuencia de la carencia de uno o más nutrientes esenciales, siendo la deficiencia de hierro en el organismo una de sus causas más importantes ya que origina una disminución de hemoglobina en la sangre (Organización Mundial de la Salud [OMS], 2020), contribuyendo a aproximadamente el 90% de los tipos de anemia existentes (Fondo de las Naciones Unidas para la Infancia [UNICEF], 1998). Este trastorno nutricional afecta a todas las etapas de la vida, sin embargo, los niños preescolares son los más propensos (OMS, 2017). En el caso de los preescolares, la Organización Mundial de la Salud (OMS) considera que 293,1 millones padecen de anemia a nivel mundial y la mitad de estos casos es debido a la carencia de hierro, esto sucede porque los niños en esta etapa tienen necesidades de hierro que no son suplidas en el proceso de crecimiento (OMS, 2015).

Asimismo, la OMS en su informe de “Worldwide prevalence of anemia 1993-2005” estimó que a nivel global la prevalencia de anemia en niños en edad preescolar fue de 47,4%, cabe resaltar que cuando la prevalencia es mayor al 40% se considera un problema de Salud Pública severo, entre 20,0 a 39,9% como moderado, y entre 5,0 a 19,9% como leve (OMS, 2008), lo que convierte a la anemia en un problema de salud pública a nivel mundial. En el Perú la prevalencia de anemia en niños de 6 a 35 meses según la ENDES en el 2020 fue 40,0%, afectando más a los que residen en el área rural (48,4%), viven en la región Sierra (48,5%) y pertenecen al quintil de riqueza inferior (50,5%) (INEI, 2021). Estos datos convierten al Perú como un país con problemas de salud pública severos, donde pese a los grandes esfuerzos (Ministerio de salud [MINSA], 2017) existen dificultades en la implementación de medidas efectivas para poder erradicarla en poblaciones más vulnerables (INEI, 2021).

La alta prevalencia de anemia a nivel mundial afecta principalmente a los países subdesarrollados con consecuencias para la salud humana que comprometería irreversiblemente el desarrollo y el crecimiento de los niños, así como una disminución de la función inmune que lo expone a infecciones, disminución de la capacidad de respuesta y actividad causando una pérdida de productividad cuando

sean adultos y un alto porcentaje de partos prematuros impactando la economía del país (OMS, 2020). Es por esta razón que la OMS ha realizado búsquedas sobre la prevalencia de la anemia en preescolares, sin embargo, existe un gran interés en las causas o variables contribuyentes que deben ser identificadas y abordadas para erradicar la anemia (OMS, 2015; OMS, 2008; Sanou & Ngnie-Teta, 2012; Balarajan et al., 2011). Es así que existen modelos conceptuales planteados en diversos estudios sobre los factores que afectan la anemia. Para el presente trabajo nos basamos en los modelos empleados por Sanou & Ngnie-Teta y Balarajan et al. para la construcción del marco teórico del problema, categorizar y seleccionar las variables independientes (Sanou & Ngnie-Teta, 2012; Balarajan et al., 2011).

Pese a que la anemia se considera como un problema importante de salud pública, las investigaciones científicas sobre anemia en niños menores de 3 años realizadas en el Perú son pocas a comparación de otros países (Ortiz et al., 2021; Shenton, 2020), y generalmente han involucrado tamaños de muestra pequeños, de grupos que no fueron representativos de todo el país, es por eso que se hace necesario el empleo de un algoritmo de “machine learning” para modelos de clasificación, siendo uno de ellos “Random Forest”, que son la misma idea con un giro. Al decidir qué nodo usar para una división, no busca la mejor función para dividir entre todas las funciones posibles. En su lugar, considera un subconjunto aleatorio (muestreo con reemplazo) de funciones y elige la mejor función de ese subconjunto (Wilmott, 2019). Estos algoritmos son útiles cuando las bases de datos son grandes como la del presente estudio.

Al ser los niños menores de tres años un grupo de la edad preescolar con una alta prevalencia de anemia en el Perú (INEI, 2021) como se menciona en el Plan Nacional para la Reducción y Control de la Anemia Materno Infantil y la Desnutrición Crónica Infantil 2017-2021 (MINSA, 2017), se hace necesario la elaboración de estudios que ayuden a los diseñadores de políticas públicas a una mejor toma de decisiones para prevenir la anemia, ya que al hallar el mejor procedimiento alternativo de clasificación de seis propuestos para la predicción de anemia considerando los criterios de balanceo de datos, reajuste de parámetros y selección de variables ayudará a discernir mejor la tenencia de anemia en niños. Por estas razones, el objetivo del estudio fue determinar el mejor procedimiento alternativo de

clasificación sobre la tenencia de anemia en niños de 6 a 35 meses del Perú al aplicar el algoritmo “random forest”.

En el presente trabajo de investigación se ha elaborado y aplicado un modelo de clasificación basado en el algoritmo “random forest”, con la finalidad de predecir la tenencia de anemia (variable respuesta) en niños de 6 a 35 meses del Perú, utilizando la base de datos recolectada por el Instituto Nacional de Estadística e Informática (INEI), a través de la Encuesta Demográfica y de Salud Familiar (ENDES), durante los años 2015 al 2019.

El presente trabajo de tesis se encuentra compuesto en 5 capítulos, en el primer capítulo se aborda la problemática de la investigación que incluye la justificación, delimitación y los objetivos de la investigación. El segundo capítulo está compuesto por el marco teórico que incluye las teorías fundamentales de las variables estudiadas y el marco conceptual. Así mismo, el tercer capítulo está conformado por el marco metodológico del presente trabajo de tesis, el cual incluye las hipótesis de investigación, las variables de estudio, el método de investigación, el diseño de investigación, la población y muestra, así como aspectos de los instrumentos de medición, los procedimientos de recolección de datos y las técnicas de procesamiento y análisis de datos. Finalmente, en el capítulo 4 se presentan los resultados y la discusión, en la cual se presentan tablas y figuras estadísticas, así como la aplicación del algoritmo de clasificación “random forest” en los seis procedimientos alternativos, todo lo anterior sirve para el desarrollo del capítulo 5 que comprende las conclusiones y recomendaciones a las cuales arriba el presente trabajo de tesis.

CAPÍTULO I

PROBLEMA DE INVESTIGACIÓN

1.1. Planteamiento y fundamentación del problema de investigación

Si bien en el Perú durante los últimos años se observa una disminución en la pobreza, la prevalencia de la anemia continúa siendo alta, afectando a 40,0% de los niños de 6 a 35 meses de edad (INEI, 2021). A nivel departamental, se observa un mayor porcentaje de niños menores de 6 a 35 meses de edad con anemia en el departamento de Puno (69,4%), seguido por los departamentos de Ucayali (57,2%) y Madre de Dios (55,0%). Mientras que el departamento con menor anemia es Tacna (29,2%). En cuanto a nivel región natural, la región con mayor anemia es la sierra (48,5%), seguida por la selva (46,3%). Asimismo, por área de residencia, un mayor porcentaje de niños con anemia se encuentra en el área rural (48,4%) (INEI, 2021).

Debido al impacto que ocasiona esta enfermedad en la salud de las personas y en la sociedad, en el año 2017 se aprobó en el país el "Plan nacional para la reducción y control de la anemia materno infantil y desnutrición crónica infantil 2017-2021" (MINSA, 2017), donde se priorizan intervenciones preventivas en niños menores de tres años.

En el año 2018 el gobierno creó: "*el Plan Multisectorial de Lucha Contra la Anemia que establece las acciones e intervenciones efectivas que deben ser implementadas de manera articulada, intersectorial e intergubernamental por las entidades del gobierno nacional, de los gobiernos regionales y los gobiernos locales, así como por la sociedad civil y la comunidad organizada, para la prevención y reducción de la anemia en niños menores de 36 meses. El Plan es implementado en toda la población con énfasis en ámbitos priorizados que muestran las mayores brechas de pobreza y anemia infantil"* (MIDIS, 2018).

Sin embargo, pese a todos los programas que instala el gobierno la anemia sigue avanzando, ya que: *"es un problema estructural que se acentúa por las desigualdades económicas, sociales y culturales, que se manifiestan en pobreza, precariedad de las condiciones de la vivienda (en especial respecto del acceso a*

agua y saneamiento), desconocimiento de las familias sobre la importancia de la alimentación saludable y las prácticas de higiene, entre otros factores. Todo ello atenta contra el desarrollo integral de los niños y, por ende, contra el ejercicio de sus derechos en el presente y en el futuro" (MIDIS, 2018).

Frente a todos los esfuerzos por reducir la anemia se hace necesario su predicción, utilizando variables relacionadas con la salud del niño, la familia, el hogar y la comunidad.

Uno de los problemas más comunes en el ámbito de las ciencias sociales es obtener un buen método de clasificación de la variable respuesta (niño no tiene o tiene anemia), es decir, etiquetar a un sujeto mediante ciertas características que lo describen, dependiendo mucho del objetivo trazado, creando un buen modelo predictivo de clasificación de datos sobre tenencia de anemia en niños que indique probabilísticamente si tiene o no la enfermedad de modo que las instituciones generen programas de prevención. Si se clasifica a un sujeto u objeto, a partir de parámetros o patrones medidos u observados, tal clasificación está asociada a un error, introduciendo una metodología probabilística, que permita medir o cuantificar el error asociado. La pregunta planteada fue: ¿Qué tipo de procedimiento alternativo de clasificación será más efectivo?

1.2. Antecedentes de la investigación

1.2.1. Antecedentes utilizando técnicas estadísticas

Velásquez *et al.* (2015) determinaron los factores sociodemográficos y las características del cuidado materno-infantil asociadas con la anemia en niños de seis a 35 meses de edad en Perú. Utilizaron datos sobre hemoglobina sanguínea registrados en la Encuesta Demográfica y de Salud Familiar (ENDES) 2007-2013, en niños entre los 6 y los 35 meses de edad. Mediante un análisis multivariado de regresión logística identificaron doce factores asociados con la anemia, entre ellos: vivir fuera de Lima y Callao; en un hogar con bajo nivel socioeconómico, madre adolescente, bajo nivel educativo, ser de sexo masculino con edad menor de 24 meses y antecedentes de fiebre reciente, falta de control prenatal en el primer trimestre, falta de suplemento de hierro durante el embarazo o administrado durante un periodo breve, parto en el domicilio, anemia en la madre y ausencia de tratamiento antiparasitario preventivo en el niño.

Moschovis *et al.* (2018) analizaron la anemia a nivel de la población en niños pequeños en África subsahariana (SSA) midiendo el impacto relativo de los factores de riesgo: individuales, maternos y domésticos para la anemia en toda la región. Reunieron datos a nivel de hogar de las encuestas demográficas y de salud realizadas en 27 países de África subsahariana (SSA) entre 2008 y 2014. Realizaron análisis bivariados y multivariados con los modelos de regresión lineal y logística. En los modelos de regresión multivariante, la edad avanzada, el sexo femenino, una mayor riqueza, menos miembros del hogar, mayor estatura para la edad, mayor edad materna, mayor índice de masa corporal materna, embarazo materno actual, mayor hemoglobina en madres, y la ausencia de fiebre reciente se asociaron con mayor hemoglobina en niños con anemia. Los factores demográficos, socioeconómicos, la estructura familiar, el agua / saneamiento, el crecimiento, la salud materna y las enfermedades recientes se asociaron significativamente con la presencia de anemia infantil. Estos grupos de factores de riesgo explican una fracción significativa de anemia (que varía de 1,0% a 16,7%) a nivel poblacional.

Ogunsakin *et al.* (2020) evaluaron y modelaron los determinantes de la prevalencia de la anemia entre los niños de 6 a 59 meses de edad en Nigeria. Utilizaron datos de la Encuesta demográfica y de salud de Nigeria (NDHS) 2018 con modelos de regresión logística binaria de un nivel y de varios niveles para su análisis. Identificaron que el sexo de los niños, la educación de la madre, la religión, el estado de riqueza del hogar, el total de niños nacidos vivos, la edad de los niños, el lugar de residencia y la región tienen un efecto estadísticamente significativo en la prevalencia de anemia.

Shenton *et al.* (2020) identificaron determinantes proximales y distales de anemia grave-moderada y anemia leve relacionados con el nivel socioeconómico, la nutrición y el acceso a la salud. Utilizaron datos de las Encuestas Demográficas y de Salud de Ghana (GDHS) de 2003, 2008 y 2014, evaluaron las probabilidades de anemia grave-moderada y anemia leve en comparación con la ausencia de anemia, en relación con varios factores de riesgo hipotéticos, quienes fueron evaluados utilizando una regresión logística multinomial. Encontraron que para el año 2014, los niños más pequeños (de 6 a 11 meses de edad), que habían tenido fiebre en las 2 semanas anteriores, provenían de familias más pobres y cuyas madres tenían menos educación tenían mayores probabilidades de tener anemia grave o

moderada. Estos resultados siguieron siendo significativos al controlar otros factores de riesgo. Los predictores de anemia en Ghana permanecieron relativamente consistentes entre los tres períodos de tiempo cuando se administró la GDHS.

Ortiz *et al.* (2021) determinaron la prevalencia del nivel de anemia y sus factores asociados en niños menores de tres años utilizando un modelo multicausal en la población peruana. Realizaron un análisis secundario con la base de datos de ENDES 2019. La variable respuesta fue el nivel de anemia. Utilizaron un modelo de regresión ordinal para el análisis de datos. Encontraron que los factores: presencia de diarrea en las últimas dos semanas, edad de 12 meses de vida, no iniciar el control prenatal, sexo masculino, madre con anemia, madre de 15 a 24 años, pozo de tierra como fuente de agua, lengua materna aymara se asociaron al nivel de anemia. Se halló una relación inversa con dos variables: los niños que alguna vez amamantaron y estar en el quintil superior.

1.2.2. Antecedentes utilizando técnicas de “Machine Learning”

El “machine learning” (aprendizaje automático) es una rama de la inteligencia artificial que permite que las máquinas aprendan sin ser expresamente programadas para ello. Es una habilidad indispensable para hacer sistemas capaces de identificar patrones entre los datos con el fin de hacer predicciones.

Khalilia *et al.* (2011) presentaron un método que utiliza la base de datos de la muestra nacional de pacientes hospitalizados (NIS), que están disponibles públicamente a través del Proyecto de utilización y costo de la atención médica (HCUP) para predecir el riesgo de enfermedad de las personas basado en su historial de diagnóstico médico. Dado que los datos de HCUP están no balanceados, emplearon un enfoque de aprendizaje conjunto basado en submuestreos aleatorios repetidos. Esta técnica divide los datos de entrenamiento en múltiples submuestras, a la vez que garantiza que cada submuestra esté completamente balanceada. Compararon el rendimiento de la máquina de soporte vectorial (SVM), “bagging”, “boosting” y “random forest” para predecir el riesgo de ocho enfermedades crónicas: cáncer de mama, diabetes sin complicación, diabetes con complicación, hipertensión, arterioesclerosis coronaria, arterioesclerosis periférica, otras enfermedades circulatorias y osteoporosis. Utilizaron varias variables independientes entre ellas: la edad, la raza, el sexo y 15 categorías de

diagnóstico. En general, en términos del promedio del indicador AUC (área bajo la curva) de la curva ROC (características operativas del receptor) obtenido en todas las enfermedades con el método de aprendizaje del conjunto de “random forest” (89,05%) superó al SVM (86,90%), “bagging” (87,25%) y “boosting” (87,51%). Además, “random forest” tiene la ventaja de calcular la importancia de cada variable en el proceso de clasificación. Los investigadores usaron la medida Gini de disminución media (“Mean Decrease Gini”) para encontrar la importancia de cada variable. Las cuatro variables más importantes en el presente estudio reportadas por el “random forest” fueron: edad, sexo, diabetes con complicación e hipertensión.

Puente *et al.* (2014) presentaron en su investigación métodos rápidos de procesamiento para clasificación en conjuntos de datos no balanceados. Detallan el gran problema que tienen los modelos de clasificación para poder predecir la categoría de interés de la variable respuesta cuando esta está desbalanceada, puesto que los modelos tradicionales para el aprendizaje supervisado siempre otorgan un pronóstico erróneo, ya que el algoritmo aprende de la categoría mayoritaria. La metodología utilizada en la técnica de balanceo como los enlaces Tomek, cuyo tiempo de ejecución es muy reducido con respecto a otras técnicas de balanceo, los resultados fueron comparados utilizando el AUC. Para probar el desempeño del método propuesto, y realizar comparaciones con el método de fuerza bruta, se utilizaron los conjuntos de datos que se encuentran disponibles públicamente en el repositorio Keel http://sci2s.ugr.es/keel/datasets.php.

Thangamani & Sudha (2014) realizaron un análisis de la desnutrición basado en la ingesta de alimentos, el índice de riqueza, el grupo de edad, el nivel de educación, la ocupación, etc. Utilizaron técnicas de machine learning supervisadas: árboles de decisión y redes neuronales artificiales para clasificar el conjunto de datos de la encuesta de salud familiar y las técnicas de clasificación y predicción proporcionan métodos apropiados y flexibles para procesar una gran cantidad de datos para especificar la detección y prevención precisas de la desnutrición en el conjunto de datos de la encuesta. El resultado de las técnicas supervisadas de minería de datos en la base de datos de nutrición proporciona el estado nutricional de los niños menores de cinco años. Encontró un porcentaje de precisión más alto en el modelo perceptrón multicapa (77,17%) presentando una tasa de error más baja (22,83%) y un tiempo de construcción del modelo de 5,77 segundos. Encontró la misma

precisión con el modelo “random forest” (77,17%) presentando una tasa de error más baja (22,83%) y un tiempo de construcción del modelo de 0,02 segundos.

Luo *et al.* (2016) usaron el analito ferritina en una prueba de concepto, extrajeron datos de laboratorio clínico de pruebas de pacientes y aplicaron una variedad de algoritmos de machine learning para predecir los resultados de las pruebas de ferritina usando los resultados de múltiples pruebas. Compararon los resultados previstos con los medidos y revisaron casos seleccionados para evaluar el valor clínico de la ferritina prevista. Demostraron que los datos demográficos de los pacientes y los resultados de múltiples pruebas de laboratorio pueden discriminar los resultados de ferritina normales de los anormales con un alto grado de precisión (área bajo la curva (AUC) de hasta 0,97, datos de prueba retenidos). La revisión de casos indicó que los resultados de ferritina pronosticados a veces pueden reflejar mejor el estado de hierro subyacente que la ferritina medida.

Jaiswal *et al.* (2019) investigaron algoritmos supervisados de Machine Learning (Naive Bayes, “random forest” y algoritmo de árbol de decisión) para la predicción de anemia utilizando datos del hemograma completo (CBC) recopilados de centros de patología. Los resultados muestran que la técnica de Naive Bayes supera en términos de precisión (96,0909) en comparación con C4.5 (95,4602) y “random forest” (95,3241).

Khan *et al.* (2019) consideraron los algoritmos de “machine learning” para predecir el estado de anemia entre los niños menores de cinco años utilizando factores de riesgo comunes como características. Realizaron una evaluación sistemática de los algoritmos en términos de precisión, sensibilidad, especificidad y área bajo la curva (AUC). Encontraron que el algoritmo “random forest” logró la mejor precisión de clasificación (68,53%) con una sensibilidad de 70,73%, especificidad de 66,41% y AUC de 0,6857. Por otro lado, el algoritmo clásico de Regresión Logística alcanzó una precisión de clasificación del 62,75% con una sensibilidad del 63,41%, especificidad del 62,11% y AUC de 0,6276. Entre todos los algoritmos considerados, el K-NN dio menores rendimientos con precisión, sensibilidad y especificidad de 61,95%, 65,85% y 58,20%, respectivamente. Concluyeron que los métodos “machine learning” se pueden considerar además de las técnicas de regresión clásicas cuando la predicción de la anemia es el enfoque principal.

Çil *et al.* (2020) desarrollaron un sistema de apoyo a la toma de decisiones para distinguir entre la enfermedad de β-talasemia (β-TT) y deficiencia de hierro (IDA), ambos son síntomas de la anemia. Propusieron un sistema donde se utilizaron los algoritmos de clasificación: Regresión logística, K-vecinos más cercanos, Máquina de soporte vectorial, Máquina de aprendizaje extremo y Máquina de aprendizaje extremo regularizado. El rendimiento de la clasificación se evaluó con los parámetros de precisión, sensibilidad, f-medida y especificidad utilizando los parámetros hemoglobina, glóbulos rojos, HCT (hematocrito), MCV (Volumen Celular Medio), MCH (hemoglobina celular media), MCHC (concentración media de hemoglobina celular) y RDW (ancho de distribución de glóbulos rojos) obtenidos de 342 pacientes. Se obtuvo una precisión del 96,30% para el sexo femenino, del 94,37% para el masculino y del 95,59% en la coevaluación de pacientes masculinos y femeninos.

Ayyildiz & Tuncer (2020) realizaron un diagnóstico diferencial de la anemia por deficiencia de hierro (IDA) y β-talasemia mediante el uso de índices de recuento de glóbulos rojos (RBC) y técnicas de “machine learning” que incluyen Máquina de soporte vectorial (SVM) y K-Vecinos más cercanos (KNN). Los índices de RBC se utilizaron como parámetros de entrada para el clasificador y los rendimientos de SVM y KNN se evaluaron por separado para determinar la efectividad de ambas técnicas. Se proporcionaron menos parámetros como entradas a los algoritmos de “machine learning” y se logró un mayor rendimiento. Por otro lado, se utilizó una técnica de selección de características, el algoritmo de selección de características de análisis de componentes de vecindario (NCA), para seleccionar características de los conjuntos de datos, y los parámetros seleccionados a través de NCA proporcionaron un alto rendimiento (97% de área bajo la curva ROC [AUC]). Los indices RBC que utilizaron mostraron un rendimiento superior en comparación con los informados en la literatura. Además, el estudio reveló que diferentes parámetros de CBC fueron eficientes para distinguir entre IDA y β-talasemia en pacientes masculinos y femeninos.

1.3. Formulación del problema de investigación

Considerando los criterios de balanceo de datos y reajuste de parámetros al aplicar el algoritmo “random forest” ¿Cuál es el mejor procedimiento para un modelo de clasificación sobre la predicción de la tenencia de anemia de niños de 6 a 35 meses del Perú?

1.4. Delimitación del estudio

Delimitación de las variables: Para el presente estudio se contó con datos que corresponden a la medición de la variable dependiente: tenencia de anemia y a 33 variables independientes organizadas en tres grupos: a) variables sociodemográficas (área de residencia, altitud de la ciudad donde vive el niño, región natural, índice de bienestar o riqueza al que pertenece el hogar, edad materna, grado de instrucción de la madre, lengua materna de la madre, conexión domiciliaria de agua potable, conexión domiciliaria de desagüe, material predominante del piso de vivienda); b) variables relacionadas con el niño (sexo, edad, número de niños menores de cinco años en el hogar, número de personas que viven en el hogar, bajo peso al nacer (< 2500 gr), orden de nacimiento, intervalo entre nacimientos, signos y síntomas (fiebre) en las dos semanas previas, diarrea durante las dos últimas semanas, tos y respiración rápida durante las dos últimas semanas), y c) variables del cuidado materno e infantil (control prenatal, control prenatal en primer trimestre, parto institucional, talla de la madre, diagnóstico de anemia en la madre en el momento de la encuesta, tiempo de consumo de suplemento de hierro en la gestación, suplemento de vitamina A, niño recibió hierro en pastillas o jarabes, medicación antiparasitaria en el niño (tratamiento antiparasitario los últimos seis meses), el niño comió algún tipo de carne el día de ayer (res, pollo, hígado, cerdo, etc.), consumo de agua hervida y control de crecimiento y desarrollo (CRED)) para todas las regiones del Perú.

Delimitación temporal: Los datos se encuentran comprendidos entre los años 2015 al 2019.

1.5. Justificación e importancia de la investigación

Dentro del contexto de la técnica del “machine learning,” el algoritmo “Random Forest” tiene múltiples aplicaciones exitosas en temas económicos, negocios, salud, entre otros, pero hay muy pocas aplicaciones para la determinación de la tenencia

de anemia, en general. Grandes empresas tales como Facebook, Google o Amazon, tienen áreas dedicadas al “Machine Learning” Amazon, por ejemplo, ha llevado la minería de datos a todas sus áreas haciendo predicciones de la demanda de sus productos, fijación de precios idóneos, recomendaciones personalizadas, optimización las rutas de distribución o inclusive detección fraude (Lange, 2016).

De acuerdo a nuestra revisión de la literatura, el algoritmo “Random Forest” no se ha aplicado para estudiar la tenencia de anemia de los niños del Perú.

La importancia de la presente investigación en nuestro medio radica en que el algoritmo “Random Forest” permite conocer los factores que se asocian a la anemia en niños para que el estado pueda tomar las medidas necesarias para disminuir la incidencia y complicaciones futuras de esta enfermedad en el desarrollo de la niñez dado que la calidad del capital humano es una base fundamental para el óptimo desarrollo socioeconómico de un país y este depende de las condiciones de salud y nutrición de la población, ya que la presencia de anemia en un niño afecta el proceso de desarrollo, por sus implicaciones funcionales en el individuo, expresada en una disminución de su rendimiento físico, capacidad de aprendizaje, productividad y desgaste en la salud. Además, la presente investigación sirve como guía a estudiantes e investigadores interesados en realizar investigación sobre “Machine Learning” aplicado a ámbitos de la salud.

El algoritmo propuesto en la presente investigación tiene como particularidad que se ejecutan varios algoritmos de árbol de decisiones en lugar de uno solo. Para clasificar un nuevo objeto basado en atributos, cada árbol de decisión da una clasificación y finalmente la decisión con mayor número de “votos” es la predicción del algoritmo.

Con el presente estudio se espera ayudar a los diseñadores de políticas públicas a una mejor toma de decisiones para prevenir la anemia, ya que al aplicar un algoritmo de clasificación mediante técnicas de balanceo y reajuste de parámetros ayudará a discernir mejor la tenencia de anemia en niños.

1.6. Objetivos de la investigación

Objetivo general

Determinar el mejor procedimiento considerando los criterios de balanceo de datos y reajuste de parámetros al aplicar el algoritmo “random forest” para un modelo de clasificación sobre la predicción de la tenencia de anemia de niños de 6 a 35 meses del Perú.

Objetivos específicos

- Describir las variables sociodemográficas, relacionadas con el niño y del cuidado materno-infantil según tenencia de anemia en niños.
- Calcular y comparar los indicadores del Área Bajo la Curva (AUC), especificidad y sensibilidad que permiten predecir la tenencia de anemia en niños de 6 a 35 meses del Perú a partir de las tablas de clasificación obtenidas con el algoritmo “random forest”.
- Identificar las variables más importantes de acuerdo a los resultados de la aplicación de algoritmo “random forest”.

CAPÍTULO II

MARCO TEÓRICO

2.1. Fundamentos teóricos de la investigación

2.1.1. Factores asociados a la anemia

Los términos anemia, deficiencia de hierro y anemia por deficiencia de hierro a menudo se usan indistintamente, pero no son equivalentes. La anemia se define como una reducción significativa en la concentración de hemoglobina, el hematocrito o el número de glóbulos rojos circulantes a un nivel inferior al que se considera normal por edad, sexo, estado fisiológico y altitud, sin considerar la causa de la deficiencia (Nestel et al., 2002). La anemia por deficiencia de hierro es una afección en la que hay anemia debido a la falta de hierro disponible para apoyar la producción normal de glóbulos rojos. Es la tercera y última etapa de la deficiencia de hierro que comienza con el agotamiento de las reservas de hierro, como se refleja en una concentración reducida de ferritina en suero. La segunda etapa es la eritropoyesis deficiente en hierro, caracterizada por una disminución de la cantidad de hierro sérico, la saturación de transferrina y la concentración de ferritina sérica, pero con una concentración de hemoglobina normal. Debido a que la anemia puede surgir de factores nutricionales y no nutricionales, se usan varios términos para clasificar la anemia, que incluyen anemia nutricional, anemia de infección, anemia de enfermedades crónicas, anemia perniciosa.

Varios factores contribuyen simultáneamente en la anemia infantil, pero sus relaciones con la aparición de la anemia no son idénticas. Por lo tanto, desde una perspectiva epidemiológica, es importante distinguir entre los diferentes factores. Un factor causal está relacionado con la aparición de una enfermedad o la condición y precede a la enfermedad. Un factor de riesgo es un elemento vinculado a una persona (biológica o hereditaria), un comportamiento, estilo de vida o entorno que aumenta la probabilidad de desarrollar la afección y se ha encontrado que está correlacionado con la afección en estudios epidemiológicos (Last, 2004). Cuando una intervención dirigida a un factor puede reducir la probabilidad de que se desarrolle la enfermedad, el factor se considera un factor de riesgo modificable. Un factor susceptible de aumentar la aparición de una condición patológica es un factor

determinante o determinante. Por ejemplo, los principales factores causales de la deficiencia de hierro que conducen a la anemia son un bajo consumo de hierro en la dieta, una inadecuada absorción de hierro, una pérdida crónica de sangre y una mayor demanda de hierro. Sin embargo, hay otros factores (relación no causal) que contribuyen a la anemia, como son: factores socioculturales: pobreza, factores maternos: enfermedades crónicas secundarias al SIDA, tuberculosis y factores genéticos: células falciformes y talasemia. Existen varios niveles de estratificación de los factores de riesgo de anemia en los niños, incluidos los factores estructurales y ambientales, los factores a nivel comunitario, los factores a nivel del hogar y los factores individuales relacionados con la salud y la nutrición. La Figura 1 resume los factores de riesgo multinivel de la anemia en niños de países en desarrollo. Existe una perspectiva antropológica que puede verse como un factor de riesgo transversal (Sanou & Ngnie-Teta, 2014).

Perspectiva antropológica

Los antropólogos creían que la revolución agraria que provocó cambios en los comportamientos dietéticos y el brote de enfermedades infecciosas hace unos 10000 años ha jugado un papel importante en la aparición y propagación de la deficiencia de hierro y la anemia (Denic & Agarwal, 2007; Wander et al., 2009). Según esta teoría, la carne era la principal fuente de energía antes de la revolución agraria. Cuando los humanos pasaron de la caza a la agricultura, la dieta se volvió deficiente en hierro biodisponible, lo que aumentó la prevalencia de la deficiencia de hierro y su posterior anemia. El cultivo de alimentos a base de plantas ha aumentado la ingesta de calorías, pero ha reducido el consumo de carne. Como resultado, la ingesta de hierro se volvió insuficiente para cumplir con los requisitos diarios individuales. Según Mann (2007), la ingesta diaria total de hierro disminuyó de 87 mg en la edad paleolítica a 15 mg en el siglo XX. Además, el mayor consumo de alimentos de origen vegetal ha reducido la ingesta de hierro absorbible porque la cantidad de hierro no heminico y los inhibidores de la absorción de hierro han aumentado en la dieta, mientras que la cantidad de hierro heminico ha disminuido.

Con la sedentarización y la cría de animales, los portadores de enfermedades infecciosas pudieron transmitirse de los animales a los humanos y provocar enfermedades infecciosas humanas emergentes o reemergentes. Posteriormente, las malas condiciones ambientales e higiénicas, el hacinamiento y los cambios en el

estilo de vida han resultado en la proliferación y propagación de estos portadores (Denic & Agarwal, 2007). Varios estudios sugirieron que la deficiencia de hierro leve a moderada puede proteger contra la infección aguda (Oppenheimer, 2001; Prentice, 2008; Sazawal et al., 2006). Por lo tanto, algunos autores presentan la hipótesis de una posible adaptación metabólica durante la cual el cuerpo humano autorregula su hierro a un estado de deficiencia, el «fenotipo deficiente en hierro», para prevenir la gravedad de las infecciones cuando la reinfección es un proceso continuo (Denic & Agarwal, 2007). Según estos autores, el avance importante en los países desarrollados para controlar la anemia es más probable debido a la erradicación exitosa de las infecciones que a la calidad de la dieta. En áreas endémicas de malaria como África, el fenotipo de deficiencia de hierro sobrevivió mejor con el tiempo (Denic & Agarwal, 2007; Wander et al., 2009). Por lo tanto, la terapia de sustitución de hierro en algunos grupos de población, como la suplementación con hierro en niños sin deficiencia funcional de hierro, puede causar más daño que bien (Sazawal et al., 2006; OMS / UNICEF, 2006).

Factores dietéticos

Los factores de riesgo dietéticos para la anemia infantil en los países en desarrollo incluyen la deficiencia única o combinada de micronutrientes como hierro, ácido fólico, vitamina B6, vitamina B12, vitamina A y cobre. Se ha encontrado asociación entre anemia y deficiencia de vitamina A, riboflavina, proteínas y otros nutrientes (Gamble et al., 2004; Semba & Bloem, 2002; Thorandenya et al., 2006; Rock et al., 1988). Aunque se cree que los factores nutricionales son los más importantes que contribuyen a la anemia infantil, su contribución exacta al riesgo de anemia no está bien establecida y puede variar con el nivel de infección y la calidad de la dieta. Magalhaes & Clements (2011) estimaron que alrededor del 37% de los casos de anemia en niños en edad preescolar en tres países de África occidental, a saber; Burkina Faso, Ghana y Malí podrían evitarse tratando únicamente los factores relacionados con la nutrición.

Sanou & Ngnie-Teta (2012) proponen un nuevo marco teórico para el análisis de los factores de riesgo de la anemia en niños en edad preescolar a partir del marco teórico propuesto por Ngnie-Teta et al. (2007) que refleja la justificación de un análisis multinivel. Las variables las definen jerárquicamente, especialmente a nivel individual, familiar y comunitario (figura 1).

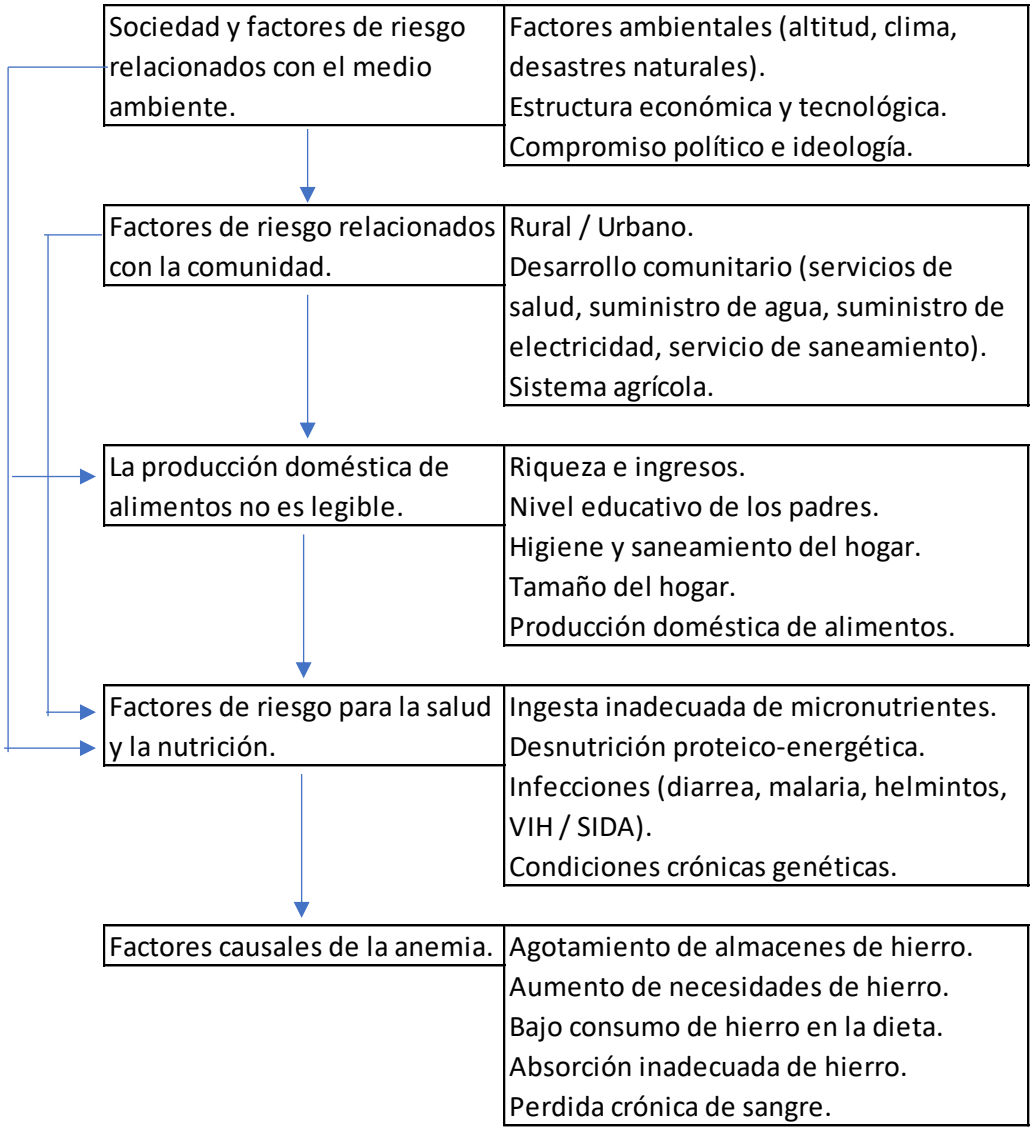


Figura 1. *Modelo para el estudio de la anemia propuesto por Sanou & Ngnie-Teta (2012).*

En la literatura internacional se plantea otro modelo causal de la anemia el cual es adoptado en el Perú por Zavaleta y Irizarry (Balarajan et al., 2011) (Figura 2). Entre las causas inmediatas se reconoce el consumo inadecuado de hierro y de otros micronutrientes a partir de los alimentos. Esta carencia de hierro y vitaminas no permitiría una apropiada formación de los glóbulos rojos y de la hemoglobina. Otras causas inmediatas de la anemia son la alta morbilidad por infecciones como la diarrea, parasitosis, malaria, etc. Esta situación está asociada a inadecuadas prácticas de higiene, de lavado de manos, limitado acceso a agua segura y saneamiento básico. Se reconoce también que la vitamina A, la vitamina B2, las vitaminas B6, B12 y el Ácido Fólico intervienen en la formación de los glóbulos rojos

en la médula ósea. Las vitaminas A, C y Riboflabina favorecerían la absorción del hierro a nivel intestinal, cumpliendo un rol movilizador del mineral a partir de las reservas; mientras que las vitaminas C y E tienen una función antioxidante para la protección de los glóbulos rojos.

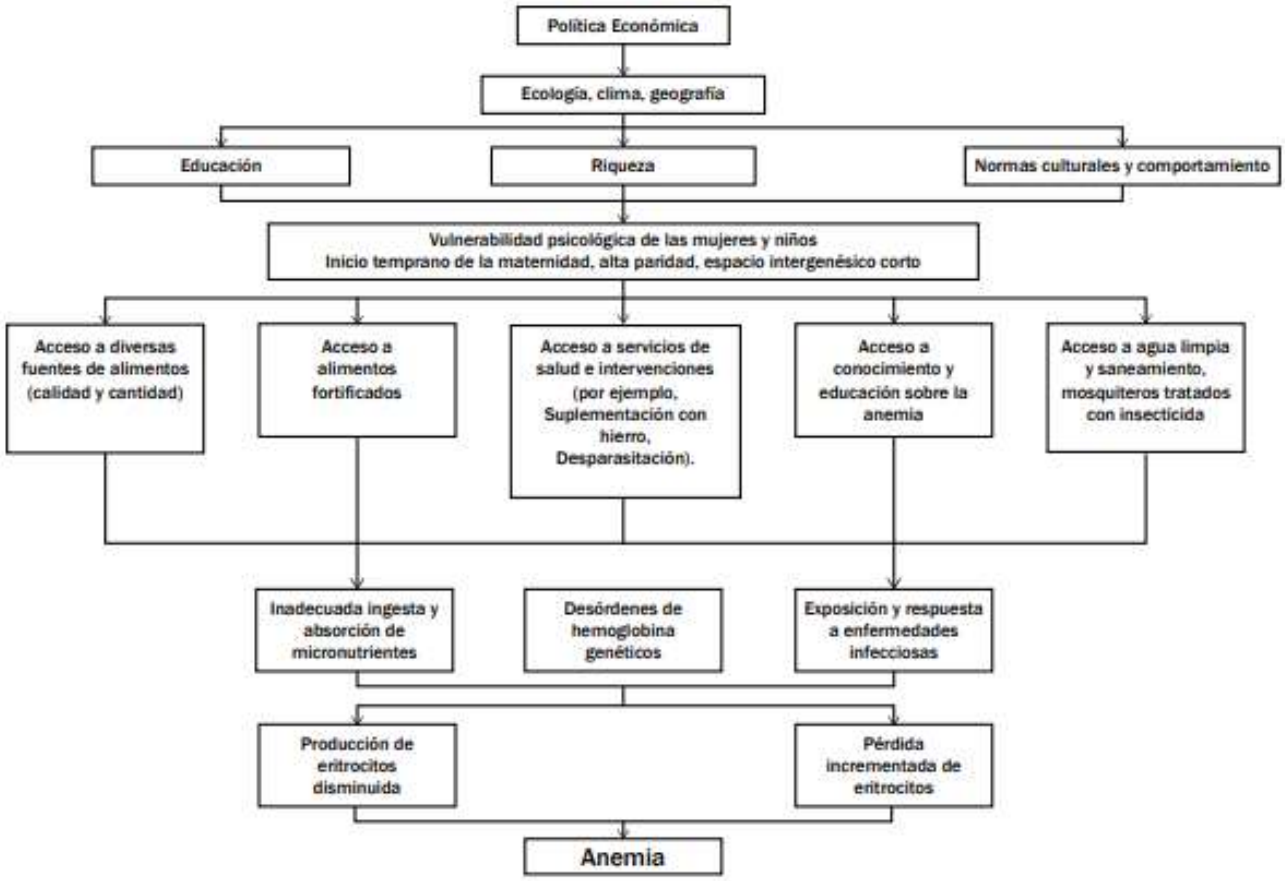


Figura 2. *Modelo causal de la anemia propuesto por Balarajan et al. (2011).*

2.1.2. “Machine learning”

El “machine learning” (aprendizaje automático) es una rama de la inteligencia artificial cuyo fin es dotar a una máquina, a través de algoritmos, la capacidad de entrenar y aprender en base a datos, sin ser explícitamente programada, imitando la capacidad que tienen las personas de aprender mediante ejemplos sin recurrir a fórmulas ni reglas entre variables y posibilitando al término del entrenamiento un modelo que permita la generalización; es decir, la obtención de resultados en nuevas situaciones no conocidas durante el aprendizaje. Este comportamiento es muy importante porque permite resolver situaciones en donde no existe o es muy difícil encontrar una fórmula que facilite una respuesta exacta a partir del

conocimiento de ciertas variables. De esta manera el modelo podrá, por ejemplo, reconocer y clasificar nuevas imágenes si se ha "entrenado" en un conjunto de casos o ejemplos para los cuales se han presentado determinadas características y la imagen correspondiente (Véliz, 2018).

El “machine learning” es esa rama de la informática que otorga a la Inteligencia Artificial la capacidad de aprender tareas. Para lograrlo, los programadores se basan en los algoritmos del “machine learning”. Dentro de estos algoritmos están los árboles de decisión.

2.1.2.1. Arboles de decisión

Son modelos que se basan en la partición recursiva de un conjunto de datos, lo que da como resultado subgrupos o patrones llamados nodos, que al ser representados gráficamente asemejan la forma de árboles. Estos modelos, que involucran relaciones si – entonces y cuyo procedimiento heurístico está relacionado con el dicho popular “divide y vencerás”, fueron creados por Breiman et al. (1984) (Véliz, 2018).

> *“Los árboles de decisión son modelos en los cuales se pueden usar variables predictoras numéricas o categóricas, son robustos a los datos atípicos y son insensibles a transformaciones monótonas de estas variables. Si la variable respuesta es categórica, el árbol es de clasificación; si es continua, el árbol es de regresión”.* (Véliz, 2018)

Los métodos basados en árboles proporcionan un mecanismo simple, intuitivo y poderoso tanto para la regresión como para la clasificación. La idea principal es dividir un espacio de características χ (potencialmente complicado) en regiones más pequeñas y ajustar una función de predicción simple a cada región. Por ejemplo, en un entorno de regresión, se podría tomar la media de las respuestas de entrenamiento asociadas con las funciones de entrenamiento que caen en esa región específica. En la configuración de clasificación, una función de predicción de uso común toma el voto mayoritario entre las variables de respuesta correspondientes. Comenzamos con un ejemplo de clasificación simple (Kroese et al, 2020).

2.1.2.1.1. Árboles de clasificación

“Se define un árbol de clasificación como una estructura en forma de diagramas de construcciones lógicas en las que las ramas representan conjuntos de decisión. Estas decisiones generan sucesivas reglas para la clasificación de un conjunto de datos en subgrupos disjuntos y exhaustivos. Las ramificaciones se realizan en forma recursiva hasta que se cumplen ciertos criterios de parada.

El objetivo de estos métodos es obtener individuos más homogéneos con respecto a la variable que se desea discriminar dentro de cada subgrupo y heterogéneos entre los subgrupos. Para la construcción del árbol se requiere información de variables explicativas o predictoras (independientes) a partir de las cuales se va a realizar la discriminación de la población en subgrupos.”

El panel izquierdo de la figura 3 muestra un conjunto de entrenamiento de 15 puntos bidimensionales (características) que se dividen en dos clases (rojo y azul). ¿Cómo debe clasificarse el nuevo vector de características (punto negro)?

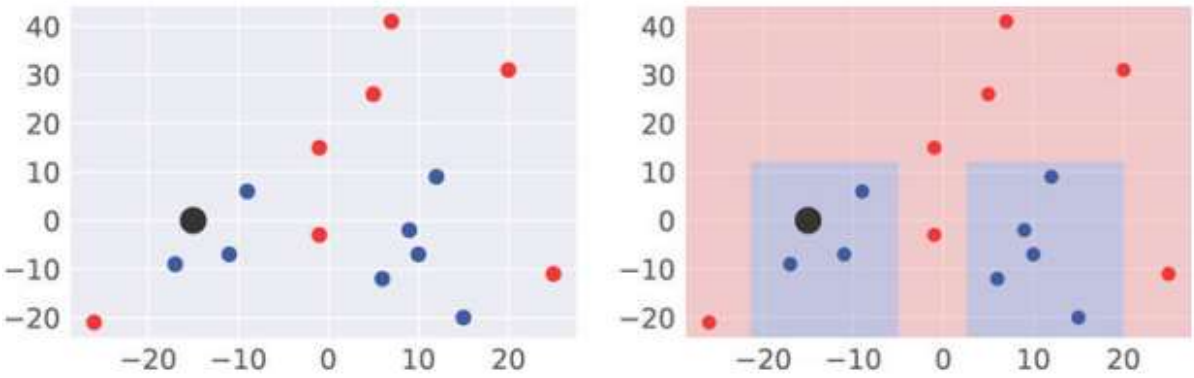


Figura 3. *Representación para la clasificación de un nuevo vector de características usando arboles de decisión.*
Izquierda: datos de entrenamiento y una nueva característica. Derecha: una partición del espacio de funciones.

No es posible separar linealmente el conjunto de entrenamiento, pero podemos dividir el espacio de características $\chi = \mathbb{R}^2$ en regiones rectangulares y asignar una clase (color) a cada región, como se muestra en el panel derecho de la Figura 3. Los puntos en estas regiones se clasifican como azules o rojos. La partición define así un clasificador (función de predicción) g que asigna a cada vector de características x una clase “roja” o “azul”. Por ejemplo, para $x = [-15, 0]^T$ (punto

negro sólido), g(x) = “azul”, ya que pertenece a una región azul del espacio de características.

Tanto el procedimiento de clasificación como la partición del espacio de características pueden representarse convenientemente mediante un *árbol de decisión* binario. Este es un árbol donde cada nodo v corresponde a una región (subconjunto) R_v del espacio de características χ — el nodo raíz correspondiente al propio espacio de características.

Cada nodo interno v contiene una condición lógica que divide a R_v en dos subregiones separadas. Los nodos hoja (los nodos terminales del árbol) no se subdividen, y sus regiones correspondientes forman una partición de χ, ya que son disjuntos y su unión es χ. Asociado con cada nodo hoja w también hay una función de predicción regional g^w en R_w.

La evaluación de las condiciones lógicas a lo largo del camino del árbol eventualmente nos llevará a un nodo hoja y su región asociada. En este caso, el proceso termina en una hoja que corresponde a la región azul izquierda en el panel derecho de la Figura 3.

De manera más general, un árbol binario T dividirá el espacio de características χ en tantas regiones como nodos hoja. Denote el conjunto de nodos hoja por W. La función de predicción general g que corresponde al árbol se puede escribir como:

$$g(x) = \sum_{\omega \epsilon W} g^{\omega}(x)\ \mathbf{1}\{x \epsilon R_{\omega}\} \qquad (1)$$

donde **1** denota la función indicadora. La representación (1) es muy general y depende de (a) cómo se construyen las regiones {Rw} a través de las condiciones lógicas en el árbol de decisión, así como (b) cómo se definen las *funciones de predicción regionales* de los nodos hoja (Kroese et al, 2020).

“Entre las ventajas de esta técnica no paramétrica de clasificación están las siguientes:

- Las reglas de asignación son legibles y por tanto la interpretación de resultados es directa e intuitiva.
- Tiene en cuenta las interacciones que pueden existir entre los datos.
- Es robusta frente a datos atípicos o individuos mal etiquetados.

- Es válida para variables explicativas de naturaleza: continua, nominal u ordinal.
- Los árboles requieren grandes masas de datos para asegurarse que la cantidad de observaciones de los nodos terminales es significativa.

Por el contrario, este método de clasificación de datos tiene las siguientes desventajas:

- Las reglas de asignación son fuertes y bastante sensibles a ligeras perturbaciones de los datos.
- Dificultad para elegir el árbol “óptimo””. (Sucari, 2018)

Se ha creado una serie de algoritmos para la construcción de los árboles de clasificación, entre los más usados están:

- El C5.0, creado por Quinlan (1992). Este algoritmo se aplica para todo tipo de variables predictoras.
- El CART (“Classification And Regression Tree”), creado por Breiman et al. (1984). Este algoritmo se ha creado para árboles binarios y para todo tipo de variables predictoras. Con el CART las particiones de cada nodo se hacen en forma recursiva hasta que se alcanza un criterio de parada.
- El CHAID (“Chi-square Automatic Interaction Detection”, Detección de Interacción Automática de Chi Cuadrado) fue desarrollado por Kass (1980) y se aplica con variables independientes categóricas, usan como criterio de parada, la prueba chi-cuadrado.

2.1.2.1.1.1. Algoritmo CHAID

“El algoritmo CHAID divide en grupos los registros de datos que presenten la misma probabilidad de los resultados, basándose en los valores de las variables independientes. El algoritmo parte de un nodo raíz y se va bifurcando en nodos descendientes hasta llegar a los nodos hoja, donde finaliza la ramificación.

La ramificación puede ser binaria, ternaria, etc. y viene determinada por la prueba Chi-cuadrado. Esta prueba se lleva a cabo mediante una tabulación cruzada entre el resultado y cada una de las variables independientes. El resultado es la probabilidad de que la hipótesis nula sea correcta, estas probabilidades se clasifican, y si el

mejor (el valor más pequeño) se encuentra bajo un umbral determinado, se realiza una ramificación del nodo raíz en esa ubicación.

Pasos para elaborar un Árbol de Clasificación mediante el algoritmo CHAID

Pasos del Algoritmo CHAID en el cual se desea clasificar la variable respuesta Y y se tiene como variables independientes X_1, X_2,…, X_k :

1. Calcular la distribución de la variable respuesta Y en el nodo raíz.
2. Para cada variable independiente X_i (i = 1, 2, …, k), hay que encontrar el par de categorías que tienen menores diferencias significativas respecto a la distribución de Y dentro del nodo. Es decir, aquel que tiene el mayor p-valor. Para calcular dicho p-valor depende del tipo de variables que estemos tratando en cada momento.
   a. La relación entre la variable independiente X_i y la variable respuesta Y dentro del nodo se representa mediante una tabla de contingencia. Se consideran todas las subtablas de contingencia posibles que se puedan formar con dos categorías de la variable independiente.
   b. El algoritmo identifica el par de categorías de X_i con mayor p-valor (p_i) asociado y lo compara con el nivel α predeterminado, normalmente α_{union} = 0.05. Si el p-valor p_i es mayor que el valor de α_{union} se agrupan dichas categorías. Se repite el literal a, considerando el par de categorías agrupadas como una única para calcular las subtablas de contingencia. En el caso de no superar el valor de la α_{union} no se realiza ninguna agrupación de las categorías y se pasa al numeral 3.
   c. De nuevo se selecciona el par de categorías con mayor p-valor y se compara con el valor α_{union}. Si es superior se vuelve a agrupar y se vuelve a calcular las subtablas de contingencia. El proceso termina en el caso en el cual el p-valor es inferior a α_{union} o se llega a dos categorías.
   d. El algoritmo calcula un p-valor ajustado empleando las categorías agrupadas obtenidas de X_i y la categoría Y usando el ajuste de Bonferroni.
3. Los pasos del literal a) al d) se repiten de nuevo con el resto de variables independientes.

4. El paso final es dividir el nodo basado en la variable independiente, con las categorías agrupadas, con el menor p-valor ajustado si el valor es menor que el prefijado $\alpha_{separacion}$. En el caso de obtener un valor superior dicho nodo no se ramifica y será un nodo terminal.
5. Se continúa ramificando el árbol hasta que se satisfaga el criterio de parada". (Sucari, 2018)

2.1.2.1.2. El método "bagging"

Es un método que podemos utilizar para reducir la varianza de los modelos CART, a veces denominado "*bootstrap aggregating"*.

Cuando creamos un único árbol de decisión, solo usamos un conjunto de datos de entrenamiento para crear el modelo.

Sin embargo, el "bagging" utiliza el método siguiente:

1. Tome b muestras de arranque del conjunto de datos original.
    - Recuerde que una muestra de arranque es una muestra del conjunto de datos original en el que las observaciones se toman con reemplazo.
2. Cree un árbol de decisión para cada muestra de arranque.
3. Promedie las predicciones de cada árbol para llegar a un modelo final.
    - Para los árboles de regresión, tomamos el promedio de la predicción realizada por los árboles B.
    - Para los árboles de clasificación, tomamos la predicción más comúnmente que ocurre hecha por los árboles B.

El "bagging" se puede usar con cualquier algoritmo de "machine learning", pero es particularmente útil para los árboles de decisión porque inherentemente tienen una varianza alta y el "bagging" puede reducir drásticamente la varianza, lo que conduce a un menor error de prueba.

Para aplicar el "bagging" a los árboles de decisión, cultivamos árboles individuales B profundamente sin podarlos. Esto da como resultado árboles individuales que tienen una varianza alta, pero un sesgo bajo. Luego, cuando tomamos las predicciones promedio de estos árboles, podemos reducir la varianza.

Considere un escenario idealizado para un árbol de regresión, donde tenemos acceso a B copias independientes e idénticamente distribuidos (i.i.d) T_1,…, T_B de un conjunto de entrenamiento T. Luego, podemos entrenar B modelos de regresión separados (B diferentes árboles de decisión) usando estos conjuntos, dando a los aprendices g_{T_1},...,g_{T_B}, y tomar su promedio:

$$\bar{g}(x) = \frac{1}{B}\sum_{b=1}^{B} g_{T_b}(x) \qquad (2)$$

Por la ley de los grandes números, cuando B → ∞, la función de predicción promedio converge a la función de predicción esperada $\mathbb{E}_{g_T}$

En la práctica, el rendimiento óptimo suele producirse con 50 a 500 árboles, pero es posible ajustar miles de árboles para producir un modelo final.

Es necesario tener en cuenta que ajustar más árboles requerirá más potencia computacional, lo que puede o no ser un problema dependiendo del tamaño del conjunto de datos.

La ventaja del “bagging” es que normalmente ofrece una mejora en la tasa de error de prueba en comparación con un único árbol de decisión.

La desventaja es que las predicciones de la colección de árboles “bagging” pueden estar altamente correlacionadas si hay un predictor muy fuerte en el conjunto de datos.

En este caso, la mayoría o todos los árboles “bagging” utilizarán este predictor para la primera división, lo que dará como resultado árboles que son similares entre sí y tienen predicciones altamente correlacionadas.

Específicamente, para algún vector de características x, sea $Z_b = g_{T_b}(x)$, b = 1,2,…, B valores de predicción i.i.d., obtenidos de conjuntos de entrenamiento independientes T_1,…, T_B. Supongamos que Var $Z_b = \sigma^2$ para todo b = 1,…, B. Entonces la varianza del valor medio de predicción $\bar{Z}_B$ es igual a σ^2/B. Sin embargo, si se utilizan conjuntos de datos Bootstrap $\{T_b^*\}$ en su lugar, las variables aleatorias correspondientes $\{Z_b\}$ estarán correlacionadas. En particular, $Z_b = g_{T_b^*}(x)$para b = 1,…, B están distribuidas idénticamente (pero no son independientes) con alguna correlación positiva por pares ϱ. Entonces se sostiene que:

$$\mathbb{V}ar\,\bar{Z}_B = \varrho\sigma^2 + \sigma^2\frac{(1-\varrho)}{B} \qquad (3)$$

Mientras que el segundo término de (3) tiende a cero a medida que aumenta el número de observaciones B, el primer término permanece constante.

Este problema es particularmente relevante para bagging con árboles de decisión. Por ejemplo, considere una situación en la que existe una función que proporciona una muy buena división de los datos. Tal característica se seleccionará y dividirá para cada $\left\{g_{T_b^*}\right\}_{b=1}^{B}$ en el nivel raíz y, en consecuencia, terminaremos con predicciones altamente correlacionadas. En tal situación, el promedio de predicción no introducirá la mejora deseada en el rendimiento del predictor bagged (Kroese et al, 2020).

Una forma de evitar este problema es utilizar “random forest” (bosque aleatorio), que utilizan un método similar al del “bagging”, pero que son capaces de producir árboles descorrelacionados, lo que a menudo conduce a tasas de error de prueba más bajas (Statology, 2020).

La idea principal de “random forest” es realizar el bagging en combinación con una "descorrelación" de los árboles al incluir solo un subconjunto de características durante la construcción del árbol. Para cada conjunto de entrenamiento de Bootstrap T_b^* construimos un árbol de decisión utilizando un subconjunto seleccionado al azar de m $\leqslant$ p características para las reglas de división. Esta idea simple pero poderosa descorrelacionará los árboles, ya que los predictores fuertes tendrán menos posibilidades de ser considerados en los niveles de raíz (Kroese et al, 2020).

2.1.2.1.3. El método “Random forest” (bosque aleatorio)

Según Breiman (2005) los bosques aleatorios son una combinación de árboles de decisión de manera que cada uno depende de los valores de un vector aleatorio muestreado de forma independiente y con la misma distribución para todos los árboles del bosque. Específicamente, un bosque aleatorio es un clasificador que consta de una colección de clasificadores estructurados en árbol tal como lo muestra en la ecuación (4):

$$\{h(x, \theta_k), k = 1, \dots\} \qquad (4)$$

donde $\{\theta_k\}$ son vectores estocásticos, independientes e idénticamente distribuidos (i.i.d) y cada árbol genera un voto unitario para la clase más popular en la entrada x. En el bosque aleatorio, cada árbol está completamente desarrollado, lo que significa que no se emplean métodos de poda. El usuario selecciona la cantidad de características a considerar en cada nodo y la cantidad de árboles para crecer. Por lo tanto, en cada nodo, solo se buscan las entidades seleccionadas para obtener la mejor división. Cada nueva instancia se transmite a cada uno de los N árboles. El bosque elige la clase que tiene más N votos para ese caso (Pal, 2005).

La definición general de bosques aleatorios, dada por Breiman (2001), es la siguiente:

Definición ("*Random Forest*") (*Bosque aleatorio*). Sea ($\hat{h}(., \Theta_1)$, . . . , $\hat{h}(., \Theta_q)$) una colección de predictores de árbol, con $\Theta_1,...,\Theta_q$ q variables aleatorias independientes i.i.d de L_n. El predictor "random forest" $\hat{h}_{RF}$ se obtiene agregando esta colección de árboles aleatorios. La agregación se realiza de la siguiente manera:

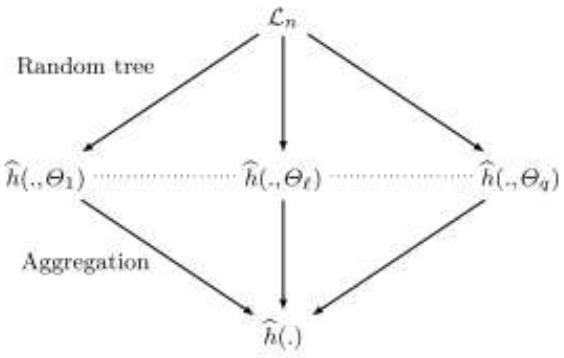


Figura 4. *Esquema general de "random forest" (Genuer & Poggi, 2020)*.

- $\hat{h}_{RF}(x) = \frac{1}{q}\sum_{l=1}^{q} \hat{h}(x, \Theta_l)$ (promedio de predicciones de árboles individuales) en regresión.
- $\hat{h}_{RF}(x) = \underset{1\leq c\leq C}{arg\,max} \sum_{l=1}^{q} \mathbf{1}_{(\hat{h}(x,\Theta_l)=c}$ (voto mayoritario entre las predicciones de árboles individuales) en la clasificación (Genuer & Poggi, 2020).

"Random Forest" es un algoritmo de aprendizaje automático que tiene la capacidad de realizar tareas de regresión y clasificación. Un clasificador de bosque aleatorio hace crecer una serie de árboles de decisión que se entrenan en diferentes partes del mismo conjunto de entrenamiento para mejorar la tasa de clasificación y superar el problema de sobreajuste (Mahboob et al., 2017).

“Random Forest” elige los atributos al azar para crear un número K de árboles cada vez con diferentes atributos sin podar. En el árbol de decisión, los datos de prueba se probarán en el único árbol construido, a diferencia de “Random Forest”, donde los datos de prueba se probarán en todos los árboles construidos y luego se asignará la salida más frecuente a esa instancia (Mishra et al., 2014). Generalmente, si el bosque tiene más árboles, entonces será más robusto. “Random Forest Classifier” tiene la misma idea, la mejor precisión la darán las técnicas de “Random Forest” si tiene una mayor cantidad de árboles en el bosque. Además, los valores perdidos pueden ser manejados por “Random Forest”, este algoritmo nunca se preocupa por la cantidad de árboles en el bosque, nunca se ajusta demasiado y también puede manejar valores categóricos.

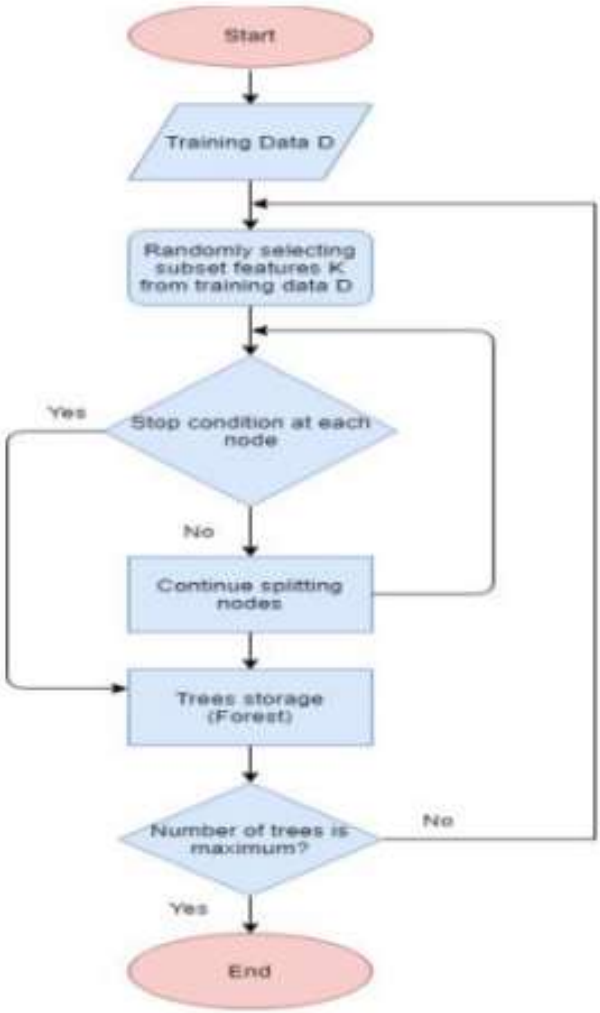


Figura 5. *Algoritmo de Random Forest (Mahboob et al., 2017).*

Se presenta a continuación el proceso del algoritmo:

1) Selecciona individuos al azar (usando muestreo con reemplazo) para crear diferentes conjuntos de datos (Ver figura 6).
2) Al crear los árboles se eligen variables al azar en cada nodo del árbol, dejando crecer el árbol en profundidad (sin podar).
3) Crea un árbol de decisión con cada conjunto de datos, obteniendo diferentes árboles, ya que cada conjunto contiene diferentes individuos y diferentes variables.
4) Predice los nuevos datos usando la "categoría mayoritaria", donde clasificará como "positivo" si la mayoría de los arboles predicen la observación como positiva (Breiman, 2001).

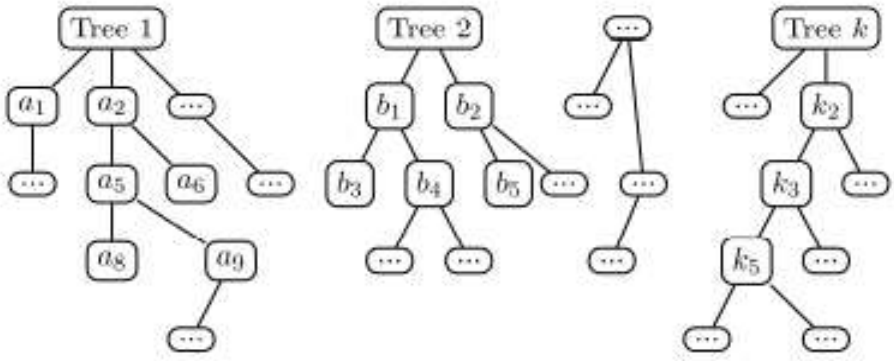


Figura 6. *La idea básica del "random forest" (Yang, 2019).*

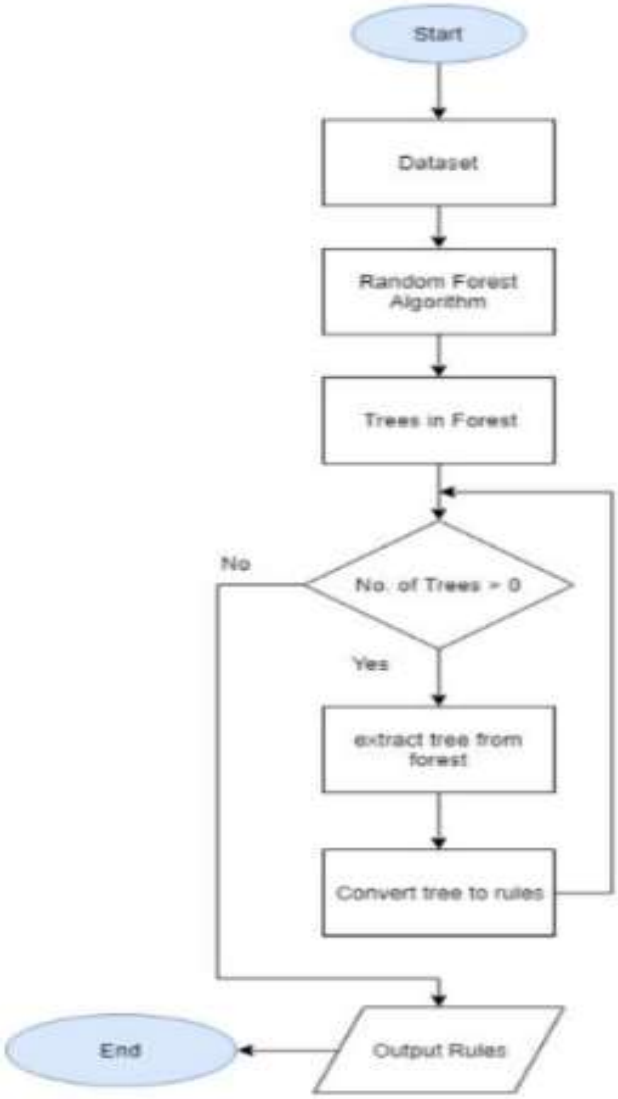


Figura 7. *Proceso de extracción de reglas (Mahboob et al., 2017).*

Asimismo, el algoritmo de “random forest” es uno de los algoritmos de clasificación de datos más usados puesto que una de sus mejores virtudes es que aporta una estimación interna de exactitud mediante una forma de validación cruzada, aportando conocimiento sobre el procedimiento moderno de trabajo en paralelo que es precisamente la distribución en cores del computador para grandes volúmenes de datos.

Estimación del error con “random forest”

Se define la tasa de error fuera de muestra (OOBi) de una observación, como el error obtenido al ser clasificada por los árboles del bosque construidos sin su intervención, es decir, dejando fuera a la muestra no modelada.

La estimación OOB del error es el promedio de todos los OOBi para todas las observaciones del conjunto de datos, es mejor estimador que el error aparente. Parecida a la estimación por validación cruzada, la medida se puede extrapolar al problema de regresión describiéndola en términos del Error Cuadrático Medio (ECM)" (Cárdenas, 2019).

Criterio de división

"Random Forest" utiliza la medida de impureza de Gini para seleccionar la división con la impureza más baja en cada nodo (Breiman, 2003). La impureza de Gini es una medida de la distribución de etiquetas de clase en el nodo. La "impureza de Gini" toma valores en [0, 1], donde se obtiene 0 cuando todos los elementos de un nodo son de la misma clase. Formalmente, la medida de impureza de Gini para la variable $X = \{x_1, x_2, ..., x_j\}$ en el nodo t, donde j es el número de niños en el nodo t, N es el número de muestras, n_{ci} es el número de muestras con valor x_i perteneciente a la clase c, a_i es el número de muestras con valor x_i en el nodo t, entonces la "impureza de Gini" viene dada por: (Breiman, 1984)

$$I\left(t_{x_i}\right) = 1 - \sum_{c=0}^{c} \left(\frac{n_{ci}}{a_i}\right)^2 \qquad (5)$$

El índice de Gini de una división es el promedio ponderado de la medida de Gini sobre los diferentes valores de la variable X, que viene dado por:

$$Gini\,(t, X) = \sum_{i=1}^{j} \frac{a_i}{N} I\left(t_{x_i}\right) \qquad (6)$$

La decisión del criterio de división se basará en el valor de impureza de Gini más bajo calculado entre las m variables. En "Random forest", cada árbol emplea un conjunto diferente de m variables para construir las reglas de división.

Importancia de las variables

Una de las características más importantes del algoritmo "random forest" es la salida de la "importancia variable". La "importancia de la variable" mide el grado de asociación entre una determinada variable y el resultado de la clasificación. "Random Forest" tiene cuatro medidas para la variable importancia: puntuación de importancia bruta para la clase 0, puntuación de importancia bruta para la clase 1,

disminución de la precisión y el índice de Gini. Para estimar la importancia de la variable para alguna variable j, las muestras fuera de la bolsa (OOB) se pasan por el árbol y se registra la precisión de la predicción. Luego, los valores de la variable j se permutan en las muestras OOB y la precisión se mide nuevamente. Estos cálculos se realizan árbol por árbol a medida que se construye el "Random Forest". La disminución promedio en la precisión de estas permutaciones se promedia sobre todos los árboles y se usa para medir la importancia de la variable j. Si la precisión de la predicción disminuye sustancialmente, sugiere que la variable j tiene una fuerte asociación con la respuesta (Hastie, 2009). Después de medir la importancia de todas las variables, "random forest" devolverá una lista clasificada de la "importancia de la variable".

Formalmente, sean β_t las muestras OOB para el árbol t, t ∈ {1,..., ntree}, y'^{t}_{i} es la clase predicha para la instancia i antes de la permutación en el árbol t y $y'^{t}_{i,\alpha}$ es la clase predicha por ejemplo i después de la permutación. La "importancia de la variable" (VI) para la variable j en el árbol t viene dada por:

$$VI_j^t = \frac{\sum_{i=1}^{N} \beta_t I\left(y_i = y_t^{'t}\right)}{|\beta_t|} - \frac{\sum_{i=1}^{N} \beta_t I\left(y_i = y_{i,\alpha}^{'t}\right)}{|\beta_t|} \qquad (7)$$

El "valor de importancia" sin procesar para la variable j se promedia sobre todos los árboles en el "random forest".

$$VI_j = \frac{\sum_{t=1}^{ntree} VI_m^t}{ntree} \qquad (8)$$

La "importancia variable" utilizada es el Gini de disminución media "Mean Decrease Gini" (MDG), que se basa en el criterio de división de Gini. Los MDG miden la disminución ΔI (ecuación 5) que resulta del desdoblamiento. Para un problema de dos clases, el cambio en I (ecuación 10) en el nodo t se define como la impureza de clase (ecuación 9) menos el promedio ponderado de la medida de Gini (ecuación 6) (Bjoern, 2009; Mingers, 1989).

$$I(t) = 1 - \sum_{c=0}^{c} \left(\frac{n_i}{N}\right)^2 \qquad (9)$$

$$\Delta I(t) = I(t) - Gini(t, X) \quad (10)$$

La disminución en la “impureza de Gini” se registra para todos los nodos t en todos los árboles (ntree) en “random forest” para todas las variables y luego se calcula la Importancia de Gini (GI) como: (Bjoern, 2009).

$$GI = \sum_{ntree} \sum_{t} \Delta I(t) \quad (11)$$

(Khalilia et al., 2011)

2.1.2.2. Desbalanceo de datos

El objetivo de un algoritmo de clasificación de datos es intentar aprender un separador o clasificador, que pueda distinguir las dos categorías de la variable respuesta. Por lo general, se muestran ejemplos como la figura 8 en dos dimensiones, con puntos que representan datos y diferentes colores de los puntos que representan la categoría. Hay muchas maneras de hacerlo, basadas en varias suposiciones matemáticas, estadísticas o geométricas (Fawcett, 2016) como se muestra en la figura 8.

a) b) c)

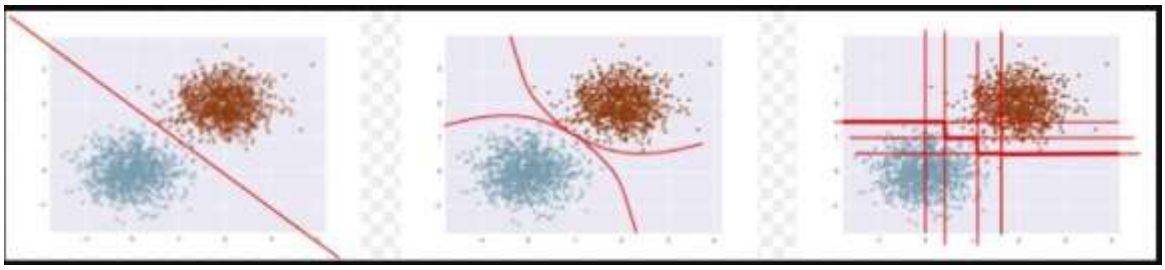


Figura 8. *Clasificadores posibles para separar 2 categorías de la variable respuesta (Fawcett, 2016)*
Nota. a) Suposición matemática; b) Suposición estadística; c) Suposición geométrica.

Pero cuando se comienza a trabajar con datos reales, una de las primeras observaciones que resalta es la desigualdad de proporción de las dos categorías de la variable respuesta, como se muestra en la figura 9.

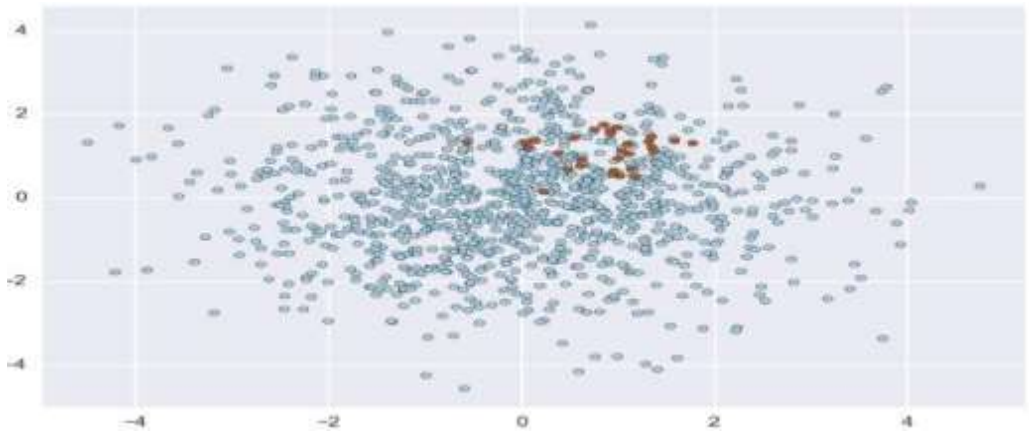


Figura 9. *Categorías desproporcionadas de la variable respuesta (Fawcett, 2016)*

El problema principal es que estas categorías están desbalanceadas: los puntos rojos son superados en gran medida por el azul.

Algunos ejemplos claros donde radica principalmente este problema es:

- ✓ Alrededor del 2% de las cuentas de tarjetas de crédito son defraudadas por año. (La mayoría de los dominios de detección de fraude están muy desbalanceados).
- ✓ El examen médico para una afección generalmente se realiza en una gran población de personas sin esta afección, para detectar una pequeña minoría que lo acompaña (por ejemplo, la prevalencia del virus de la inmunodeficiencia humana (VIH) en EE. UU. es aproximadamente de 0,4%).
- ✓ Las fallas de la unidad de disco son aproximadamente 1% por año.
- ✓ Las tasas de defectos de producción en fábrica normalmente se ejecutan alrededor de 0,1%.

Los algoritmos convencionales o tradicionales a menudo están sesgados hacia la categoría mayoritaria porque sus funciones de pérdida intentan optimizar cantidades tales como la tasa de error, sin tener en cuenta la distribución de datos.

En el peor de los casos, los ejemplos de minorías se tratan como valores atípicos de la categoría mayoritaria e ignorada. El algoritmo de aprendizaje simplemente genera un clasificador trivial que clasifica cada ejemplo como la categoría mayoritaria.

La solución según Fawcett (2016), para este problema es:

1) Equilibre el conjunto de entrenamiento de alguna manera:
   - Sobremuestra de la categoría minoritaria.
   - Dé muestras de la categoría mayoritaria.
   - Sintetiza nuevas categorías de minorías.
2) Replicar los ejemplos de minorías y cambie a un marco de detección de anomalías.
3) En el nivel del algoritmo, o después de él:
   - Ajuste el peso de la categoría (costos de clasificación errónea).
   - Ajusta el umbral de decisión.
   - Modifique un algoritmo existente para que sea más sensible a las categorías raras.
4) Construya un algoritmo completamente nuevo para obtener buenos resultados en datos desequilibrados.

Las técnicas más comunes que se usa en estos tipos de problemas son: "*oversampling"*, "*undersampling"* y *combinación*.

Los enfoques más fáciles requieren pocos cambios en los pasos de procesamiento, y simplemente implican ajustar los conjuntos de ejemplos hasta que estén equilibrados.

2.1.2.2.1. Sobremuestreo ("Oversampling")

Esta técnica replica aleatoriamente instancias minoritarias (categoría menor en proporción de la variable respuesta) para aumentar su población y así equilibrar a la categoría mayoritaria. En la figura 10 se entiende el procedimiento.

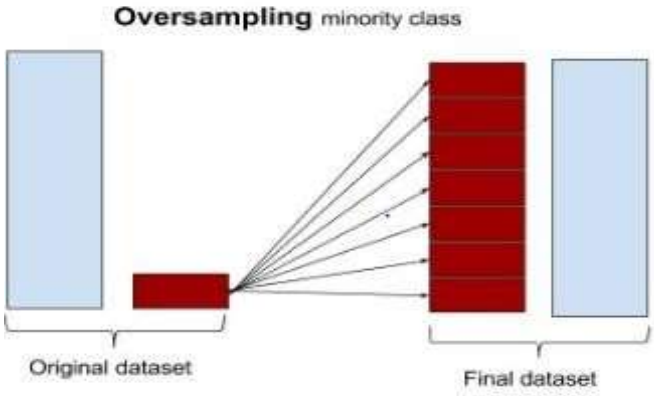


Figura 10. *Funcionamiento del "oversampling" (Fawcett, 2016)*
Nota. La categoría minoritaria en la base de datos original es la barra de color rojo, se hace aumentar en la base de datos final para que se equilibre con la categoría mayoritaria de color celeste.

2.1.2.2.2. Submuestreo ("Undersampling")

A comparación de la técnica anterior, este proceso realiza lo contrario, es decir, aleatoriamente reduce la categoría de la mayoría hasta completar la categoría minoritaria y así equilibrar las muestras para el entrenamiento del modelo de "machine learning*"* tal como se muestra en la figura 11.

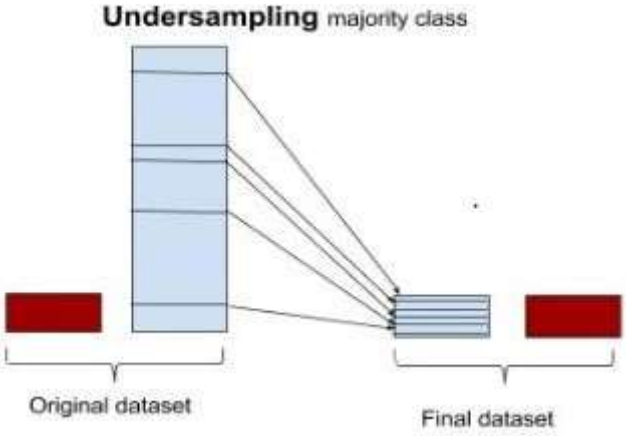


Figura 11. *Funcionamiento del "undersampling" (Fawcett, 2016)*
Nota. La categoría mayoritaria en la base de datos original es la barra de color celeste, se hace reducir en la base de datos final para que se equilibre con la categoría minoritaria de color rojo.

2.1.2.2.3. Combinación ("Bothsampling")

Cuando usamos esta técnica, la categoría minoritaria es sobremuestreada, mientras que la mayoritaria es submuestreada sin reemplazamiento. Esta técnica aplica en

simultáneo un algoritmo de “undersampling” y otro de “oversampling” a la vez al conjunto de datos.

Las técnicas más usadas para el empleo de la técnica **combinación** son: **SMOTE** para “oversampling”: busca puntos vecinos cercanos y agrega puntos “en línea recta” entre ellos. Y **Tomek** para “undersampling” que quita los de distinta categoría que sean “nearest neighbor” y deja ver mejor la decisión “boundary” (la zona limítrofe de nuestras categorías).

2.1.2.3. SMOTE

“SMOTE (Técnica de Sobremuestreo de Minorías Sintéticas) consiste en la síntesis de elementos para la categoría minoritaria, basados en los que ya existen. Funciona eligiendo al azar un punto de la categoría minoritaria y calcula los k vecinos más cercanos para este punto. Los puntos sintéticos se agregan entre el punto elegido y sus vecinos” (Cárdenas, 2019). La idea es crear nuevos ejemplos minoritarios interpolando entre los existentes. En la figura 12 se ilustra el procedimiento con la técnica SMOTE.

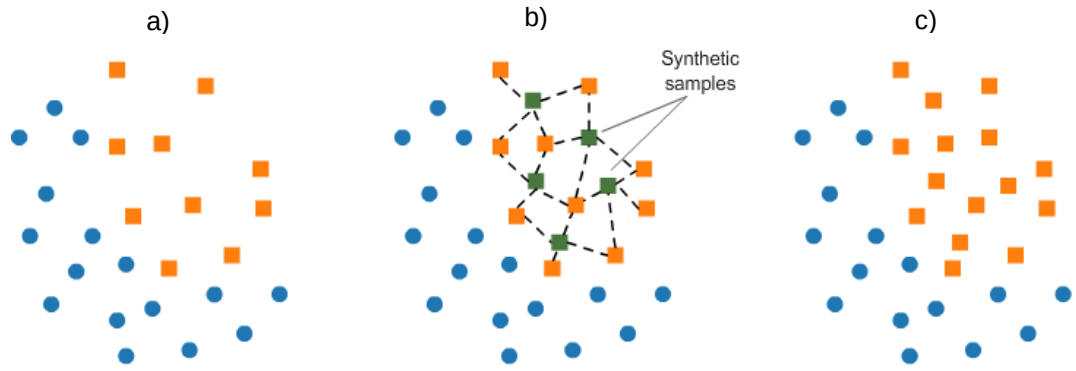


Figura 12. *Funcionamiento del SMOTE (Fawcett, 2016)*
Nota. (a) La categoría minoritaria son los puntos de color anaranjado; (b) Se crea los puntos sintéticos verdes interpolando con los k vecinos más cercanos de color anaranjado; (c) los puntos verdes pasan a formar parte de los puntos anaranjado (categoría minoritaria).

2.1.2.4. “Tomek Links”

“Tomek Links” (Técnica de “undersampling”), se eliminan las instancias de la categoría mayoritaria que sean redundantes o que se encuentren muy cerca de instancias de categoría minoritaria (Tomek, 1976).

“Tomek links” elimina ejemplos ruidosos y cercarnos a la frontera de la clase minoritaria. Los ejemplos cercarnos a la frontera se consideran inseguros, pues al poseer alguna cantidad de ruido puede conllevar a que el caso caiga al otro lado de la frontera de decisión (Kubat, 1997).

El método SMOTE + Tomek links: inicialmente se realiza el “oversampling” con la categoría minoritaria y luego se aplica el “Tomek Link” a ambas categorías (Batista, 2002).

2.2. Marco conceptual

2.2.1. Encuesta Demográfica y de Salud Familiar (ENDES)

ENDES es una de las investigaciones estadísticas más importantes que ejecuta de manera continua el Instituto Nacional de Estadística e Informática y que se realiza en el marco del programa mundial de las Encuestas de Demografía y Salud, conocido en la actualidad como MEASURE. La ENDES continúa los esfuerzos iniciados con la Encuesta Mundial de la Fecundidad y la Encuesta de Prevalencia de Anticonceptivos, en 1977-78 y 1981, respectivamente, para obtener información actualizada y efectuar análisis del cambio, tendencias y determinantes de la fecundidad, mortalidad y la salud en los países en vías de desarrollo.

> *“Desde el año 2004, la ENDES se ejecuta anualmente en el marco del Presupuesto por Resultado, luego de firmar un convenio con el Ministerio de Economía para contar con información que permita estimar de manera oportuna y confiable los indicadores identificados en los Programas Estratégicos a nivel departamental. Entre los principales indicadores del Programa Presupuestal Articulado Nutricional que se investigan en la ENDES son: Desnutrición Crónica infantil, lactancia exclusiva, anemia, enfermedad respiratoria aguda, bajo peso al nacer, mortalidad infantil y en la niñez, vacunas básicas completas, entre otros. Respecto al Programa Presupuestal a la salud materno neonatal, se recoge información sobre tasa de mortalidad neonatal, tasa global de fecundidad, parto institucional, parto por cesárea, controles prenatales, preferencias reproductivas, nutrición en las mujeres en edad fértil, prevalencia de infecciones de transmisión sexual en mujeres en edad fértil, entre otros”.* (ENDES, 2016)

El método de recolección de datos es por entrevista directa, con personal debidamente capacitado y entrenado para tal fin y que visita las viviendas seleccionadas durante el período de recolección de información.

La información se recoge durante todo el año (todos los meses) siendo su población de estudio: hogares particulares y sus miembros residentes habituales, mujeres de 12 a 49 años, niñas y niños menores de 6 años, niñas y niños menores de 12 años y personas de 15 y más años. Asimismo, el tipo de muestra es bietápica, probabilística de tipo equilibrado, estratificada e independiente a nivel regional, por área urbana y rural.

> *"En el Perú, desde 1975 hasta 1991, con la Encuesta Demográfica Nacional (EDEN-PERU) y el levantamiento de la ENDES 1986, la ENDES 1991-92 y la ENDES 2000, se ha tenido la oportunidad de conocer aproximadamente cada cinco años, el nivel, tendencia y diferenciales de la fecundidad, mortalidad, prevalencia anticonceptiva y de la salud familiar; conocimiento que ha sido y es fundamental para el diseño y orientación de las políticas y programas de población".* (Instituto Nacional de Estadística e Informática [INEI], 2020)

La medición de la hemoglobina en la ENDES se rige según las normas y procedimientos plasmados en las normativas técnicas del sector (Instituto Nacional de Salud [INS], 2013). Para el análisis de hemoglobina de las(os) niñas(os), mujeres de 12 a 14 años y de las mujeres en edad fértil (MEF) en la ENDES se utiliza el hemoglobinómetro HemoCue modelo Hb 201+. Cabe mencionar, que antes del análisis de hemoglobina se mencionan algunas recomendaciones y pasos previos que el personal de la ENDES debe cumplir durante la medición:

Antes de salir a campo verifique todo el material necesario para el trabajo del día, incluyendo material para análisis adicionales sin alterar el cierre de ésta ni la estabilidad del equipo. Es responsabilidad de la antropometrista contar con todo el material para la evaluación.

Lea el consentimiento informado

Si la madre no acepta el Consentimiento Informado, entonces no realice el análisis de hemoglobina al niño o niña si:

- La madre se opone.

- El niño o niña tiene fiebre muy alta o está con diarrea

- El niño o niña tiene alguna limitación que imposibilite la evaluación

La secuencia para el análisis de hemoglobina se recomienda en el siguiente orden. Sin embargo, no es determinante y se manejará de acuerdo a la situación que se presente en la vivienda:

- Primero: los niños y niñas de 12 a 71 meses; la madre puede observar todo el procedimiento y perder el temor cuando se le toma la muestra de sangre.

- Segundo: los niños y niñas de 4 a 11 meses, por lo general no perciben la toma de la muestra de sangre y no lloran.

- Tercero: paciente adulto o mujeres en edad fértil.

El hemoglobinómetro y los materiales deben estar colocados sobre una superficie plana y evitar la exposición directa a la luz solar o el viento.

La madre debe estar cómodamente sentada y cargar correctamente al niño o niña, ello permitirá al analista manipular adecuadamente los materiales y evitar dificultades en el procedimiento del análisis.

2.2.2. Determinación de la hemoglobina

La determinación de la hemoglobina en la ENDES se hace mediante el método colorimétrico, con un equipo portátil HemoCue ® (HemoCue AB, Angelhome, Suecia). El método se basa en una reacción modificada de la azida-metahemoglobina, a partir del método de Vanzetti (Vanzetti, 1966). El equipo utiliza microcubetas que contienen el reactivo, constituido por desoxicolato de sodio, nitrito de sodio y azida de sodio.

> *La muestra de sangre capilar se obtiene del dedo anular o medio de la mano (en niños menores de seis meses la muestra es del talón), la cual se vierte por capilaridad en la microcubeta. La cubeta con la muestra se lee en el HemoCue ® a una longitud de onda doble de 565 a 880 nm.* (ENDES, 2016)

El valor obtenido por el Hemocue ® en sangre capilar, se soluciona de acuerdo con la altitud donde viven los niños evaluados por medio de la fórmula de Dirren et al. (1994). La ENDES no incluye la recolección de datos particulares, exámenes

clínicos u otras pruebas complementarias que permitan diferenciar las probables causas clínicas de la anemia.

La determinación de la hemoglobina en sangre capilar mediante el HemoCue ® es un excelente método para el despistaje de la anemia, con una precisión bastante cercana a la obtenida por métodos directos con sangre venosa y arterial (Sanchis et al., 2013).

2.2.3. Librería "ROSE"

Esta librería del software estadístico R proporciona funciones para tratar problemas de clasificación binaria en presencia de categorías desbalanceadas. Las muestras sintéticas balanceadas se generan según ROSE (Menardi y Torelli, 2013). También se proporcionan funciones que implementan remedios más tradicionales para el desbalanceo de categorías, así como diferentes métricas para evaluar una precisión de aprendizaje. Estos se estiman mediante métodos de retención, arranque o validación cruzada.

2.2.4. Paquete "randomForest"

El paquete randomForest en R fue introducido por Liaw & Wiener (2002) y se basa en el código Fortran inicial de Breiman (2001) y Cutler (2010). El código fuente del paquete está estructurado de una manera bastante compleja, con funciones principales en R que llaman código C, que a su vez se llama código Fortran. Sin embargo, para el usuario, el paquete cuenta con la función principal *randomForest()* bien documentada, con métodos habituales de R como *print()*, *plot()*, *predict()* (no implementado por el método *summary()*) y otras funciones relacionadas que ayudan a usar o interpretar bosques aleatorios como *importance()* o *partialPlot()*.

Los principales parámetros de la función *randomForest()* son:

- *mtry*: el número de variables seleccionadas aleatoriamente en cada nodo, que por defecto es $\sqrt{p}$ en clasificación y p/3 en regresión (donde p se refiere al número de variables de entrada).
- *ntree*: el número de árboles en el bosque, indicado como q, que por defecto es 500.

- *nodesize*: el número mínimo de observaciones que debe contener una hoja de un árbol, que por defecto es 1 en clasificación y 5 en regresión (Genuer & Poggi, 2020).

En el enfoque “Random Forest” (bosque aleatorio), se crean un gran número de árboles de decisión. Cada observación se introduce en cada árbol de decisiones. El resultado más común para cada observación se utiliza como resultado final. Una nueva observación se introduce en todos los árboles y toma el dato mayoritario para cada modelo de clasificación.

Se realiza una estimación de error para los casos que no se utilizaron durante la construcción del árbol. Esto se denomina estimación de error **OOB** (fuera de la bolsa) que se menciona como porcentaje.

Error OOB. Para predecir la i-ésima observación X_i, solo agregamos predictores basados en muestras bootstrap que no contienen (X_i, Y_i). Esto proporciona una predicción Y_i para la salida de la i-ésima observación.

El error OOB se calcula entonces de la siguiente manera:

- $En\ regresión\ \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$
- $En\ clasificación\ \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{Y_i \neq \hat{Y}_i}$

(Genuer & Poggi, 2020).

2.2.5. Librería “caret”

La librería “caret” (*classification and regression training*, Kuhn (2016)) del software estadístico R incluye una serie de funciones que facilitan el uso de decenas de métodos complejos de clasificación y regresión. Utilizar esta librería en lugar de las funciones originales de los métodos presenta dos ventajas:

1) Permite utilizar un código unificado para aplicar reglas de clasificación muy distintas, implementadas en diferentes paquetes.
2) Es más fácil poner en práctica algunos procedimientos usuales en problemas de clasificación. Por ejemplo, hay funciones específicas para dividir la muestra en datos de entrenamiento (train) y datos de prueba (test) o para ajustar parámetros mediante validación cruzada.

2.2.6. Algoritmo “Hmisc”

“Hmisc contiene un conjunto de herramientas para la reducción de datos, la imputación, el cálculo de potencia y tamaño de muestra, la creación avanzada de tablas, variables de recodificación, importación e inspección de datos y gráficos generales.

El algoritmo de imputación múltiple de “*Hmisc”* toma en cuenta la incertidumbre de las imputaciones mediante “bootstraping” para aproximarse a la predicción de los valores a partir de la distribución predictiva bayesiana completa. Está basado en modelos semiparamétricos que utilizan los métodos *regresión aditiva, “bootstraping” y pareamiento por medias predictivas*.

El algoritmo funciona de la siguiente forma:

1. Para cada variable con valores perdidos (NA), inicializa esta con valores de una muestra aleatoria de los valores observados.
2. Realiza lo siguiente “burnin”+”n.impute” veces:
    a. Extrae una muestra con reemplazo del conjunto de datos completo y ajusta un modelo aditivo flexible para predecir los valores de todos los casos.
    b. Imputa cada valor faltante con el observado de la observación, cuyo valor predicho es más cercano al predicho del valor faltante (*pareamiento por medias predictivas*)” (Durán, 2019).

2.2.7. Librería “Hmisc”

“Hmisc” es una librería multiusos del software estadístico R útil para el análisis de datos, gráficos de alto nivel, imputación de valores que faltan, fabricación avanzada de tablas, ajuste y diagnóstico de modelos (regresión lineal, regresión logística y regresión cox). En medio de la amplia gama de funciones contenidas en esta librería, ofrece 2 potentes funciones para imputar valores que faltan. Estos son *impute()* y *aregImpute()*.

La *función aregImpute()* permite la imputación de medias utilizando regresión aditiva, “bootstrapping” y emparejamiento de medias predictivo.

En el “bootstrapping”, se utilizan diferentes remuestreos de arranque para cada una de las imputaciones múltiples. A continuación, se instala un modelo aditivo flexible

(método de regresión no paramétrico) en muestras tomadas con reemplazos de datos originales y los valores que faltan (actúa como variable respuesta) se predicen utilizando valores que no faltan (variable independiente).

A continuación, utiliza la coincidencia media predictiva (valor predeterminado) para imputar los valores que faltan. La coincidencia media predictiva funciona bien para datos continuos y categóricos (binarios y multinivel) sin necesidad de residuos informáticos y la máxima probabilidad de ajuste.

Estos son algunos aspectos destacados importantes de esta librería:

- Asume la linealidad en las variables que se predicen.
- El método de puntuación óptimo de Fisher se utiliza para predecir variables cualitativas.

2.2.8. Modelo de clasificación

Los modelos de clasificación utilizan como respuestas variables cualitativas y las variables independientes puede ser cualitativas o cuantitativas.

Un modelo de clasificación predice una categoría. Es para predecir cualquiera de las dos clases objetivos. Ejemplo, predecir si un alumno aprobará o no, predecir si un niño tendrá o no tendrá anemia.

Tabla de clasificación. La tabla de clasificación muestra la distribución de valores observados y pronosticados. Los valores observados son los valores reales y los valores pronosticados se obtienen a partir del modelo de clasificación. En la Tabla 1 se muestra la estructura de una tabla de clasificación.

Tabla 1

Tabla de clasificación (Fernández, 2016)

Valores observados		**Valores pronosticados**		
		A		Porcentaje correcto
		1	0	
A	1	a	b	a/(a+b) *
	0	c	d	d/(c+d) **
Porcentaje Global				(a+d)/(a+b+c+d)

***Sensibilidad:** **“**indica la capacidad que tiene un modelo para clasificar correctamente la categoría de interés de la variable respuesta” (Fernández, 2016).

****Especificidad:** **“**indica la capacidad que tiene un modelo para clasificar

correctamente la categoría que no es de interés de la variable respuesta" (Fernández, 2016).

Curva ROC: la curva ROC ("Receiver Operating Characteristic", característica operativa del receptor) es una herramienta común utilizada con los clasificadores binarios. La curva ROC traza la sensibilidad frente a 1-especificidad.

La curva ROC "indica que cuanto más alejada este de la diagonal principal mejor es el método de diagnóstico, ya que la curva ROC ideal sería la que con una especificidad de 1 tuviera una sensibilidad de 1, y cuanto más cercana esté a dicha diagonal peor será el método de diagnóstico" (Géron, 2019).

En la figura 13 se observa una línea de puntos que representa la curva ROC de un clasificador puramente aleatorio; un buen clasificador se mantiene lo más lejos posible de esa línea (hacia la esquina superior izquierda).

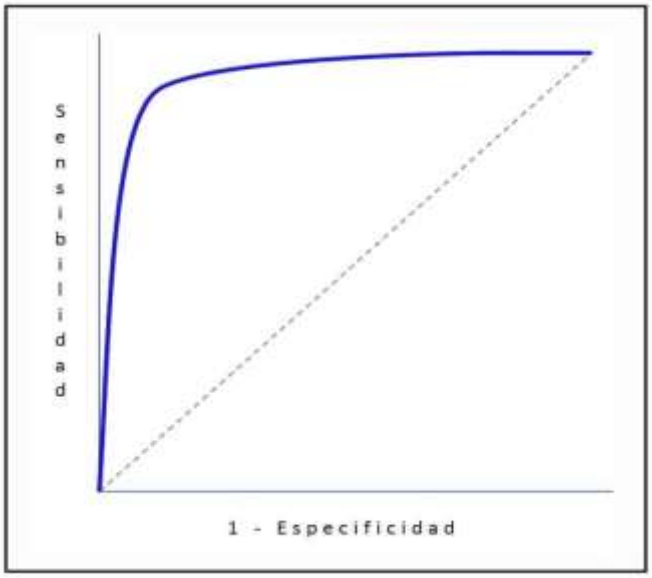


Figura 13. *Representación de la curva ROC a partir de los indicadores especificidad y sensibilidad.*

AUC: ("area under the curve", área debajo de la curva) mide el área debajo de la curva ROC para comparar clasificadores. Un clasificador perfecto tendrá un área debajo de la curva ROC igual a 1, mientras que un clasificador puramente aleatorio la tendrá igual a 0.5. Pérez (2015): "menciona que la diagonal principal es la que corresponde al peor test de diagnóstico y tiene un área bajo la curva de 0.5. Adicionalmente se ha establecido los siguientes intervalos para los valores AUC de

la curva ROC:

- [0.5 - 0.6>: Test malo
- [0.6 - 0.75>: Test regular
- [0.75 - 0.9>: Test bueno
- [0.9 - 0.97>: Test muy bueno
- [0.97 - 1>: Test excelente

Se plantean las siguientes hipótesis:

H0: el área bajo la curva ROC es igual a 0.5

H1: el área bajo la curva ROC no es igual a 0.5

Si se rechaza Ho asociado a un p-valor implica que el modelo ajustado es el adecuado" (Pérez, 2015).

Parámetros: Son valores que se configuran antes del entrenamiento del modelo y no forman parte del modelo como tal, por ejemplo, learning rate, número de capas, número de árboles, etc. Es decir, son configuraciones que no se obtienen a partir de la data.

"Grid search": es un proceso que busca exhaustivamente a través de un subconjunto especificado manualmente del espacio de parámetros del algoritmo de destino. La búsqueda aleatoria, por otro lado, selecciona un valor para cada parámetro de forma independiente utilizando una distribución de probabilidad. Permite analizar varias combinaciones de parámetros.

CAPÍTULO III

MARCO METODOLÓGÍCO

3.1. Hipótesis central de la investigación

La aplicación del algoritmo “random forest” para un modelo de clasificación permite predecir la anemia de niños de 6 a 35 meses del Perú en los años 2015 al 2019 en un porcentaje mayor al 50% mediante un modelo basado en datos balanceados con reajuste de parámetros.

3.2. Variables e indicadores de la investigación

La variable respuesta es la tenencia de anemia en niños de 6 a 35 meses de edad, la que se confirma cuando la hemoglobina sanguínea corregida registra un valor menor de 11 mg/dl (WHO, 2014).

Teniendo como referencia los modelos de Sanou & Ngnie-Teta (2012) junto con el de Balarajan et al. (2011) las variables independientes se organizaron en tres grupos: a) variables sociodemográficas (área de residencia, altitud de la ciudad donde vive el niño, región natural, índice de bienestar o riqueza al que pertenece el hogar, edad materna, grado de instrucción de la madre, lengua materna de la madre, conexión domiciliaria de agua potable, conexión domiciliaria de desagüe, material predominante del piso de vivienda); b) variables relacionadas con el niño (sexo, edad, número de niños menores de cinco años en el hogar, número de personas que viven en el hogar, bajo peso al nacer (< 2500 gr), orden de nacimiento, intervalo entre nacimientos, signos y síntomas (fiebre) en las dos semanas previas, diarrea durante las dos últimas semanas, tos y respiración rápida durante las dos últimas semanas), y c) variables del cuidado materno e infantil (control prenatal, control prenatal en primer trimestre, parto institucional, talla de la madre, diagnóstico de anemia en la madre en el momento de la encuesta, tiempo de consumo de suplemento de hierro en la gestación, suplemento de vitamina A, niño recibió hierro en pastillas o jarabes, medicación antiparasitaria en el niño (tratamiento antiparasitario los últimos seis meses), el niño comió algún tipo de carne el día de ayer (res, pollo, hígado, cerdo, etc.), consumo de agua hervida y control de crecimiento y desarrollo (CRED)).

En la Tabla 2 se presenta la relación de preguntas de la ENDES tomadas en cuenta para la formación de las diferentes variables para la presente investigación, junto con el módulo al que pertenecen y el nombre de cada variable original de donde proviene cada pregunta.

Tabla 2

Preguntas de la ENDES utilizadas en la investigación según tipo de modulo y nombre de la variable

Modulo	**Nombre de la variable**	**Preguntas**
Niños Talla/Peso/Hemoglobina (Antropometría / anemia – niños) – Cuestionario del hogar	Nivel de anemia	Número de personas con hemoglobina por debajo del límite: >11g/dl en niños (6-59 meses) y MEF gestante.
	El nivel educativo más alto de la madre	"108. ¿Cuál fue el año o grado de estudios más alto que aprobó? 0. Inicial/preescolar 1. Primaria 2. Secundaria 3. Superior no universitario 4. Superior universitario 5. Postgrado"
	Sexo	Sexo
	Edad en meses	Edad del niño (en meses)
	Número de orden de nacimiento	Número de orden de nacimiento del niño
	Intervalo de los nacimientos anteriores al niño	Intervalo de los nacimientos anteriores al niño

Tabla 2. *(Continuación a …)*

Modulo	Nombre de la variable	Preguntas
Característica de la vivienda – Cuestionario Individual	Tipo de lugar de residencia	"102. Primero me gustaría hacerle algunas preguntas acerca de Ud. y de su hogar. Antes de que Ud. cumpliera los 12 años ¿Dónde vivió la mayor parte del tiempo: En una ciudad, en un pueblo o en el campo? 1. Capital del departamento 2. Ciudad 3. Pueblo 4. Campo 5. Extranjero"
	Altitud del conglomerado	Altitud del conglomerado (en metros)
	Número de miembros del hogar	Total de personas en el hogar:
	Índice de riqueza	Se calcula utilizando el método de componentes principales
	Fuente principal de abastecimiento de agua potable que utilizan en su hogar para tomar o beber	"40. ¿Cuál es la fuente principal de abastecimiento de agua que utilizan en su hogar para tomar o beber? RED PÚBLICA 11 Dentro de la vivienda 12 Fuera de la vivienda, pero dentro del edificio 13 Pilón/Grifo Público AGUA DE POZO 21 Pozo en la casa/patio/lote 22 Pozo Público AGUA DE SUPERFICIE 31 Manantial (Puquio) 32 Río/acequia/laguna 41 AGUA DE LLUVIA 51 CAMIÓN TANQUE/AGUATERO 96 OTRO (Especifique)"
	Tipo de instalación sanitaria	"53. ¿Qué tipo de servicio higiénico tiene su hogar? CONECTADO A RED PÚBLICA 11 Dentro de la vivienda 12 Fuera de la vivienda, pero dentro del edificio 21 Pozo séptico/tanque séptico LETRINA 31 Mejorada ventilada 32 Mejorada ecológica/abonera/compostera 33 Mejorada colgante /flotante 34 Pozo ciego o negro con tratamiento de cal, ceniza, estiércol, aserrín, arena 35 Pozo ciego o negro 41 RÍO, ACEQUIA O CANAL 51 NO HAY SERVICIO (MATORRAL / CAMPO) 96 OTRO (Especifique) *Si tiene LETRINA, sondee el tipo".
	Etnicidad	"119A. ¿Cuál es el idioma o lengua materna que aprendió en su niñez? 1. Castellano 2. Quechua 3. Aymará 4. Otra lengua aborigen". 5. Idioma extranjero

Tabla 2. *(Continuación b …)*

Modulo	Nombre de la variable	Preguntas
Datos básicos del hogar - Cuestionario del hogar	Número de niños menores de 5 años	"Total de personas en el hogar: Número de niñas y niños menores de 5 años".
Específicas del País - variables simples – Cuestionario de la mujer	Región natural	"Región Natural: 1. Lima metropolitana 2. Resto Costa 3. Sierra 4. Selva alta 5. Selva baja"
Características de Hogar y vivienda - Hogar	Material predominante del piso de la vivienda	"73. Material predominante del piso de la vivienda (Por observación o indague) PISO ACABADO 11. Parquet o madera pulida 12. Láminas asfálticas, vinílicos o similares 13. Losetas, terrazos o similares 14. Cemento/ladrillo PISO RÚSTICO 21. Madera (entablados) 22. Pona PISO NATURAL 31. Tierra / arena 96. Otro (especifique)"
	El agua usualmente es tratada por: hervida	47. En su hogar, ¿Qué le hacen al agua que habitualmente utilizan para tomar o beber? 02. La hierben
Mujeres Talla/ peso/ hemoglobina – Cuestionario del Hogar	Edad de la mujer en años	"MUJERES DE 12 A 49 AÑOS Edad"
	Talla en centímetros (1 decimal)	"MEDICIÓN DE PESO Y TALLA DE MUJERES DE 12 A 49 AÑOS Columna 205. Talla (centímetros)"
	Nivel de anemia	Mujeres en Edad Fértil de 15 a 49 años de edad: grave, moderado, leve y sin anemia
Variables mujeres (cont.) – Cuestionario Individual – Mujeres de 12 a 49 años	Persona que normalmente alimenta al niño	"496. ¿Generalmente quién le da de comer a (NOMBRE)? 01. Entrevistada 02. Esposo/ compañero 03. Hijas/ Hijos mayores 04. Padres/ Suegros 05. Otros parientes 06. Vecinos/ amigos 07. Otros no parientes 08. Empleada Doméstica 09. Nadie/ Come Solo 96. Otro (Especifique)"

Tabla 2. *(Continuación c …)*

Modulo	Nombre de la variable	Preguntas
Salud niños (cont.) – Cuestionario Individual	En los últimos 7 días tomo hierro en jarabe	465E. En los últimos siete días ¿(NOMBRE) tomó: a. ¿Hierro en pastillas o jarabe? 1. Si 2. No 3. No sabes
	En los Últimos 7 días tomo hierro en polvo como Micronutrientes - (chispitas, estrellitas o NUTRIMIX)	"465E. En los últimos siete días ¿(NOMBRE) tomó: b. ¿Hierro en polvo como chispitas o estrellitas? 1. Si 2. No 3. No sabes"
	En los Últimos 7 días tomo hierro en gotas	"465E. En los últimos siete días ¿(NOMBRE) tomó: c. ¿Hierro en gotas? 1. Si 2. No 3. No sabes"
	En los Últimos 7 días tomo hierro en otra presentación	465E. En los últimos siete días ¿(NOMBRE) tomó: d. ¿Hierro en otra presentación? 1. Si 2. No 3. No sabes
	Le hicieron algún control de crecimiento y desarrollo	"466. En los últimos 6 meses ¿le hicieron a (NOMBRE) algún control de Crecimiento y Desarrollo? 1. Si 2. No 8. No sabe"

Tabla 2. *(Continuación d …)*

Modulo	Nombre de la variable	Preguntas
Salud (niños) – Cuestionario Individual	Ha tenido fiebre en las últimas dos semanas	467. En los últimos 14 días, es decir, entre el ___ y el día de ayer, ¿(NOMBRE) ha tenido fiebre? 1. Si 2. No 8. No sabe
	En los últimos 14 días, ha tenido diarrea la niña(o)	472. En los últimos 14 días, es decir, entre el ____ y el día de ayer, ¿(NOMBRE) ha tenido diarrea? 1. Si 2. No 8. No sabe
	Ha tenido tos en las últimas dos semanas	468. En los últimos 14 días, es decir, entre el ___ y el día de ayer, ¿(NOMBRE) ha tenido tos? 1. Si 2. No 8. No sabe
	Alguna vez recibió la dosis de vitamina A	465B. ¿Recibió (NOMBRE) alguna dosis de Vitamina A? 1. Si 2. No 8. No sabe
	Medicamentos para parásitos intestinales en los últimos 6 meses	477. En los últimos 12 meses, entre_____ del año pasado y _____ de este año, ¿(NOMBRE) ha recibido algún tratamiento para las lombrices o los gusanos intestinales? 1. Sí 2. No

Tabla 2. *(Continuación e …)*

Modulo	Nombre de la variable	Preguntas
Maternidad – Cuestionario Individual	Visitas prenatales por embarazo	"410. ¿Cuántos controles prenatales tuvo Ud. durante el embarazo de (NOMBRE)? Nº de controles /__/__/ No sabe ..... 98"
	Momento del primer control prenatal	"409. ¿Cuántos meses de embarazo tenía Ud. cuando se hizo su primer control prenatal? Meses /__/__/ No sabe... 98"
	Peso del niño al nacer (kilos - 3 dec.)	"430B.¿Cuánto pesó (NOMBRE)? solicite que le muestren el carné de crecimiento y desarrollo y transcriba la información. 1. Gramos del carné 2. Gramos según recuerda 99998. No sabe"
	La atendió en el parto: Médico	426. ¿Quién la atendió en el parto de (NOMBRE)? A. Médico Luego pregunte: ¿Alguien más?
	La atendió en el parto: Enfermera	"426. ¿Quién la atendió en el parto de (NOMBRE)? C. Enfermera Luego pregunte: ¿Alguien más?"
	La atendió en el parto: Obstetra	"426. ¿Quién la atendió en el parto de (NOMBRE)? B. Obstetriz Luego pregunte: ¿Alguien más?"
	Por cuantos días tomó hierro y/o cuantas inyecciones recibió	"422. Durante todo el embarazo de (NOMBRE), ¿por cuántos días tomó hierro y/o cuántas inyecciones recibió? Número de días /__/__/ No sabe 998 Número de inyecciones /__/__/ No sabe 98"
	El día de ayer o durante el día o la noche cuantas veces le dio comida sólidas o semisólidas	"448. Ayer durante el día o la noche, ¿le dió a (NOMBRE) comidas sólidas o semisólidas distintas a líquidos? Número de veces /_ /_ / 00. No come 98. No sabe"

En la Tabla 3 se presenta la operacionalización de las 34 variables construidas a partir de las 39 preguntas tomadas de los cuestionarios de la ENDES, asimismo se presenta su definición conceptual, su definición operacional (categorías), indicadores y la variable original en la ENDES, junto con el módulo al que pertenece cada variable original.

Tabla 3

Operacionalización de las variables

Variables	Definición conceptual	Definición operacional (Categoría)	Indicadores	Variable original / Modulo
Variable respuesta				
Presencia de anemia en niños de 6 a menos de 35 meses (ANEMIA)	Hemoglobina corregida por altura < 11.0 g/dL	1: Sin anemia 2: Con anemia	Prevalencia de niños con anemia	HC57 / RECH6
Variables independientes				
Nivel educativo más alto de la madre (X01)	Es el grado más elevado de estudios realizados o en curso.	1: Superior 2: Secundaria 3: Sin educación o Primaria	Proporción de niños de madres en determinado nivel educativo	HC61 / RECH6
Sexo (X02)	Conjunto de características biológicas de las personas en estudio que los definan como hombres y mujeres.	1: Mujer 2: Hombre	Proporción niños según sexo	HC27 / RECH6
Edad (En meses) (X03)	Tiempo transcurrido desde el nacimiento hasta el momento del estudio	6, 7, …, 35	Proporción de niños según edad (en meses)	HC1 / RECH6
Orden de nacimiento del niño (X04)	La madre reporta el orden de nacimiento de su hijo.	1,2, …, 15	Proporción de niños según orden de nacimiento	HC64 / RECH6
Intervalo entre nacimientos anteriores al niño (X05)	Se calcula entre las fechas de nacimientos entre hermanos.	0, 1, 2, …	Proporción de niños según intervalo entre nacimientos.	HC63 / RECH6
Lugar de residencia (X06)	Agrupaciones de viviendas con el objeto de conseguir una repartición de población en grupos homogéneos respecto a una serie de características que atañen a su modo de vida.	1: Urbana 2: Rural	Proporción de niños según área de residencia.	V102 / REC0111
Altitud del conglomerado (en metros) (X07)	Se tomará como referencia las alturas a las que se encuentran los centros poblados, capturados por el INEI.	0, 1, 2, ……	Proporción de niños según altitud del conglomerado	V040 / REC0111
Número de niños menores de 5 años (X08)	Es el número total de niños que residen en la misma vivienda y comparte una misma economía.	0, 1, 2, …	Proporción de niños según número de niños menores de cinco años en el hogar	HV014 / RECH0
Número de miembros del hogar (X09)	Conjunto de personas que viven en el mismo hogar. Número de miembros del hogar.	2, 3, 4, …	Proporción de niños provenientes de hogares según el número de personas que lo conforman	V136 / REC0111

Tabla 3. *(Continuación a …)*

Variables	Definición conceptual	Definición operacional (Categoría)	Indicadores	Variable original / Modulo
Región natural (X10)	Área continua o discontinua, en la cual son comunes o similares el mayor número de factores del medio ambiente natural.	1: Lima Metropolitana 2: Resto de costa 3: Selva 4: Sierra	Proporción de niños en cada región	SREGION/ REC91
Índice de riqueza (X11)	Se tomará el quintil de riqueza al que pertenece el hogar de donde procede el niño.	1: Más rico 2: Rico 3: Medio 4: Pobre 5: Más pobre	Proporción de niños en cada quintil de bienestar	V190 / REC0111
Fuente principal de abastecimiento de agua potable que utilizan en su hogar para tomar o beber (X12)	"Comprende la unión física entre la red de agua y el límite del predio a través de un tramo de tubería que incluye la caja del medidor" (Sunass, 2006).	1: Dentro de la vivienda 2: Agua embotellada y otro 3: Pilón, pozo, manantial, río, camión y lluvia.	Proporción de niños provenientes de hogares con conexión domiciliaria de agua potable.	V113 / REC0111
Tipo de instalación sanitaria (X13)	"Comprende la unión física (instalación de tubería y accesorios) entre la red matriz de agua y el límite de propiedad del predio a través de una tubería que incluye la caja de control y su medidor" (Sedapal, 2015).	1: Vivienda interior 2: Letrina ventilada y pozo séptico 3: Vivienda exterior y latrina (ciego o negro) 4: Río, canal, lago, sin servicio y otro.	Proporción de niños provenientes de hogares con conexión domiciliaria de desagüe.	V116 / REC0111
Material predominante del piso de la vivienda (X14)	"Se refiere al material del cual está hecha la mayor parte de los pisos del edificio o casa donde está ubicada la vivienda" (INEI, 2017).	1: Láminas asfálticas y losetas 2: Cemento / Ladrillo 3: Madera o parquet 4: Tierra / Arena, otro	Proporción de niños provenientes de hogares con viviendas de piso de cemento.	HV213 / RECH23
El agua usualmente es tratada por: hervida (X15)	La madre reporta que hierve el agua antes de consumir.	1: Si 2: No	Proporción de niños cuyas madres reportan que hierben el agua antes de consumir.	HV237A / RECH23
Edad de la madre (en años) (X16)	Proporción de niños según edad de la madre.	12, 13, 14, …, 49	Proporción de niños según edad de la madre.	HA1 / RECH5
Talla de la madre (en centímetros) (1 decimal) (X17)	Talla obtenida de la ENDES	139, 140, 141, …	Proporción de niños según talla de la madre.	HA3 / RECH5

Tabla 3. *(Continuación b …)*

Variables	Definición conceptual	Definición operacional (Categoría)	Indicadores	Variable original / Modulo
Nivel de anemia en la madre (X18)	Medición de hemoglobina en sangre (Hb corregida <12.0 g/dL) Anemia en madres	1: Sin anemia 2: Moderado, grave o leve	Proporción de niños cuya madre presenta anemia.	HA57 / RECH5
Etnicidad (X19)	La madre reporta cual es el idioma que habitualmente hablan en su hogar.	1: Castellano u otra lengua extranjera 2: Quechua 3: Aimara 4: Otra lengua nativa u originaria.	Proporción de niños hijos de madres que hablan algún idioma indígena	V131 / REC0111
Persona que normalmente alimenta al niño (X20)	La información es entregada o reportada por la entrevistada	1: Empleada doméstica 2: Padres / Suegros u otros 3: Madre e hijas / hijos mayores	Proporción de niños según el parentesco con la persona que normalmente lo alimenta	S496 / REC91
Niño tomó hierro en jarabe, polvo, gotas u otra presentación (X21)	La madre reporta si su hijo tomó hierro en pastillas o jarabes.	1: No 2: Si	Proporción de niños que tomaron hierro en pastillas o jarabes.	S465EA + S465EB + S465EC + S465ED / REC95
Le hicieron algún control de crecimiento y desarrollo (X22)	Le hicieron algún control de crecimiento y desarrollo	1: No 2: Si	Proporción niños que tuvieron un control CRED	S466 / REC95
Ha tenido fiebre en las últimas dos semanas (X23)	Se reporta por la madre de los niños menores de 6 a 35 meses, si el niño presentó fiebre las dos últimas semanas.	1: No 2: Si	Proporción de niños con fiebre en las dos últimas semanas	H22 / REC43
En los últimos 14 días, ha tenido diarrea la niña(o) (X24)	La madre reporta si su hijo presentó diarrea en las dos últimas semanas.	1: No 2: Si	Proporción de niños con presencia de diarrea en dos últimas semanas.	H11 / REC43
Ha tenido tos en las últimas dos semanas (X25)	La madre reporta si su hijo presentó tos en las dos últimas semanas.	1: No 2: Si	Proporción de niños con presencia de tos en dos últimas semanas.	H31 / REC43
Alguna vez recibió la dosis de vitamina A (X26)	La madre reporta si su hijo recibió alguna vez vitamina A.	1: No 2: Si	Proporción de niños con recepción de vitamina A.	H41B / REC43
Medicamentos para parásitos intestinales en los últimos 6 meses (X27)	Se reporta por la madre de los niños menores de 6 a 35 meses, si el niño recibió tratamiento antiparasitario en los últimos seis meses.	1: Si 2: No	Proporción de niños que recibieron tratamiento para parásitos (lombrices)	H43 / REC43

Tabla 3. *(Continuación c …)*

Variables	Definición conceptual	Definición operacional (Categoría)	Indicadores	Variable original / Modulo
Visitas prenatales por embarazo (X28)	Información obtenida por tarjeta de control o reporte de la madre.	0, 1, 2, …, 20	Proporción niños según el número de visitas prenatales que tuvo su madre durante su embarazo	M14 / REC41
Momento del primer control prenatal (X29)	La información puede ser extraída usando la tarjeta de control de la madre o por la misma madre.	0, 1, 2, …, 9	Proporción de niños según el momento del primer control gestacional durante el embarazo en la madre	M13 / REC41
Peso del niño al nacer (kilos - 3 dec.) (X30)	Es el peso del niño que le toman inmediatamente después de haber nacido.	500, 501, 502, …	Proporción de niños según peso al nacer.	M19 / REC41
Parto institucional (X31)	Es el parto realizado en establecimiento de salud (EESS) con la ayuda de un médico, enfermera u obstetra.	1: Institucional 2: No institucional	Proporción de niños producto de un parto institucional	M3A, M3B y M3C / REC41
Por cuantos días tomó hierro y/o cuantas inyecciones recibió (X32)	Reporte de la madre cuando se le preguntó sobre el tiempo de consumo de suplemento de hierro en la gestación.	0, 1, 2, …	Proporción de niños según el número de días que su madre tomó suplemento de hierro	M46 / REC41
El día de ayer o durante el día o la noche cuantas veces le dio comida sólidas o semisólidas (X33).	La madre reporta si su hijo consumió algún tipo de carne el día de ayer (¿Cuántas veces?).	0, 1, 2, …	Proporción de niños que consumieron algún tipo de carne el día de ayer.	M39 / REC41

3.3. Métodos de la investigación

“El método utilizado en la presente investigación es el cuantitativo o tradicional, ya que se fundamenta en la medición de las características de los fenómenos sociales, lo cual supone derivar de un marco conceptual pertinente al problema analizado una serie de postulados que expresen relaciones entre las variables estudiadas de forma deductiva. Este método tiende a generalizar y normalizar resultados” (Bernal, 2016).

3.4. Diseño de la investigación

Se realizó un estudio observacional (no experimental), debido a que no se realizó manipulaciones de variables, sino que se trabajó a partir de variables ya existentes (datos secundarios) (Hernández y Mendoza, 2018) porque los datos fueron descargados de la base de datos del INEI (2015-2019).

3.5. Población y muestra

La población estuvo constituida por 57410 registros de niños de 6 a 35 meses de edad del Perú recolectados a través de la Encuesta Demográfica y de Salud Familiar (ENDES) por el Instituto Nacional de Estadística e Informática (INEI), durante los años 2015 al 2019 que contaban con medición de hemoglobina sanguínea y que fueron descargados del menú “Bases de Datos” de la página web del INEI (http://iinei.inei.gob.pe/microdatos/).

En el presente estudio la muestra fue igual a la población ya que las técnicas de “machine learning” (“random forest”) han sido creadas para trabajar con grandes bases de datos.

3.6. Actividades del proceso investigativo

Se descargo la información de los registros de niños recogida con la ENDES, ubicada en la página web del INEI en la sección "Base de datos" (INEI, 2015-2019). Las bases de datos de la ENDES se encuentran organizadas por módulos de acuerdo al rubro de la información, escogiéndose los módulos: RECH6, REC0111, RECH0, REC91, RECH23, RECH5, REC95, REC43 y REC41 (ver tabla 3), que miden características de la anemia en niños, anemia en la madre, salud del niño y de la madre y características sociodemográficas del hogar y la vivienda.

Se juntó (fundió y apiló) los módulos necesarios de las ENDES del 2015 al 2019, para construir una nueva base de datos que contenga las variables necesarias para el análisis de datos, utilizando el software estadístico SPSS.

Se realizó la imputación de los datos faltantes (perdidos) en la base de datos construida a partir de la ENDES, utilizando la librería Hmisc (Ver sección 2.2.7) del software estadístico R con su función *aregImpute().*

Se recategorizó los valores de las variables independientes para poder mantener una estabilidad en el procedimiento alternativo propuesto, donde las categorías de

cada predictor se juntaron si no son significativamente distintas respecto a la variable respuesta, para ello se utilizó la técnica de árboles de decisión CHAID (ver sección 2.1.2.1.1.1) con ayuda del software estadístico R. Está técnica identifica las divisiones óptimas, generando arboles no binarios, es decir, algunas divisiones generan más de dos ramas.

Se examinó los datos recategorizados utilizando tablas estadísticas para un mejor análisis e interpretación de los resultados con los softwares estadísticos SPSS y R.

Se aplicó el muestreo estratificado para dividir la base de datos en dos partes: Train y Test con ayuda de la librería "caret" del software estadístico R.

Se aplicó el algoritmo "Random Forest" con el software estadístico R para predecir la anemia en niños menores de 6 a 35 meses de edad de acuerdo a los siguientes 6 procedimientos alternativos obtenidos a partir de una combinación de los criterios de balanceo de datos y reajuste de parámetros para la predicción de anemia:

- Procedimiento alternativo A: se planteó sin que la variable respuesta este balanceada, considerando todas las variables y utilizando los parámetros del algoritmo "random forest" por defecto.
- Procedimiento alternativo B: se planteó balanceando la variable respuesta, considerando todas las variables y utilizando los parámetros del algoritmo "random forest" por defecto.
- Procedimiento alternativo C: se planteó sin que la variable respuesta este balanceada, utilizando un reajuste de los parámetros, mediante un "*grid search*" para encontrar mejores parámetros, reduciendo la cantidad de árboles a 300 y con todas las variables.
- Procedimiento alternativo D: se planteó balanceando la variable respuesta, utilizando un reajuste de los parámetros, mediante un "*grid search*" para encontrar mejores parámetros, reduciendo la cantidad de árboles a 300 y con todas las variables.
- Procedimiento alternativo E: se planteó sin que la variable respuesta este balanceada, utilizando un reajuste de los parámetros mediante un "*grid search*" para encontrar mejores parámetros del algoritmo "random forest", reduciendo la cantidad de árboles a 300 y con selección de variables.
- Procedimiento alternativo F: se planteó balanceando la variable respuesta,

utilizando un reajuste de los parámetros mediante un “*grid search*” para encontrar mejores parámetros del algoritmo “random forest”, reduciendo la cantidad de árboles a 300 y con selección de variables.

Para el balanceo de la variable respuesta se utilizó el método combinación (“bothsampling”) con ayuda de la librería “ROSE” del software estadístico R.

Se evaluó los seis procedimientos alternativos propuestos con los indicadores de AUC, sensibilidad y especificidad mediante las librerías “caret” y “randomForest” del software estadístico R para saber cuál es el mejor procedimiento alternativo.

Por último, se calculo la importancia relativa de cada una de las variables según procedimiento alternativo utilizando el indicador “Mean Decrease Gini”, que es el promedio (media) de la disminución total de la impureza del nodo de una variable, ponderada por la proporción de muestras que llegan a ese nodo en cada árbol de decisión individual en el “random forest”. Un mayor “Mean Decrease Gini” indica una mayor importancia de la variable.

3.7. Técnicas e instrumentos de la investigación

La técnica de recolección de datos primarios utilizada por el INEI en la ENDES 2015-2019 fue la encuesta, conformada por tres instrumentos: “a) el cuestionario del hogar, que incluye el listado de miembros del hogar y las características de la vivienda; b) el cuestionario individual para mujeres de 15 a 49 años de edad, que recopila datos de reproducción, anticoncepción, embarazo, parto, puerperio y lactancia, inmunización y salud, nupcialidad, preferencias de fecundidad, antecedentes del cónyuge y trabajo de la mujer, HIV/sida y otras infecciones de transmisión sexual, mortalidad materna y violencia doméstica, y c) el cuestionario de traumatismos y enfermedades crónicas, para todas las personas de 40 o más años de edad. Además, incluye un módulo en el que se registran todas las mediciones y pruebas que se hacen: peso y talla para mujeres de 15 a 49 años de edad y para niños de 0 a 5 años de edad; prueba de hemoglobina para mujeres de 15 a 49 años de edad y para niños de 0 a 5 años de edad, y medición de la presión arterial para las personas de 40 o más años de edad” (ENDES, 2015-2019). Estos cuestionarios se encuentran en el menú “Bases de datos” de la página web del INEI.

3.8. Procedimiento para la recolección de datos

Los datos fueron descargados del menú “Bases de Datos” de la página web del INEI (http://iinei.inei.gob.pe/microdatos/), en la sección Encuesta Demográfica y de Salud Familiar – ENDES, escogiéndose los módulos que contenían información sobre las variables mencionadas en la tabla 3: RECH6, REC0111, RECH0, REC91, RECH23, RECH5, REC95, REC43 y REC41, para los años 2015 al 2019.

3.9. Técnicas de procesamiento y análisis de datos

Los datos se analizaron con los softwares estadísticos IBM SPSS versión 26 para la limpieza de la base de datos y R versión 3.6.1 para la corrida de los diferentes procedimientos alternativos propuestos.

En la presente investigación se estimó diversos procedimientos alternativos de clasificación de datos, teniendo en cuenta la metodología CRISP-DM (Cross Industry Standard Process for Data Mining). Esta es una metodología de minería de datos para el desarrollo de proyectos analíticos. CRISP-DM consta de 6 fases que son necesarias para abordar un proyecto de analítica avanzada con las máximas garantías posibles:

- **Comprender el negocio:** Los niños menores de tres años de edad presentan un elevado riesgo de anemia ferropénica debido a la alta demanda de hierro para su crecimiento, y además por el aporte insuficiente y baja disponibilidad del mineral en la dieta, debiéndose principalmente a variables: sociodemográficas, relacionadas con el niño y del cuidado materno e infantil.
- **Conocer los datos:** Las bases de datos de la ENDES están organizadas en módulos, necesarios para el análisis de distintas variables de salud, siendo una de ellas la tenencia de anemia en niños.
- **Limpieza de datos:** Las bases de datos de las ENDES cuentan con datos perdidos, donde la persona encuestada muchas veces no quiso responder determinadas preguntas o no le correspondían, por lo que se hace necesario una limpieza de datos, eliminando registros para los casos donde no se contaba con el dato de la tenencia de anemia e imputando datos perdidos de las diversas variables elegidas para el análisis y que pertenecen a registros donde se contaba con información de la tenencia de anemia.

- **Modelización:** Se utilizó el algoritmo “Random Forest” para determinar el mejor procedimiento alternativo de clasificación de datos a partir de los 6 procedimientos alternativos propuestos.
- **Evaluación:** Para evaluar los procedimientos alternativos de clasificación de datos propuestos se utilizó los indicadores AUC, sensibilidad y especificidad.
- **Implementación:** El mejor procedimiento alternativo de clasificación de datos para predecir la tenencia de anemia será informado a los encargados de elaborar las políticas públicas del Perú, de tal manera que diseñen de una mejor manera la implementación de los diferentes proyectos que se estén desarrollando en bien de la reducción de la tenencia de anemia en niños a nivel nacional.

CAPITULO IV

RESULTADOS Y DISCUSIÓN

Se realizó un análisis del comportamiento univariado y bivariado de las variables asociadas a la anemia así como también un análisis multivariado para encontrar el mejor procedimiento alternativo de clasificación por medio del algoritmo “random forest”, cuyos resultados etapa por etapa se presentan a continuación:

4.1. Análisis descriptivo de la tenencia de anemia según variables independientes

En el anexo A se presenta los resultados al procesar la base de datos imputada a partir de la información recogida con la ENDES cuentan con 33 variables independientes y la variable respuesta (anemia). La variable cuantitativa edad del niño (X3) tiene una mediana de 21 meses, es decir, el 50% de los datos se encuentran alrededor de los 21 meses y la media de la edad es de 20,76 ≈ 21 meses por lo que se sospecha que tiende a una distribución simétrica. La edad de la madre (X16) en promedio es de 30,42 ≈ 30 años y su mediana es de 30 años con lo que se llega a sospechar que también tiende a una distribución simétrica. El peso del niño al nacer (X30) presenta un promedio de 3246,44 gramos y una mediana de 3250 gramos, mientras tanto las variables cualitativas lugar de residencia tiene 2 categorías, índice de riqueza tiene 5 categorías, etnicidad tiene 4 categorías, etc.

Otra de las actividades que se realizó en el presente estudio fue la recategorización de cada variable independiente respecto a la variable respuesta utilizando el algoritmo CHAID (ver Anexo B).

Los árboles de decisión ordenan cada variable ya sea cuantitativa o cualitativa en grupos con propensión hacia la categoría de interés. Como una ilustración de la recategorización, en la figura B1 del Anexo B, se presenta la recategorización de la variable nivel educativo más alto de la madre (X01), por ejemplo el nodo 2 nos muestra en primer lugar la cantidad de casos en el nodo (n=7798), las probabilidades de cada categoría por ejemplo la probabilidad del nodo de ser cero o no tener anemia es 0,673 y la probabilidad de ser uno o tener anemia es 0,327,

donde esté último se tomó en cuenta para el orden de la variable recategorizada, es decir al nivel educativo más alto de la madre con educación primaria y sin educación tomara valor “3” puesto que de todos los nodos del árbol, es el que tiene mayor probabilidad de tener anemia y así sucesivamente se recodificará cada nodo (ver figura B1 del Anexo B).

En el anexo B se presentan las recategorizaciones del resto de variables independientes.

En la Tabla 4 se presentan frecuencias de todas las variables recategorizadas mediante la metodología con árboles de decisión CHAID junto con los porcentajes de niños con anemia para cada categoría de las variables. De acuerdo a los datos de la variable nivel educativo más alto de las madres de los niños (X01), la mayoría tiene una educación secundaria (64,6%), un menor porcentaje se observa en el nivel superior (13,6%), un 21,8% de los niños tiene madres sin educación o con educación primaria. Un mayor porcentaje de niños con anemia se encontró en aquellos hijos de madres sin educación o con educación primaria (52,8%).

Por otro lado, en la variable sexo del niño (X02) se observa que un mayor porcentaje de niños en la muestra eran hombres (51,3%), asimismo, más niños anémicos se encontró en los hombres (47,7%) comparados con las mujeres (43,6%).

En cuanto a la edad de los niños, se muestra que un mayor porcentaje tuvieron de 6 a 16 meses (35,5%), seguido por los que tuvieron más de 30 hasta los 35 meses (17,4%) y los que tuvieron más de 25 hasta 30 meses (17,0%). Un mayor porcentaje de niños con anemia se encontró en el grupo con una edad de 6 hasta 16 meses (61,1%), seguido por el grupo que tenían más de 16 hasta 18 meses (53,6%). A medida que la edad era menor en los niños el porcentaje de niños anémicos es menor.

En el orden de nacimiento, la mayoría de niños son de primer orden (33,1%), seguido por aquellos que son de segundo orden al nacimiento (30,5%), solo un 5% fueron de un orden mayor a 5. Por consiguiente, un 56,0% de los niños de un orden mayor al 5 eran anémicos, conforme el orden al nacimiento del niño se incrementaba el porcentaje de niños anémicos en dicho grupo también se incrementaba.

En lo que respecta al intervalo entre nacimientos anteriores al niño, la mayoría han tenido un intervalo no mayor a 11 meses (33,3%), seguido por aquellos que han tenido más de 56 hasta 109 meses (24,2%), asimismo un 21,5% de los niños han tenido un intervalo mayor a 11 meses hasta 41 meses, sólo un 10,5% han tenido un intervalo entre nacimientos mayor a 109 meses. De esta manera, un 52,7% de los niños con un intervalo entre nacimientos mayor de 11 hasta 41 meses tenían anemia, además un 49,2% de los niños con un intervalo entre nacimientos mayor a 41 hasta 56 meses estaban anémicos.

Tabla 4.

Porcentaje de niños con anemia según recategorización de las variables independientes del periodo 2015-2019 (INEI, 2015-2019)

Variables independientes	Categoría	Niños	Porcentaje	Niños con anemia	% de niños con anemia
Nivel educativo más alto de la madre					
Superior	1	7798	13.6	2548	32.7%
Secundaria	2	37072	64.6	17076	46.1%
Sin educación o Primaria	3	12540	21.8	6622	52.8%
Sexo del niño					
Mujer	1	27983	48.7	12199	43.6%
Hombre	2	29427	51.3	14047	47.7%
Edad del niño (En meses)					
>30	1	9971	17.4	2854	28.6%
<25 - 30]	2	9735	17.0	3181	32.7%
<22 - 25]	3	5789	10.1	2205	38.1%
<20 - 22]	4	3807	6.6	1623	42.6%
<19 - 20]	5	1982	3.5	900	45.4%
<18 - 19]	6	1902	3.3	970	51.0%
<16 - 18]	7	3824	6.7	2051	53.6%
<=16]	8	20400	35.5	12462	61.1%
Orden de nacimiento del niño					
<=1	1	19024	33.1	7960	41.8%
<1 - 2]	2	17501	30.5	7813	44.6%
<2 - 3]	3	10389	18.1	4926	47.4%
<3 - 5]	4	7640	13.3	3947	51.7%
>5]	5	2856	5.0	1600	56.0%
Intervalo entre nacimientos (en meses)					
>109	1	6030	10.5	2397	39.8%
<=11]	2	19137	33.3	8050	42.1%
<56 - 109]	3	13904	24.2	6338	45.6%
<41 - 56]	4	6005	10.5	2955	49.2%
<11 - 41]	5	12334	21.5	6506	52.7%
Lugar de residencia					
Urbana	1	40677	70.9	17378	42.7%
Rural	2	16733	29.1	8868	53.0%

Tabla 4. *(Continuación a …)*

Variables independientes	Categoría	Niños	Porcentaje	Niños con anemia	% de niños con anemia
Altitud del conglomerado (en metros)					
<=74	1	11933	20.8	4760	39.9%
<99 - 132]	2	2193	3.8	894	40.8%
<378 - 3046]	3	18815	32.8	7809	41.5%
<74 - 99]	4	2808	4.9	1270	45.2%
<132 - 378]	5	10750	18.7	5136	47.8%
<3046 - 3403]	6	4529	7.9	2388	52.7%
<3403 - 3753]	7	2929	5.1	1709	58.3%
>3753	8	3453	6.0	2280	66.0%
N° de niños menores de 5 años de edad en el hogar					
<=1	1	39194	68.3	17070	43.6%
2	2	15214	26.5	7523	49.4%
3	3	2423	4.2	1296	53.5%
>3	4	579	1.0	357	61.7%
N° de miembros del hogar					
<=3	1	10889	19.0	4650	42.7%
4	2	14615	25.5	6454	44.2%
<4 - 6]	3	20475	35.7	9420	46.0%
<6 - 8]	4	7516	13.1	3649	48.5%
<8 - 12]	5	3503	6.1	1827	52.2%
>12	6	412	0.7	246	59.7%
Región natural					
Lima metropolitana	1	6729	11.7	2460	36.6%
Resto Costa	2	17481	30.4	7047	40.3%
Selva	3	14403	25.1	7031	48.8%
Sierra	4	18797	32.7	9708	51.6%
Índice de riqueza					
Más rico	1	5629	9.8	1656	29.4%
Rico	2	8667	15.1	3145	36.3%
Medio	3	11983	20.9	5139	42.9%
Pobre	4	16204	28.2	8147	50.3%
Más pobre	5	14927	26.0	8159	54.7%
Fuente principal de abastecimiento de agua potable que utilizan en su hogar para tomar o beber					
Dentro de la vivienda	1	39891	69.5	17467	43.8%
Agua embotellada y otro	2	7435	13.0	3474	46.7%
Pilón, Pozo, Manantial, Río, Camión y LLuvia	3	10084	17.6	5305	52.6%
Tipo de instalación sanitaria					
Vivienda interior	1	32289	56.2	12920	40.0%
Letrina ventilada y pozo séptico	2	5199	9.1	2542	48.9%
Vivienda exterior y Latrina (ciego o negro)	3	14527	25.3	7732	53.2%
Río, Canal, Lago, Sin servicio y otro	4	5395	9.4	3052	56.6%

Tabla 4. *(Continuación b …)*

Variables independientes	Categoría	Niños	Porcentaje	Niños con anemia	% de niños con anemia
Material predominante del piso de la vivienda					
Láminas ásfalticas y Losetas	1	7839	13.7	2537	32.4%
Cemento / Ladrillo	2	27427	47.8	12042	43.9%
Madera o parquet	3	5178	9.0	2595	50.1%
Tierra / Arena, Otro	4	16966	29.6	9072	53.5%
El agua usualmente es tratada por: hervida					
Si	1	44971	78.3	20318	45.2%
No	2	12439	21.7	5928	47.7%
Edad de la madre (en años)					
>30	1	27277	47.5	11512	42.2%
<25 - 30]	2	12709	22.1	5784	45.5%
<22 - 25]	3	6833	11.9	3320	48.6%
<19 - 22]	4	5632	9.8	2865	50.9%
<=19	5	4959	8.6	2765	55.8%
Talla de la madre (en cm)					
>163,3	1	1504	2.6	519	34.5%
<159,1 - 163,3]	2	4486	7.8	1746	38.9%
<153,5 - 159,1]	3	16084	28.0	6950	43.2%
<151,8 - 153,5]	4	6981	12.2	3185	45.6%
<145,3 - 151,8]	5	22077	38.5	10610	48.1%
<=145,3	6	6278	10.9	3236	51.5%
Nivel de anemia en la madre					
Sin anemia	1	45313	78.9	19412	42.8%
Moderado, grave o leve	2	12097	21.1	6834	56.5%
Etnicidad					
Castellano u otra lengua extranjera	1	30257	52.7	13135	43.4%
Quechua	2	24361	42.4	11387	46.7%
Aimara	3	1842	3.2	1101	59.8%
Otra lengua nativa u originaria	4	950	1.7	623	65.6%
Persona que normalmente alimenta al niño					
Empleada doméstica	1	636	1.1	189	29.7%
Padres/Suegros u otros	2	12473	21.7	4632	37.1%
Madre e hijas/hijos mayores	3	44301	77.2	21425	48.4%
Niño tomó hierro en jarabe, polvo, gotas u otra presentación					
No	1	40643	70.8	18118	44.6%
Si	2	16767	29.2	8128	48.5%
Le hicieron algún control de crecimiento y desarrollo					
No	1	11802	20.6	5343	45.3%
Si	2	45608	79.4	20903	45.8%

Tabla 4. *(Continuación c …)*

Variables independientes	Categoría	Niños	Porcentaje	Niños con anemia	% de niños con anemia
Ha tenido fiebre en las últimas dos semanas					
No	1	45460	79.2	20449	45.0%
Si	2	11950	20.8	5797	48.5%
En los últimos 14 días, ha tenido diarrea la niña(o)					
No	1	48887	85.2	21959	44.9%
Si	2	8523	14.8	4287	50.3%
Ha tenido tos en las últimas dos semanas					
No	1	37994	66.2	17273	45.5%
Si	2	19416	33.8	8973	46.2%
Alguna vez recibió la dosis de vitamina A					
No	1	42050	73.2	19123	45.5%
Si	2	15360	26.8	7123	46.4%
Medicamentos para parásitos intestinales en los últimos 6 meses					
Si	1	16364	28.5	6686	40.9%
No	2	41046	71.5	19560	47.7%
Visitas prenatales por embarazo					
>15	1	934	1.6	292	31.3%
<12 - 15]	2	4672	8.1	1781	38.1%
<9 - 12]	3	19520	34.0	8517	43.6%
<7 - 9]	4	15162	26.4	6981	46.0%
<5 - 7]	5	10517	18.3	5177	49.2%
<3 - 5]	6	4367	7.6	2237	51.2%
<=3	7	2238	3.9	1261	56.3%
Momento del primer control prenatal					
<=1	1	15306	26.7	6083	39.7%
<1 - 2]	2	16699	29.1	7472	44.7%
<2 - 3]	3	13007	22.7	6189	47.6%
<3 - 4]	4	5682	9.9	2835	49.9%
<4 - 6]	5	5080	8.8	2702	53.2%
>6	6	1636	2.8	965	59.0%
Peso del niño al nacer (en gramos)					
<= 1640	1	460	0.8	181	39.3%
> 3825	2	6674	11.6	2796	41.9%
<3113 - 3825]	3	28257	49.2	12709	45.0%
<1640 - 3113]	4	22019	38.4	10560	48.0%
Parto institucional					
Institucional	1	53944	94.0	24290	45.0%
No institucional	2	3466	6.0	1956	56.4%
Por cuantos días tomó hierro y/o cuantas inyecciones recibió					
> 209	1	7775	13.5	3040	39.1%
<116 - 209]	2	19176	33.4	8311	43.3%
<44 - 116]	3	17088	29.8	8101	47.4%
<= 44	4	13371	23.3	6794	50.8%

Tabla 4. *(Continuación d …)*

Variables independientes	Categoría	Niños	Porcentaje	Niños con anemia	% de niños con anemia
El día de ayer o durante el día o la noche cuantas veces le dio comida sólidas o semisólidas					
0	1	1068	1.9	465	43.5%
> 4	2	26187	45.6	10933	41.7%
4	3	15273	26.6	6994	45.8%
3	4	11659	20.3	6050	51.9%
[1-2]	5	3223	5.6	1804	56.0%

Un 70,9% de los niños vivían en el área urbana y el resto en el área rural (29,1%). Por otro lado, un 53,0% de los niños que vivían en el área rural tenían anemia y en el área urbana este porcentaje fue de 42,7%.

La mayoría de los niños viven en conglomerados a una altitud de más de 378 hasta 3046 metros (32,8%), asimismo un 20,8% vivía en conglomerados que se encuentran a una altitud hasta los 74 metros. Se encontró que un 66,0% de los niños que residen en un conglomerado por encima de los 3753 metros tenían anemia, de igual manera, con los niños que residen en conglomerados que se encuentran a una altitud mayor a 3403 hasta 3753 metros, un 58,3% se encuentran anémicos.

Un 68,3% de los niños vivían en hogares con un niño menor de 5 años, asimismo un 26,5% de los niños vivían en hogares con dos niños menores de 5 años. Por otro lado, un mayor porcentaje de niños anémicos se encuentran en hogares con más de 3 niños menores de 5 años (61,7%), seguido por el grupo de niños de hogares con 3 niños menores de 5 años, donde el 53,5% tienen anemia.

El mayor porcentaje de niños forman parte de hogares compuestos por más de 4 hasta 6 miembros (35,7%), seguido por un 25,5% de niños que habitan en hogares con 4 miembros, asimismo, un 19,0% de los niños residían en hogares con hasta 3 miembros. Por otro lado, un 59,7% de los niños que residían en hogares con más de 12 miembros tenían anemia. Se observa que a medida que el número de miembros por hogar aumenta, el porcentaje de niños con anemia también aumenta (Moschovis et al, 2018).

La mayoría de niños residen en la Sierra (32,7%), seguido por un 30,4% que residen en el resto de la Costa, un 25,1% residen en la Selva y un menor porcentaje residen

en Lima Metropolitana (11,7%). Asimismo, un mayor porcentaje de niños con anemia se observa en la región de la Sierra (51,6%), seguido por la Selva (48,8%), esta variable estaría relacionada con estructurales de exclusión social en el país (MIDIS, 2012).

En cuanto al índice de riqueza, se observa que la mayoría de niños habitan hogares catalogados como pobres (28,2%), seguidos por aquellos que habitan hogares más pobres (26,0%), un menor porcentaje de niños se encontraron en hogares más ricos (9,8%). De igual manera, un mayor porcentaje de niños con anemia se encontró en hogares más pobres (54,7%), seguido por niños de hogares pobres, donde el 50,3% tenían anemia. En resumen, conforme el índice de riqueza disminuye el porcentaje de niños con anemia aumenta (Moschovis et al, 2018).

Analizando la fuente principal de abastecimiento de agua potable que utilizan los hogares de los niños para tomar, se observa que la mayoría tiene su fuente principal dentro de la vivienda (69,5%), mientras que el agua embotellada u otro solo es fuente principal para un 13,0% de los hogares de los niños. De igual modo, un 52,6% de los niños que habitan hogares donde su fuente principal de abastecimiento de agua potable es pilón, pozo, manantial, río, camión y lluvia se encuentran anémicos, seguidos por un 46,7% de niños que habitan hogares donde su fuente principal de abastecimiento es agua embotellada y otro son anémicos, finalmente, un menor porcentaje de niños con anemia habitan hogares donde su fuente principal de abastecimiento es dentro de la vivienda (43,8%).

En el caso del tipo de instalación sanitaria, se observa que un 56,2% del total de niños tiene su instalación sanitaria en el interior de la vivienda, seguido por un 25,3% que tiene su instalación sanitaria en el exterior de la vivienda y latrina (ciego o negro). Por consiguiente, un 56,6% de los niños que tienen su instalación sanitaria en el río, canal, lago, sin servicio y otro tienen anemia, además un 53,2% de los niños cuya instalación sanitaria es en el exterior de la vivienda y latrina (ciego o negro) se encuentran anémicos. En resumen, mientras la instalación sanitaria se encuentre más al exterior de la vivienda mayor será el porcentaje de niños con anemia (Moschovis et al, 2018).

En cuanto al material predominante del piso de la vivienda, la mayoría de niños reside en viviendas con piso de cemento / ladrillo (47,8%) y un menor porcentaje de

niños se encuentra en viviendas con piso de madera o parquet (9,0%). También, el 53,5% de los niños que residen en viviendas donde el piso es de tierra/arena u otro se encuentran anémicos, en cambio, en niños que residen en viviendas cuyo piso es de láminas asfálticas y losetas el porcentaje de anémicos es de 32,4%, en conclusión, a medida que el piso de la vivienda es más pobre, el porcentaje de niños con anemia aumenta.

La mayoría de niños provienen de hogares donde el agua la hierven antes de tomarla (78,3%). Por otro lado, un mayor porcentaje de niños anémicos se observa en hogares donde el agua no la hierven antes de tomarla (47,7%).

Un 47,5% del total de niños son hijos de madres con más de 30 años de edad, seguidos por los hijos de madres con edades de más de 25 hasta 30 años (22,1%). Por otro lado, un mayor porcentaje de niños anémicos se encuentran en los que son hijos de madres con edades hasta 19 años (55,8%), seguido por aquellos que son hijos de madres con edades mayor a 19 años hasta 22 años (50,9%). Resumiendo, se observa que a medida que las madres tienen menor edad, el porcentaje de niños anémicos aumenta (Moschovis et al, 2018).

También se estudia el nivel de anemia en la madre, donde el 78,9% de niños son hijos de madres sin anemia y el resto son hijos de madres con un nivel de anemia moderado, grave o leve (21,1%). Por otro lado, un alto porcentaje de niños anémicos son hijos de madres con anemia moderada, grave o leve (56,5%), en cambio, en los niños de madres sin anemia, un 42,8% tienen anemia.

En cuanto a la etnicidad de los niños, un 52,7% de los niños provienen de hogares donde la lengua materna es el castellano u otra lengua externa, seguido por un 42,4% del total de niños cuya lengua materna es el quechua. Por ende, un 65,6% de los niños provenientes de hogares cuya lengua materna es otra lengua nativa u originaria se encuentran anémicos, asimismo, el 59,8% de niños cuya lengua materna es el aimara se encuentran anémicos. En niños provenientes de hogares donde la lengua materna es el castellano u otra lengua extranjera el porcentaje de anémicos es menor (43,4%).

Asimismo, se observa que la mayoría de los niños (77,2%) son alimentados normalmente por la madre e hijas/hijos mayores, seguidos por un 21,7% del total de niños que son alimentados normalmente por los padres/suegros u otros. Sin

embargo, el 48,4% de los niños alimentados normalmente por la madre e hijas/hijos mayores tienen anemia, por otro lado, el 37,1% de los niños alimentados normalmente por los padres/suegros u otros tienen anemia. En el caso de los niños alimentados normalmente por la empleada doméstica el 29,7% se encontraban anémicos.

El 70,8% del total de niños no tomó hierro en jarabe, polvo, gotas u otra presentación; el resto si tomó hierro. Sin embargo, el 48,5% del total de niños que si tomaron hierro son anémicos.

Al 79,4% de los niños le hicieron algún control de crecimiento y desarrollo, sin embargo, el porcentaje de niños con anemia tanto en los que no les hicieron algún control de crecimiento y desarrollo como en los que si les hicieron el control son parecidos (45,3% versus 45,8%).

Un 79,2% de los niños no han tenido fiebre en las últimas dos semanas, el resto si tuvo fiebre. En consecuencia, un 48,5% de los niños que si tuvieron fiebre en las últimas semanas tuvieron anemia frente a un 45,0% de niños anémicos en aquellos que no tuvieron fiebre en las últimas dos semanas.

Se observa que un 85,2% del total de niños no han tenido diarrea en los últimos 14 días, el resto si tuvo diarrea. Por consiguiente, un 50,3% de los niños que si tuvieron diarrea en los últimos 14 días se encontraban anémicos, en cambio, en los niños que no tuvieron diarrea en los últimos 14 días el porcentaje de anémicos fue menor (44,9%).

El 66,2% del total de niños no tuvieron tos en las últimas dos semanas, el resto si tuvo tos. Por ende, el 46,2% del total de niños que si han tenido tos se encontraban anémicos, en cambio, en los niños que no tuvieron tos en las últimas dos semanas el porcentaje de niños anémicos fue menor (45,5%).

De igual manera, el 73,2% del total de niños alguna vez ha recibido la dosis de vitamina A, el resto no ha recibido la dosis de vitamina A. Sin embargo, el porcentaje de niños anémicos es casi similar al comparar los que no recibieron la dosis de vitamina A con los que si la recibieron (45,5% frente a 46,4%).

El 71,5% del total de niños no han recibido tratamiento con medicamentos para parásitos intestinales en los últimos 6 meses, el resto si ha recibido. Por ende, el 47,7% del total de niños que no recibieron el tratamiento se encuentran anémicos en

comparación con el 40,9% del total de niños que si recibieron el tratamiento con medicamentos para parásitos que también se encuentran anémicos.

El 34,0% del total de niños son hijos de madres con más de 9 hasta 12 visitas prenatales durante su embarazo del niño, seguido por un 26,4% de niños hijos de madres con más de 7 hasta 9 visitas prenatales durante el embarazo del niño. Un menor porcentaje de niños son hijos de madres que recibieron hasta 3 controles prenatales por embarazo (3,9%). En consecuencia, el porcentaje más alto de niños anémicos se encuentra en el grupo de hijos de madres que recibieron hasta 3 visitas prenatales durante su embarazo (56,3%), seguido por aquellos niños hijos de madres que recibieron más de tres hasta cinco visitas prenatales durante su embarazo (51,2%). En conclusión, mientras menos visitas prenatales tenga la madre durante su embarazo mayor será el porcentaje de niños con anemia.

Un 29,1% del total de niños son hijos de madres cuyo primer control prenatal fue en el segundo mes, seguido por aquellos niños hijos de madres cuyo primer control prenatal fue en el primer mes, asimismo, un 22,7% del total de niños son hijos de madres cuyo primer control prenatal fue en el tercer mes. Sin embargo, un 59,0% del total de niños hijos de madres cuyo primer control prenatal fue después de los seis meses de embarazo fueron anémicos, seguido por un 53,2% del total de niños hijos de madres cuyo primer control prenatal fue en el quinto o sexto mes de embarazo fueron anémicos. En conclusión, mientras más se demore la madre en hacer su primer control prenatal durante su embarazo mayor será la probabilidad que el niño sufra de anemia.

La mayoría de niños tenía un peso mayor a 3113 hasta 3825 gramos (49,2%) seguido por un 38,4% que tuvieron un peso mayor a 1640 hasta 3113 gramos, un menor porcentaje de niños se encuentran en el intervalo con hasta 1640 gramos. Por consiguiente, un mayor porcentaje de niños anémicos se encuentra en aquellos que han nacido con un peso mayor a 1640 hasta 3113 gramos (48,0%), así como también, en aquellos que nacieron con un peso de 3113 a 3825 gramos, donde el 45,0% se encontraron anémicos.

El análisis del parto institucional arroja que el 94,0% del total de niños provienen de un parto institucional, es decir, que fue realizado con el apoyo de un médico, enfermera u obstetra, el resto de niños provienen de un parto no institucional. Sin

embargo, el 56,4% del total de niños provenientes de un parto no institucional se encontraron anémicos, y en los niños de partos institucionales este porcentaje fue del 45,0%.

De la misma manera, se observa que el 33,4% del total de niños tomaron hierro y/o recibieron inyecciones por más de 116 a 209 días, asimismo, un 29,8% los recibió por un periodo de más de 44 hasta 116 días, un menor porcentaje de niños los recibieron por más de 209 días (13,5%). Por ende, un 50,8% del total de niños que recibieron hierro y/o recibieron inyecciones por hasta 44 días se encontraron anémicos, seguidos por un 47,4% de niños que recibieron hierro y/o inyecciones por más de 44 hasta 116 días se encuentran anémicos. Finalmente, a medida que el número de días en que los niños toman hierro disminuye, el porcentaje de niños anémicos se incrementa.

Además, se observa que el 45,6% del total de niños recibieron más de 4 comidas sólidas o semisólidas un día anterior a la encuesta, por otro lado, un 26,6% del total de niños recibieron 4 comidas sólidas o semisólidas, asimismo, un 20,3% del total de niños recibieron 3 comidas sólidas o semisólidas. Por ende, el 56,0% del total de niños que recibieron de 1 a 2 comidas sólidas o semisólidas se encontraron anémicos, seguidos por un 51,9% del total de niños que recibieron 3 comidas sólidas o semisólidas quienes también se encontraban anémicos.

En la Tabla 5 se observa un porcentaje alto de niños de 6 a 35 meses de edad con anemia (45,7%).

Tabla 5

Distribución de la tenencia de anemia en los niños del periodo 2015-2019 (INEI, 2015-2019)

Prevalencia de Anemia	Categoría	Niños	Porcentaje
Sin anemia	0	31164	54.3
Con anemia	1	26246	45.7

Posteriormente se procedió a la partición muestral estratificada en base a la variable respuesta para garantizar que se mantenga la proporción de la variable respuesta en los datos de entrenamiento (train) y evaluación (test). Los datos se dividieron en un 70% para los datos de entrenamiento y un 30% para los datos de evaluación, es

decir, para observar el desempeño del modelo realizado. Esta partición muestral se realizó mediante la librería “caret”, luego para realizar los procedimientos alternativos y compararlos se utilizó la librería “randomForest” (ver parte muestral del código en el anexo F).

En la Tabla 6 se observa la proporción de las categorías de la variable respuesta, donde claramente al comparar ambas tablas se mantiene la proporción original de la variable respuesta, tanto para los datos de entrenamiento (train) como para la de evaluación (test) (ver código del modelado de la data mediante el algoritmo “random forest” en el anexo F).

Tabla 6

Tabla de contingencia de la variable respuesta (Anemia) para los datos de entrenamiento (Train) y de evaluación (Test)

Anemia	Train		Test	
	Cantidad	% del total	Cantidad	% del total
0	21815	54.3%	9349	54.3%
1	18373	45.7%	7873	45.7%

Donde: 0: Sin anemia; 1: Con anemia.

4.2. Ejecución y comparación de los procedimientos alternativos mediante indicadores

A partir de las tablas de clasificación de las categorías estimadas y la variable respuesta (Ver Anexo C) se calculan los indicadores (AUC, especificidad y sensibilidad).

En la tabla 7 se observan los indicadores (AUC, especificidad y sensibilidad) de los procedimientos alternativos propuestos para el conjunto de datos de evaluación.

Se procedió a realizar el primer procedimiento alternativo de clasificación (procedimiento alternativo A) con el algoritmo “random forest”, donde se planteó sin que la variable respuesta este balanceada y utilizando los parámetros del algoritmo “random forest” por defecto.

El segundo procedimiento alternativo de clasificación (procedimiento alternativo B) con el algoritmo “random forest”, se planteó mediante el balanceo de la variable respuesta y utilizando los parámetros del algoritmo “random forest” por defecto.

Tabla 7

Comparación de Indicadores de los procedimientos alternativos propuestos para los datos de evaluación

Indicador	Procedimiento alternativo A	Procedimiento alternativo B	Procedimiento alternativo C	Procedimiento alternativo D	Procedimiento alternativo E	Procedimiento alternativo F
AUC	70.49%	70.13%	70.36%	69.88%	70.48%	70.09%
especificidad	58.59%	63.46%	58.35%	63.62%	58.24%	63.62%
sensibilidad	71.87%	66.22%	71.28%	66.10%	71.26%	65.88%

El tercer y cuarto procedimiento alternativo de clasificación (Procedimiento alternativo C y D) con el algoritmo “random forest”, se plantearon utilizando un reajuste de los parámetros con el fin de mejorar las predicciones, según los resultados anteriores se sugiere reducir la cantidad de árboles (se redució a 300 árboles) y basándonos en el error se aplica el procedimiento alternativo y se hace una comparación con datos balanceados y no balanceados.

El tercer procedimiento alternativo de clasificación (procedimiento alternativo C) con el algoritmo “random forest”, se planteó en un primer momento sin que la variable respuesta esté balanceada y utilizando como parámetro 300 arboles.

El cuarto procedimiento alternativo de clasificación (procedimiento alternativo D) con el algoritmo “random forest”, se planteó con la variable respuesta balanceada y utilizando como parámetro 300 árboles.

El quinto y sexto procedimiento alternativo de clasificación de datos (procedimientos alternativos E y F) con el algoritmo “random forest”, se plantearon utilizando un reajuste de parámetros mediante un análisis “*grid search”* para encontrar los mejores parámetros del algoritmo “random forest”. Se ha considerado el mejor de los distintos árboles, utilizando entonces 300 árboles y aplicando selección de variables se hace una comparación con datos balanceados y no balanceados. Se ha utilizado un “*grid search”* por desgaste computacional.

En un primer momento, se encuentra el procedimiento alternativo E utilizando datos sin balancear, reduciendo la cantidad de árboles a 300 y con selección de variables o atributos.

En un segundo momento, se encuentra el procedimiento alternativo F utilizando datos balanceados, reduciendo la cantidad de árboles a 300 y con selección de variables o atributos.

Como se observa en la Tabla 7, los 6 procedimientos alternativos planteados mediante el algoritmo de “random forest”, comparando en primer lugar balanceo y luego “*reajuste”* de parámetros del procedimiento alternativo se detalla lo siguiente:

Indicador AUC

Los 6 procedimientos alternativos propuestos obtuvieron un indicador AUC alrededor de 0.70, donde según las definiciones hechas por Pérez (2015) los 6 procedimientos alternativos se encuentran en una calidad de predicción Regular (Ver sección 2.2.8).

Indicador de especificidad

En base a este indicador existen diferencias en los resultados, este indicador evalúa la predicción sobre la categoría que no es de interés, en este caso el “0” (sin anemia), para los procedimientos alternativo D y F se obtuvieron un mayor indicador, esto es esperable puesto los procedimientos alternativos que se plantearon solo aprendieron de la categoría que presenta mayor frecuencia, es decir sobre la categoría de los ceros (sin anemia), encontrando los mejores parámetros del algoritmo “random forest” mediante el “*grid search”*, donde se obtuvo un mejor acierto sobre este indicador.

Indicador de sensibilidad

En base a este indicador existen diferencias en los resultados, este indicador evalúa la predicción de la categoría que es de interés, en este caso el “1” (con anemia), para los procedimientos alternativos A y C se obtuvo un mayor indicador, esto es esperable puesto que la especificidad para estos procedimientos alternativos es baja, siendo necesario el balanceo de la variable respuesta para que los procedimientos alternativos que se plantearon aprendan de ambas categorías.

Estos resultados son comparables con lo encontrado por Khan et al (2019) quienes encontraron con el mismo algoritmo “random forest” una sensibilidad del 70,73%, una especificidad del 66,41% y un AUC de 0,6857.

Este estudio tiene algunas limitaciones. Como los procedimientos alternativos predictivos utilizados en este estudio se establecieron utilizando datos de encuestas demográficas y de salud familiar transversales, no se dispuso de información adicional sobre otras variables clínicas y dietéticas potencialmente relevantes. La incorporación de esas variables probablemente habría mejorado la precisión predictiva. Dado que algunos atributos (es decir, el estado de diarrea y fiebre de los niños en las últimas dos semanas desde la fecha de la entrevista de la encuesta) fueron autoinformados, hubo posibilidades de sesgo de recuerdo. Finalmente, de los numerosos algoritmos “machine learning” que podrían haber sido aplicados en este contexto, el algoritmo “random forest” fue elegido en base al juicio subjetivo. Sin embargo, este estudio proporciona evidencia de que los algoritmos de “machine learning” se pueden utilizar para predecir la tenencia de anemia en función de los factores de riesgo comunes, lo que puede ayudar en el desarrollo de intervenciones para prevenir la anemia en los niños.

La integración de técnicas de “machine learning” para predecir la supervivencia del paciente y el estado de la enfermedad se ha vuelto cada vez más popular en la investigación de la salud pública (Alghamdi et al., 2017, Khan et al, 2019, Meena et al., 2019), lo que tiene como resultado un impacto positivo en la mejora de la planificación de la atención médica. Sin embargo, hasta la fecha, se han realizado muy pocas investigaciones sobre el uso de algoritmos de “machine learning” para predecir el estado de la enfermedad utilizando datos de encuestas de salud y demográficas transversales (Khare et al., 2017, Khan et al, 2019). Además, ninguna investigación ha explorado el potencial del “random forest” para predecir el estado de anemia de los niños de 6 a 35 meses en Perú a partir de la ENDES. Descubrimos que la anemia infantil se puede predecir con bastante precisión utilizando un conjunto de características sociodemográficas y de salud de la población que se recopilan de forma rutinaria en las encuestas de demografía y salud familiar (ENDES).

4.3. Importancia de variables

La anemia es una enfermedad potencialmente mortal que afecta la producción de hemoglobina y es especialmente (potencialmente) mortal en los niños. Por lo tanto, el uso del algoritmo “random forest” en este estudio mostró que la probabilidad de anemia infantil se puede minimizar sustancialmente interviniendo en ciertos factores

sociodemográficos y relacionados con la salud. Estos procedimientos alternativos también se pueden utilizar no solo como una guía para monitorear futuras iniciativas de control de la anemia, sino también para formular programas de nutrición infantil y políticas de salud. Además, esto puede ayudar a construir un sistema basado en el conocimiento para predecir la incidencia de anemia infantil en los niños que residen en Perú, pero no puede reemplazar la intuición y las habilidades interpretativas del médico.

Una de las virtudes del algoritmo de “random forest” es el de importancia de variables. Las variables con alta importancia tienen una fuerte asociación con los resultados de la predicción. Para medir la importancia de cada una de las variables se utiliza el indicador Gini de disminución media (“Mean Decrease Gini”) (MDG).

En las tablas 8 al 13 se observa la importancia de cada una de las 5 variables más importantes para cada uno de los procedimientos alternativos de acuerdo a su indicador de importancia relativa (“Mean Decrease Gini”).

Tabla 8

Nivel de importancia de las 5 variables del procedimiento alternativo A con mayor puntaje

Etiqueta	Variable	Importancia Relativa
X03	Edad del niño (en meses)	2018.04
X07	Altitud del conglomerado (en metros)	1162.69
X28	Visitas prenatales por embarazo	1123.07
X17	Talla de la madre (en centímetros)	1087.57
X29	Momento del primer control prenatal	1075.87

Tabla 9

Nivel de importancia de las 5 variables del procedimiento alternativo B con mayor puntaje

Etiqueta	Variable	Importancia Relativa
X03	Edad del niño (en meses)	2065.33
X07	Altitud del conglomerado (en metros)	1187.07
X28	Visitas prenatales por embarazo	1153.93
X29	Momento del primer control prenatal	1106.53
X17	Talla de la madre (en centímetros)	1072.55

Tabla 10

Nivel de importancia de las 5 variables del procedimiento alternativo C con mayor puntaje

Etiqueta	Variable	Importancia Relativa
X03	Edad del niño (en meses)	2007.57
X07	Altitud del conglomerado (en metros)	1175.24
X28	Visitas prenatales por embarazo	1128.99
X17	Talla de la madre (en centímetros)	1086.82
X29	Momento del primer control prenatal	1078.04

Tabla 11

Nivel de importancia de las 5 variables del procedimiento alternativo D con mayor puntaje

Etiqueta	Variable	Importancia Relativa
X03	Edad del niño (en meses)	2064.54
X07	Altitud del conglomerado (en metros)	1182.95
X28	Visitas prenatales por embarazo	1149.08
X29	Momento del primer control prenatal	1106.09
X17	Talla de la madre (en centímetros)	1082.85

Tabla 12

Nivel de importancia de las 5 variables del procedimiento alternativo E con mayor puntaje

Etiqueta	Variable	Importancia Relativa
X03	Edad del niño (en meses)	2010.72
X07	Altitud del conglomerado (en metros)	1168.38
X28	Visitas prenatales por embarazo	1131.23
X17	Talla de la madre (en centímetros)	1082.82
X29	Momento del primer control prenatal	1081.72

Tabla 13

Nivel de importancia de las 5 variables del procedimiento alternativo F con mayor puntaje

Etiqueta	Variable	Importancia Relativa
X03	Edad del niño (en meses)	2062.11
X07	Altitud del conglomerado (en metros)	1184.67
X28	Visitas prenatales por embarazo	1158.65
X29	Momento del primer control prenatal	1112.65
X17	Talla de la madre (en centímetros)	1078.62

Las 5 variables más importantes para los procedimientos alternativos y que son necesario tener en cuenta para el lanzamiento de futuras políticas públicas para la disminución del porcentaje de niños con anemia son: edad del niño (en meses),

altitud del conglomerado (en metros), visitas prenatales por embarazo, momento del primer control prenatal y talla de la madre (en centímetros).

En este estudio, se exploró el conjunto de datos de la ENDES 2015-2019 para proporcionar una idea inicial de la aplicabilidad potencial del algoritmo “random forest” para predecir el estado de anemia de los niños de seis a 35 meses de edad en función de las características sociodemográficas y de salud. Los hallazgos sugirieron claramente que la mayoría de los atributos relacionados con la salud materna e infantil, como la edad del niño, visitas prenatales por embarazo, momento del primer control prenatal, talla de la madre (en centímetros), tienen una fuerte relación con el estado de anemia de los niños. Además, las características del hogar, como la altitud del conglomerado (en metros), muestra una clara asociación con la incidencia de anemia. Durante la experimentación se construyeron varios procedimientos alternativos que podían predecir el riesgo de anemia infantil basándose en estas variables significativamente relacionadas.

CAPÍTULO V

CONCLUSIONES Y RECOMENDACIONES

5.1. Conclusiones

En la investigación se desarrollaron 6 procedimientos alternativos. En el procedimiento A se usaron datos sobre la tenencia de anemia sin balancear. En el procedimiento B se consideraron datos balanceados. En los procedimientos C y D se reajustó los parámetros encontrando los mejores valores mediante un "*grid search"* reduciendo el número de árboles de decisión (se redujo a 300 árboles) y con todas las variables, haciendo una comparación con datos balanceados y no balanceados. En los procedimientos E y F se encuentra los mejores parámetros mediante un "*grid search"*, se reduce la cantidad de árboles a 300 y se aplica selección de variables no balanceando y balanceando los datos.

1) De todos los procedimientos alternativos, mejores resultados se obtuvieron con los procedimientos B (todas las variables) y F (reducción de la cantidad de árboles a 300 y con selección de variables), ambos balanceados y tomando en consideración el procedimiento F (que tiene un área bajo la curva de 70,1% y tanto su especificidad como su sensibilidad tienen porcentajes más similares y donde se utiliza las variables de mayor importancia).

2) A partir del análisis descriptivo, mayores porcentajes de niños con anemia se encuentran en casos donde los niños tienen menor edad, residen en conglomerados con mayor altitud, pertenecen a hogares con mayor cantidad de niños menores de 5 años de edad, con mayor número de miembros en el hogar y provienen de hogares donde la lengua materna es nativa u originaria.

3) En el caso del indicador AUC (área bajo la curva) todos los procedimientos superaron el valor mínimo (0,60), por lo que las predicciones de la anemia son de calidad "regular."

   Respecto a la especificidad, que mide la predicción de la clase que no es de interés (sin anemia), los mayores valores se obtuvieron con el procedimiento D (63,6%) y procedimiento F (63,6%). Estos resultaron ser los mejores procedimientos cuando se mejoraron los parámetros mediante el "*grid search"*

con balanceo.

En el caso de sensibilidad, que mide la predicción de la clase que es de interés (con anemia), los mayores valores se obtuvieron con el procedimiento A (71,9 %) y el procedimiento C (71,3%).

4) Las 5 variables más importantes para el mejor procedimiento (procedimiento F) y que es necesario tener en cuenta para el lanzamiento de futuras políticas públicas para la disminución del porcentaje de niños con anemia son: variables relacionadas con el niño (edad del niño, en meses), variables sociodemográficas (altitud del conglomerado, en metros), variables del cuidado materno e infantil (número de visitas prenatales por embarazo, meses de embarazo del primer control prenatal y talla de la madre en centímetros).

5.2. Recomendaciones

1) Entrenar los procedimientos alternativos respecto a un indicador que mida las probabilidades como lo es el AUC en el "*grid search"*, nos ayudaría a encontrar un mejor performance para este indicador.
2) Un alto valor de la especificidad ayudará a conocer a los niños que no presentan anemia, por lo que se recomienda hacer un estudio más profundo para estos niños.
3) Un alto valor de la sensibilidad ayudará hacer la gestión más eficiente y priorizar a los niños más propensos.
4) Inflar la curva ROC agregando nuevas variables y así obtener mejores resultados.
5) Optar por otros algoritmos como redes neuronales, máquina de soporte vectorial y modelos bayesianos (Naives Bayes) para futuras investigaciones con estas bases de datos ayudará a encontrar posiblemente mejores indicadores.

REFERENCIAS BIBLIOGRÁFICAS

Alghamdi M, Al-Mallah M, Keteyian S, Brawner C, Ehrman J. & Sakr S. (2017) *Predicting diabetes mellitus using SMOTE and ensemble machine learning approach:* The Henry Ford Exercise Testing (FIT) project. PloS one;12(7): e0179805.

Aprende Machine Learning (2019). *Aprende Machine Learning*. https://www.aprendemachinelearning.com/clasificacion-con-datos-desbalanceados/

Astete, L., Velásquez, J. y Loyola, J. (2014). *Factores asociados con la anemia en niños menores de tres años Perú. 2007-2013*. https://www.researchgate.net/publication/264436051

Ayyildiz, H. & Tuncer, S.A. (2020). *Determination of the effect of red blood cell parameters in the discrimination of iron deficiency anemia and beta thalassemia via Neighborhood Component Analysis Feature Selection-Based machine learning.* Chemometrics and Intelligent Laboratory Systems. Volume 196, 15 January 2020, 103886. DOI: https://doi.org/10.1016/j.chemolab.2019.103886

Balarajan, Y., Ramakrishnan, U., Ozaltin, E., Shankar, A. & Subramanian, S. (2011). *Anaemia in low-income and middle-income countries.* The Lancet; 378(9809):2123–35. DOI: https://doi.org/10.1016/S0140-6736(10)62304-5

Batista, G. (2002). *A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data.* International Conference on Machine Learning, (págs. 139-146).

Bernal, C. (2016). *Metodología de la investigación. Administración, economía, humanidades y ciencias sociales*. Editorial Pearson. Cuarta edición. Colombia.

Bjoern, M, A. (2009). C*omparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data.* BMC Bioinformatics. [PMC free article] [PubMed]

Black, R., Victora, C., Walker, S., Bhutta, Z., Christian, P., De Onis, M., Ezzati, M., Grantham-McGregor, S., Katz, J., Martorell, R. & Uauy, R. (2013). *Maternal and child undernutrition and overweight in low-income and middle-income countries*. Lancet, 2013. 382(9890): p. 427-51.

Breiman, L. (1984). *Classification and regression trees.* Vol. 358. Wadsworth. Inc., Belmont, CA. [Google Scholar]

Breiman, L. (2001). *Random Forest*. Machine Learning. California: Statistics Department. 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L. (2003). *Manual-Setting Up, Using, and Understanding Random Forests V4. 0.* ftp://ftpstat.berkeley.edu/pub/users/breiman

Brownlee, J. (2015). *Tactics to Combat Imbalanced Classes in Your Machine Learning Dataset*. Obtenido de http://machinelearningmastery.com/tactics-to-combat- imbalanced-classes-in-your-machine-learning-dataset/

Buja, A., Stuetzle, W. & Shen, Y. (2005). *Loss Functions for Binary Class Probability Estimation and Classification: Structure and Applications*. Pennsylvania.

Cárdenas, J. (2019). *Clasificación de aceptación de campañas para una entidad financiera, usando random forest con datos balanceados y datos no balanceados*. Tesis URP.

Cutler, A. (2010). *Remembering Leo Breiman. The Annals of Applied Statistics*, 4(4), 1621–1633.

Defensoría del Pueblo (2018). *Intervención del Estado para la reducción de la anemia infantil: Resultados de la supervisión nacional.* Serie Informes de Adjuntía – Informe de Adjuntía 012-2018-DP/AAE. https://www.defensoria.gob.pe/wp-content/uploads/2018/12/Informe-de-Adjunt%C3%ADa-012-2018-DP-AAE-Intervenci%C3%B3n-del-Estado-para-la-reducci%C3%B3n-de-la-anemia-infantil.pdf

Denic, S., Agarwal, M.M. (2007). *Nutritional Iron deficiency: an evolutionary perspective.* Nutrition 23(7-8):603-614. https://www.researchgate.net/publication/6254786_Nutritional_iron_deficiency_An_evolutionary_perspective

Dirren, H., Logman, M., Barclay, D. & Freire, W. (1994). Altitude correction for hemoglobin. Eur J Clin Nutr. 1994;48:625-32.

Durán, B. (2019). Comparación de metodologías de imputación aplicadas a ingresos *laborales de la ENOE*. INEGI Vol.10, Núm.3. Recuperado el 13 de Julio de 2021, de Comparación de metodologías de imputación aplicadas a ingresos laborales de la ENOE - REALIDAD, DATOS Y ESPACIO REVISTA INTERNACIONAL DE ESTADÍSTICA Y GEOGRAFÍA (inegi.org.mx)

Fawcett, T. (2016). Datos Desbalanceados. Recuperado el 25 de Noviembre de 2017, de https://svds.com/learning-imbalanced-classes/

Fernández, R. (2016). *Regresión bayesiana con enlaces asimétricos para la clasificación de clientes con propensión a caer en mora en una entidad bancaria.* Lima: Escuela de Posgrado UNAM.

Fundo de las Naçiones Unidas para la Infancia (UNICEF) (1998). El estado mundial de la infancia 1998: un informe de UNICEF. Desnutrición: causas, consecuencias y soluciones. Nutr Rev. 1998;56:115-23.

Gamble, M.V., Palafox, N.A., Dancheck, B., Ricks, M.O., Briand, K. & Semba, R.D. (2004). *Relationship of vitamin A deficiency, iron deficiency, and inflammation to anemia among preschool children in the Republic of the Marshall Islands.* Eur J Clin Nutr 58(10) :1396-401.

Genuer, R. & Poggi, J.M. (2020). *Random Forest with R*.

Géron, A. (2019). *Aprende Machine Learning con Scikit-Learn, Keras y TensorFlow. Conceptos, herramientas y técnicas para conseguir sistemas inteligentes*. Segunda edición. Ed. Anaya Multimedia. Madrid – España.

Çil, B., Ayyıldız, H. & Tuncer, T. (2020). *Discrimination of β-thalassemia and iron deficiency anemia through extreme learning machine and regularized extreme learning machine based decision support system.* Medical Hypotheses 138 (2020) 109611. https://doi.org/10.1016/j.mehy.2020.109611

Hastie, T. (2009). *The elements of statistical learning: data mining, inference and prediction.* pp. 605–622.

Hernández, R. y Mendoza (2018). *Metodología de la investigación. Las rutas cuantitativa, cualitativa y mixta*. Mc Graw Hill. México.

Instituto Nacional de Estadística e Informática (INEI) (2016). *Encuesta Demográfica y de Salud Familiar - ENDES 2015.* Lima, Perú.

Instituto Nacional de Estadística e Informática (INEI) (2017). *Encuesta Demográfica y de Salud Familiar - ENDES 2016.* Lima, Perú.

Instituto Nacional de Estadística e Informática (INEI) (2018). *Encuesta Demográfica y de Salud Familiar - ENDES 2017.* Lima, Perú.

Instituto Nacional de Estadística e Informática (INEI) (2019). *Encuesta Demográfica y de Salud Familiar - ENDES 2018.* Lima, Perú.

Instituto Nacional de Estadística e Informática (INEI) (2020). *Encuesta Demográfica y de Salud Familiar - ENDES 2019.* Lima, Perú.

Instituto Nacional de Estadística e Informática (INEI) (2021). *Encuesta Demográfica y de Salud Familiar – ENDES 2020.* Lima, Perú. https://proyectos.inei.gob.pe/endes/documentos.asp.

Instituto Nacional de Salud [INS] (2013). *Procedimiento para la determinación de la hemoglobina mediante hemoglobinómetro portátil.* Lima-Perú. Guía Técnica: procedimiento para la determinación de hemoglobina mediante hemoglobinómetro portátil (ins.gob.pe).

Jaiswal, M., Srivastava, A. & Siddiqui, T.J. (2019). *Machine Learning Algorithms for Anemia Disease Prediction: Select Proceedings of IC3E 2018.* ResearchGate. DOI:10.1007/978-981-13-2685-1_44 https://www.researchgate.net/publication/329484705

Khalilia, M., Chakraborty, S. & Popescu, M. (2011). *Predicting disease risks from highly imbalanced data using random forest.* BMC Medical Informatics and Decision Making 2011, 11:51. doi: 10.1186/1472-6947-11-51. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3163175/

Khan, J.R., Chowdhury, S., Islam, H. & Raheem, E. (2019). *Machine learning algorithms to predict the childhood anemia in bangladesh.* Journal of Data Science,17(1). P. 195 - 218. DOI:10.6339/JDS.201901_17(1).0009. https://www.researchgate.net/publication/328368657_Machine_Learning_Algorithms_To_Predict_The_Childhood_Anemia_In_Bangladesh#:~:text=We%20considered%20machine%20learning%20(ML,common%20risk%20factors%20as%20features.&text=We%20found%20that%20the%20RF,66.41%25%20and%20AUC%20of%200.6857.

Khare S., Kavyashree S., Gupta D. & Jyotishi A. (2017). *Investigation of Nutritional Status of Children based on Machine Learning Techniques using Indian Demographic and Health Survey Data.* Procedia Computer Science.115:338-49.

Kroese, D.P., Botev, Z.I., Taimre, T. & Vaisman, R. (2020). *Data Science and Machine Learning. Mathematical and Statistical Methods.* Chapman & Hall / CRC Press. Taylor & Francis Group. Machine Learning & Pattern Recognition.

Kubat, M.M. (1997). *Addressing the Course of Imbalanced Training Sets: One-sided Selection.* ICML, 179- 186.

Landry, M. (2018). *Machine Learning with R and H2O*. United States of America: H2O.

Lozoff, B., Beard, J., Connor, J., Felt, B., Georgieff, M. & Schallert, T. (2006). *Long-lasting neural and behavioral effects of iron deficiency in infancy*. Nutrition Reviews. 2006;64:S34–S43.

Lutter, C. (2008). *Iron deficiency in young children in low-income countries and new approaches for its prevention.* J Nutr. 2008;138:2523-8. http://dx.doi.org/10.3945/jn.108.095406

Luo, Y., Szolovits, P., Dighe, A.S. & Baron, J.M. (2016). *Using Machine Learning to Predict Laboratory Test Results.* © American Society for Clinical Pathology 2016; 145:778-788. DOI: 10.1093/AJCP/AQW064

Magalhães, R.J.S. & Clements, A.C.A. (2011). *Mapping the risk of anemia in preschool-age children: the contribution of malnutrition, malaria, and helminth*

infections in West Africa. PLoS Medicine, 8(6): e1000438 doi:10.1371/journal.pmed.1000438

Mahboob, T., Irfan, S., & Karamat, A. (2017). A machine learning approach for student assessment in E-learning using Quinlan's C4.5, Naive Bayes and Random Forest algorithms. Proceedings of the 2016 19th International MultiTopic Conference, INMIC 2016. https://doi.org/10.1109/INMIC.2016.7840094

Mann, N. (2007). *Meat in the human diet: an anthropological perspective.* Nutr Diet 64 (Suppl. 4): S102–S107.

Meena, K., Tayal, D. K., Gupta, V. & Fatima, A. (2019). *Using Classification Techniques for Statistical Analysis of Anemia, Artificial Intelligence In Medicine*, https://doi.org/10.1016/j.artmed.2019.02.005

Mingers, J. (1989). *An empirical comparison of selection measures for decision-tree induction. Machine learning.* 1989;3(4):319–342. [Google Scholar]

Ministerio de Desarrollo e Inclusión Social (2012). *Una política para el desarrollo y la inclusión social en el Perú.* Lima: MIDIS.

Ministerio de Salud de Perú. (2017). *Plan Nacional para la Reducción y Control de la Anemia Materno Infantil y Desnutrición Crónica Infantil 2017-2021.* Resolución Ministerial N° 249-2017/MINSA. Lima: MINSA; 2017.

Ministerio de Desarrollo e Inclusión Social (2018). *Plan multisectorial de lucha contra la anemia.* Lima: MIDIS. plan-multisectorial-de-lucha-contra-la-anemia-v3.pdf (www.gob.pe)

Ministerio de Salud (MINSA) (2017). *Plan Nacional para la Reducción y Control de la Anemia Materno Infantil y la Desnutrición Crónica Infantil y la desnutrición crónica infantil en el Perú: 2017-2021*. Lima, Perú. http://bvs.minsa.gob.pe/local/MINSA/4189.pdf

Mishra, T., Kumar, D., & Gupta, S. (2014). Mining students' data for prediction performance. International Conference on Advanced Computing and Communication Technologies, ACCT, 255–262. https://doi.org/10.1109/ACCT.2014.105

Moreno, Q. (2001). *Aplicación de técnicas de minería de datos en la Construcción y Validación de modelos predictivos y Asociativos a partir de especificaciones de requisitos de software.* Real: Universidad de Salamanca.

Moreno, M., Quintales, L., García, F. y Martín, M. (2002). *Obtención y Validación de Modelos de Estimación de Software Mediante Técnicas de Minería de Datos.* Revista Colombiana de Computación – RCC, Vol 3, N° 1.

Moschovis, P., Wiens, M., Arlington, L., Antsygina, O., Hayden, D. & Dzik, W. (2018). *Individual, maternal and household risk factors for anaemia among young children in sub-Saharan Africa: a crosssectional study.* BMJ Open. https://bmjopen.bmj.com/content/bmjopen/8/5/e019654.full.pdf

Nestel, P., Briend, A., De Benoist, B., Decker, E., Ferguson, E., Fontaine, O., Micardi, A. & Nalubola, R. (2003). *Complementary food supplements to achieve micronutrient adequacy for infants and young children.* Journal of Pediatric Gastroenterology and Nutrition, 2003, 36:316–328.

Nestel, D., Muir, E., Plant, M., Kidd, J. & Thurlow, S. (2002). *Modelling the lay expert for first-year medical students: the actor-patient as teacher.* Med Teach 24:562–564.

Ngnie-Teta, I., Receveur, O. & Kuate-Defo, B. (2007). *Risk factors for moderate to severe anemia among children in Benin and Mali: Insights from a multilevel*

analysis. Food and Nutrition Bulletin, vol. 28, no. 1. The United Nations University. https://doi.org/10.1177/156482650702800109

Obasohan, P.E., Walters, S.J., Jacques, R. & Khatab, K. (2022). *Individual, household and area predictors of anaemia among children aged 6–59 months in Nigeria*. Public Health in Practice, Vol. 3. https://doi.org/10.1016/j.puhip.2022.100229

Ogunsakin, R.E., Babalola, B.T. & Akinyemi, O. (2020). *Statistical Modeling of Determinants of Anemia Prevalence among Children Aged 6–59 Months in Nigeria: A Cross-Sectional Study.* Hindawi Anemia Volume 2020, Article ID 4891965, https://doi.org/10.1155/2020/4891965 https://www.hindawi.com/journals/anemia/2020/4891965/

Organización Mundial de la Salud (OMS) (2008). Worldwide prevalence of anaemia 1993-2005. Ginebra. https://apps.who.int/iris/bitstream/handle/10665/43894/9789241596657_eng.pdf?sequence=1

Organización Mundial de la Salud (OMS) (2015). Administración de suplementos de hierro en niños de 6 a 23 meses de edad. Ginebra. https://www.who.int/elena/titles/iron_supplementation_children/es/

Organización Mundial de la Salud (OMS) (2017). Metas mundiales de nutrición 2025. Documento normativo sobre anemia. Ginebra: https://apps.who.int/iris/bitstream/handle/10665/255734/WHO_NMH_NHD_14.4_spa.pdf?ua=1

Organización Mundial de la Salud (OMS) (2020). Carencia de micronutrientes. Ginebra. https://apps.who.int/nutrition/topics/ida/es/index.html

Oppenheimer, S.J. (2001). *Iron and its relation to immunity and infectious disease.* J Nutr 131: 616S–635S.

Ortiz, K.J., Ortiz, Y.J., Escobedo, J.R., Neyra, L. y Jaimes, C.A. (2021). *Análisis del modelo multicausal sobre el nivel de la anemia en niños de 6 a 35 meses en Perú.* Enfermería global. Revista electrónica trimestral de Enfermería. vol.20 no.64. https://dx.doi.org/10.6018/eglobal.472871 https://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S1695-61412021000400426#B8

Pal, M. (2005). Random forest classifier for remote sensing classification. International Journal of Remote Sensing, 26(1), 217–222. https://doi.org/10.1080/01431160412331269698

Pardo, C. (2015). *Minería de datos y combinación de regresores*. España.

Perez, C. y Santín, D. (2007). *Minería de Datos: Técnicas y Herramientas*. Madrid: Ediciones Paraninfo, S.A.

Pérez, J. (2015). *Bayesian Asymmetric Logit Model for Detecting Risk Factors in Motors Ratemaking.* Principado de Asturias: Facultad de Ciencias, Universidad de Oviedo.

Prentice, A.M. (2008). *Iron metabolism, malaria, and other infections: what is all the fuss about?* J Nutr 138 (12): 2537-2541.

Puente-Maury, L., López-Chau, A. y Cruz-Santos, W. (2014). *Método rápido de preprocesamiento para clasificación en conjuntos de datos no balanceados.* México.

Raschka, S. & Mirjalili, V. (2019). *Python Machine Learning. Editorial Marcombo.* Segunda edición. España.

Rimachi, N. y Longa, J. (2013). *Factores de riesgo asociados a anemia a menores de 5 años usuarios del consultorio de crecimiento y desarrollo – Centro de Salud Mi Perú – Ventanilla, 2013.* UAP. Lima- Perú.

Rock, E., Gueux, E., Mazur, A., Motta, C. & Rayssiguier, Y. (1995). *Anemia in copper-deficient rats: role of alterations in erythrocyte membrane fluidity and oxidative damage.* Am J Physiol 269(5 Pt 1):C1245-9.

Rodríguez, O. (2010). *Un Aprendizaje Supervisado: Árboles de Decisión.*

Samuel, D. y Pacheco, L. (2005). *El clasificador de Naive Bayes en la extracción de conocimiento de bases de datos.*

Sanou, D. & Ngnie-Teta, I. (2012). *Risk Factors for Anemia in Preschool Children in Sub-Saharan Africa.* Researchgate. https://www.researchgate.net/profile/Dia_Sanou/publication/221926483_Risk_Factors_for_Anemia_in_Preschool_Children_in_Sub-Saharan_Africa/links/547c4c8a0cf293e2da2da619/Risk-Factors-for-Anemia-in-Preschool-Children-in-Sub-Saharan-Africa.pdf

Sanchis-Gomar, F., Cortell-Ballester, J., Pareja-Galeano, H., Banfi, G. & Lippi, G. (2013) *Hemoglobin point-of-care testing: The HemoCue system.* J Lab Autom.;18:198-205. http://dx.doi.org/10.1177/2211068212457560

Sazawal, S., Black, R. & Ramsan, M. (2006). *Effects of routine prophylactic supplementation with iron and folic acid on admission to hospital and mortality in preschool children in a high malaria transmission setting: community-based, randomised, placebo-controlled trial.* Lancet 367, 133–143.

Semba R.D. & Bloem M.W., (2002). *The anemia of vitamin A deficiency: epidemiology and pathogenesis.* Eur J Clin Nutr 56:271-281.

Shenton L.M., Jones A.D., Wilson M.L. (2020). *Factors Associated with Anemia Status Among Children Aged 6-59 months in Ghana, 2003-2014*. Maternal and Child Health Journal. 2020;24:483-502. DOI: https://doi.org/10.1007/s10995-019-02865-7

Statology (2020). *An Introduction to Bagging in Machine Learning.* An Introduction to Bagging in Machine Learning - Statology

Sucari, R. (2018). *Comparación del análisis discriminante no métrico, árboles de clasificación CHAID y la regresión logística multinomial*. [Tesis de Maestría]. UNALM. Lima – Perú.

Talukder & Ahammed (2020) *Machine Learning Algorithms for Predicting Malnutrition among Under-Five Children in Bangladesh*. Revista Nutrition. Editorial: Elsevier.

Thangamani, D. & Sudha, P. (2014*). Identification Of Malnutrition With Use Of Supervised Datamining Techniques –Decision Trees And Artificial Neural Networks*. International Journal Of Engineering And Computer Science ISSN:2319-7242. Volume - 3 Issue -9 September, 2014 Page No. 8236-8241. https://www.researchgate.net/publication/329012021_Identification_Of_Malnutrition_With_Use_Of_Supervised_Datamining_Techniques-Decision_Trees_And_Artificial_Neural_Networks

Thorandeniya, T., Wickremasinghe, R., Ramanyake, R. & Atukorala, S. (2006). *Low folic acid status and its association with anemia in urban adolescent girls and women of childbearing age in Sri Lanka.* Brit J Nutr. 95 (3): 511-516.

Tomek, I. (1976). *Two modifications of CNN. IEEE Transactions on Systems, Man and Cybernetics.* Volume SMC-6, Issue 11, 769-772.

Vanzetti, G. (1966). *An azide-methemoglobin method for hemoglobin determination in blood.* J Lab Clin Med. 1966;67:116-26.

Vargas, M. (s.f.). *Clasificador Bayesiano usando Distribución Normal Multivariado para predecir el riesgo académico de pregrado en la Universidad de Colombia*.

Velásquez, J.E., Rodríguez, Y., Gonzales, M., Astete-Robilliard, L., Loyola-Romaní, J., Vigo, W.E. y Rosas-Aguirre, A.M. (2016). *Factores asociados con la anemia en niños menores de tres años en Perú: análisis de la Encuesta Demográfica y de Salud Familiar, 2007-2013.* Revista del Instituto Nacional de Salud: Biomédica. Vol. 36 Núm. 2. Colombia. https://revistabiomedica.org/index.php/biomedica/article/view/2896

Véliz, C. (2018). *Aprendizaje automático. Análisis para la minería de datos y big data.* PUCP. Lima-Perú.

Wander, K., Shell-Duncan, B. & McDade, T. (2009). *Evaluation of iron deficiency as a nutritional adaptation to infectious disease: An evolutionary medicine perspective*. *American Journal of Human Biology*, 2009; 21(2):172-179.

Wilmott, P. (2019). *Machine Learning: An applied mathematics introduction.* First Edition. Panda Ghana Publishing.

World Health Organization. (2004). *Comparative quantification of health risks. Geneva:* WHO; 2004.

World Health Organization. (2001). *Iron deficiency anaemia: Assessment, prevention and control. A guide for programme managers*. http://www.who.int/nutrition/publications/en/ida_assessment_prevention_control.pdf.

WHO/UNICEF. (2006). *Iron supplementation of young children in regions where malaria transmission is intense and infectious disease highly prevalent.* Joint statement by the World Health Organization and the United Nations Children's Fund. Geneva. 2 p.

WHO/UNICEF. (2006). *Iron supplementation of young children in regions where malaria transmission is intense and infectious disease highly prevalent.* WHO/UNICEF Joint statement. Geneva.

Yang, K. (2019). *Introduction to Algorithms for Data Mining and Machine Learning*. Middlesex University School of Science and Technology London, United Kingdom.

Zavaleta, N. y Irizarry, L. (2016). *Nutrición en el Perú 2016. Situación nutricional y sus Implicancias de Política Pública. Nota Técnica. Banco Interamericano de Desarrollo – BID*. División de Protección Social y Salud.

ANEXOS

ANEXO A

Tabla A. *Estadísticas de resumen de los datos obtenidos de la ENDES del periodo 2015-2019 (INEI, 2015-2019)*

Etiqueta	Variable	Tipo de variable	Media	Mediana	Moda	Desv. Est.*	Mínimo	Máximo	N° de categorias
X1	Nivel educativo más alto de la madre	Cualitativa	-	2	2	-	1	3	3
X2	Sexo del niño	Cualitativa	-	-	2	-	1	2	2
X3	Edad del niño (en meses)	Cuantitativa	20.76	21	33	8.61	6	35	-
X4	Orden de nacimiento del niño	Cuantitativa	2.41	2	1	1.56	1	15	-
X5	Intervalo entre nacimientos anteriores al niño	Cuantitativa	45.57	35	0	47.33	0	388	-
X6	Lugar de residencia del niño	Cualitativa	-	-	1	-	1	2	2
X7	Altitud del conglomerado (en metros)	Cuantitativa	1265.31	430	12	1413.58	2	5037	-
X8	Número de niños menores de 5 años en el hogar	Cuantitativa	1.38	1	1	0.63	1	10	-
X9	Número de miembros del hogar	Cuantitativa	5.17	5	4	2.04	2	21	-
X10	Región natural	Cualitativa	-	-	4	-	1	4	4
X11	Índice de riqueza	Cualitativa	-	4	4	-	1	5	5
X12	Fuente principal de abastecimiento de agua potable	Cualitativa	-	-	1	-	1	3	3
X13	Tipo de instalación sanitaria	Cualitativa	-	-	1	-	1	4	4
X14	Material predominante del piso de la vivienda	Cualitativa	-	-	2	-	1	4	4
X15	El agua usualmente es tratada por: hervida	Cualitativa	-	-	1	-	1	2	2
X16	Edad de la madre (en años)	Cuantitativa	30.42	30	30	8.17	12	49	-
X17	Talla en centímetros (1 decimal)	Cuantitativa	152.07	151.9	150	56.59	39.5	176.7	-
X18	Nivel de anemia en la madre	Cualitativa	-	-	1	-	1	2	2
X19	Etnicidad de la madre	Cualitativa	-	-	1	-	1	4	4
X20	Persona que normalmente alimenta al niño	Cualitativa	-	-	3	-	1	3	3
X21	Niño tomó hierro en jarabe, polvo, gotas u otra presentación	Cualitativa	-	-	1	-	1	2	2
X22	Le hicieron algún control de crecimiento y desarrollo	Cualitativa	-	-	2	-	1	2	2
X23	Ha tenido fiebre en las últimas dos semanas	Cualitativa	-	-	1	-	1	2	2
X24	En los últimos 14 dìas, ha tenido diarrea la niña(o)	Cualitativa	-	-	1	-	1	2	2
X25	Ha tenido tos en las últimas dos semanas	Cualitativa	-	-	1	-	1	2	2
X26	Alguna vez recibió la dosis de vitamina A	Cualitativa	-	-	1	-	1	2	2
X27	Medicamentos para parásitos intestinales en los últimos 6 meses	Cualitativa	-	-	2	-	1	2	2
X28	Visitas prenatales por embarazo	Cuantitativa	8.98	9	10	3.10	0	20	-
X29	Momento del primer control prenatal	Cuantitativa	2.61	2	2	1.56	0	9	-
X30	Peso del niño al nacer (kilos - 3 dec.)	Cuantitativa	3246.44	3250	3500	544.09	9	6000	-
X31	Parto institucional	Cualitativa	-	-	1	-	1	2	2
X32	Por cuantos días tomó hierro y/o cuantas inyecciones recibió	Cuantitativa	109.42	90	90	71.47	0	360	-
X33	Comidas sólidas o semisólidas	Cuantitativa	4.14	4	5	1.23	0	7	-
Anemia	Prevalencia de Anemia en el niño	Cualitativa	-	-	1	-	1	2	2

* Desv.Est: desviación estándar.

Anexo B

Recategorizaciones de las variables mediante arboles de decisión CHAID

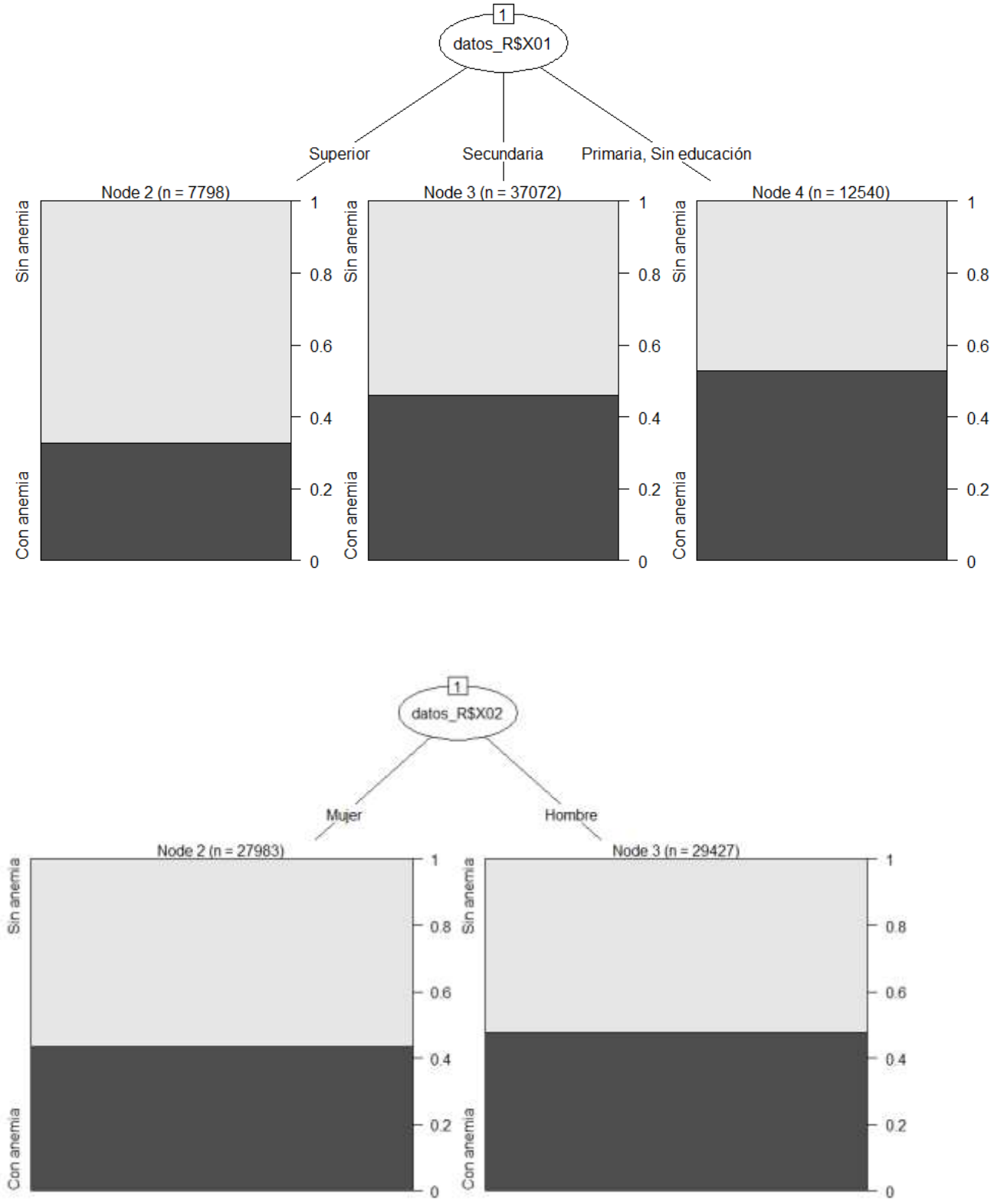


Figura B1. *Árbol de decisión CHAID correspondiente a la variable X01 (nivel de educación, arriba) y X02 (sexo, abajo)*

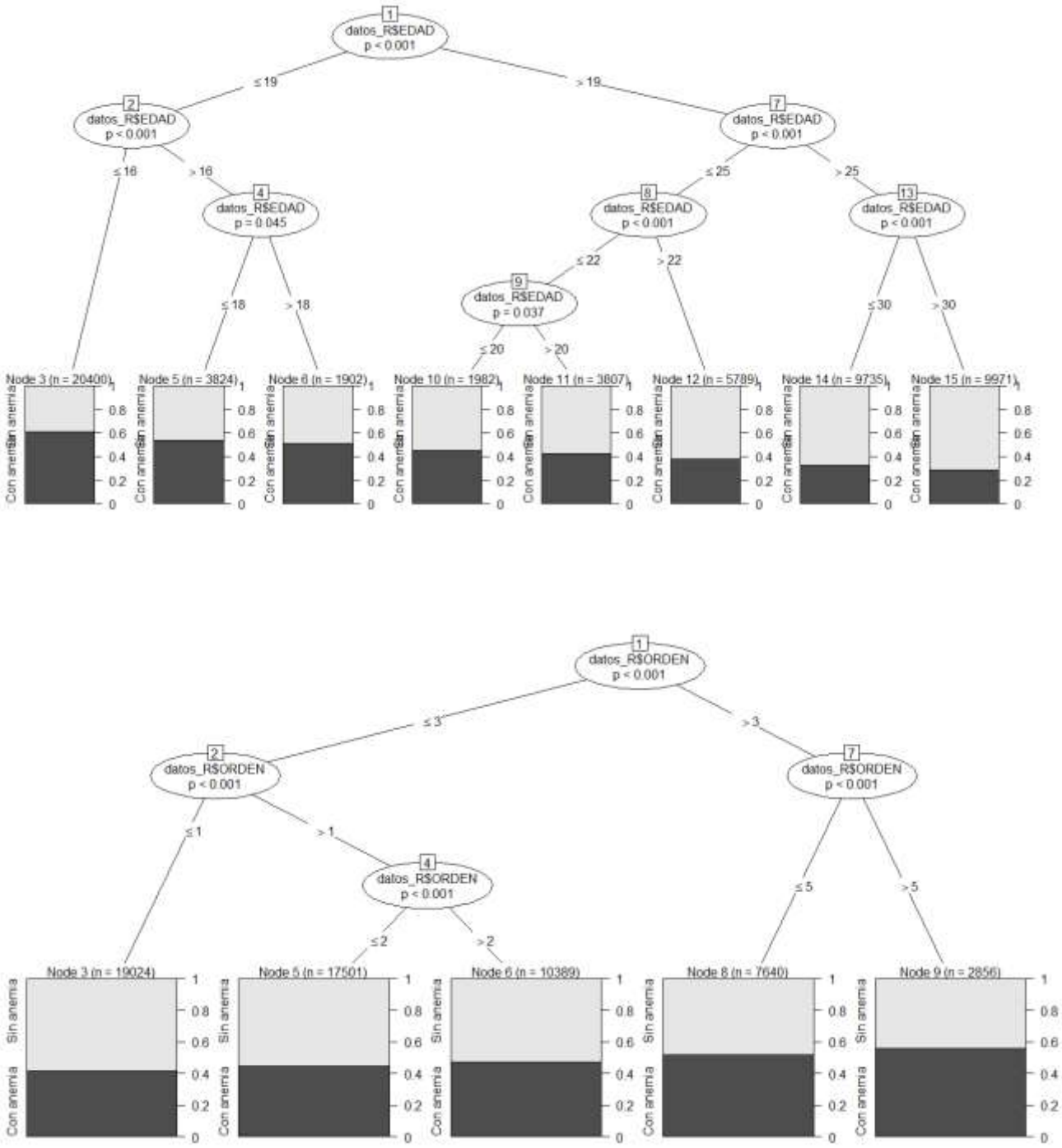


Figura B2. *Árbol de decisión CHAID correspondiente a la variable EDAD (edad del niño, arriba) y ORDEN (orden de nacimiento, abajo)*

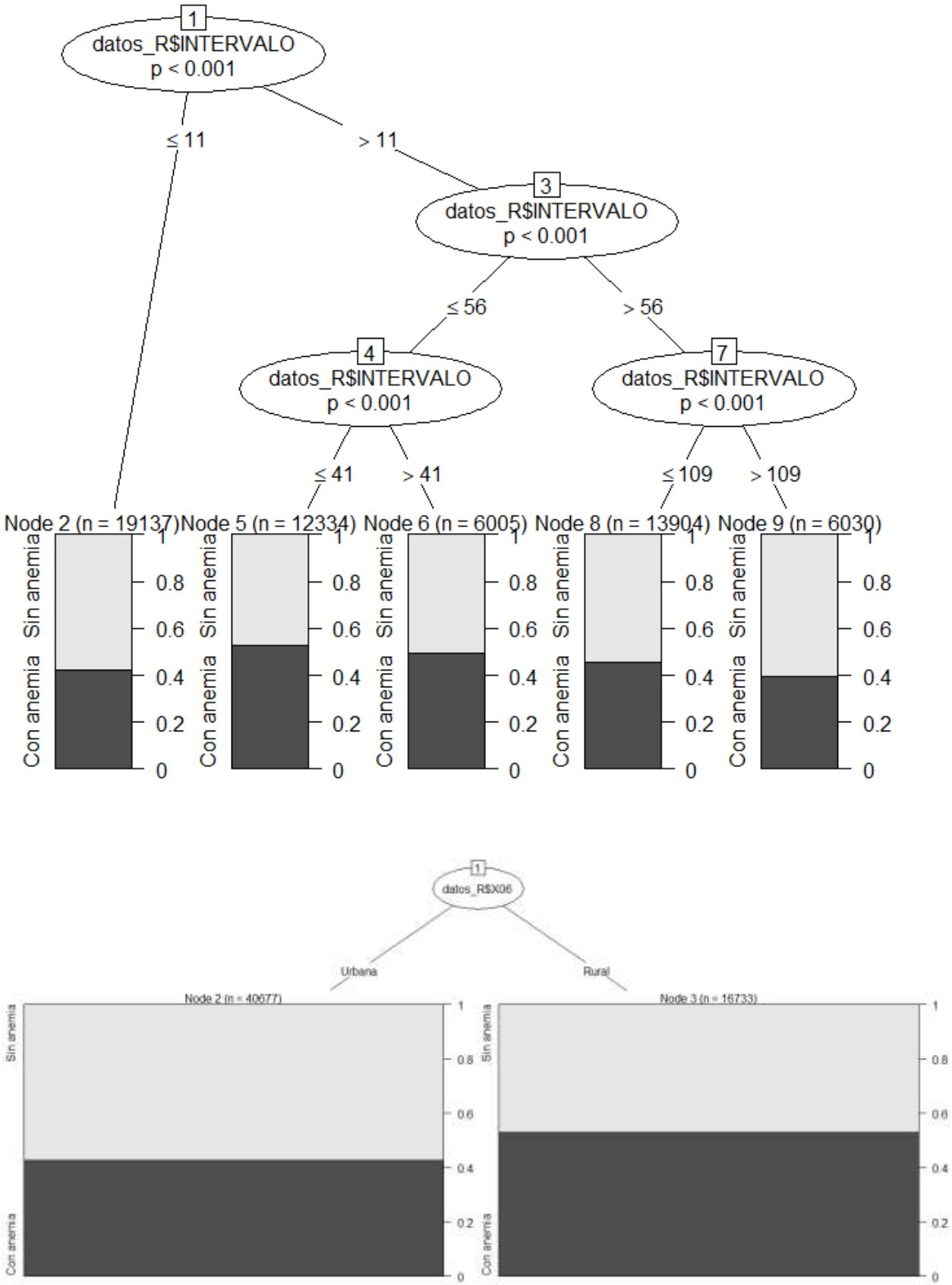


Figura B3. *Árbol de decisión CHAID correspondiente a la variable INTERVALO (intervalo entre nacimientos anteriores al niño, arriba) y X06 (lugar de residencia del niño, abajo)*

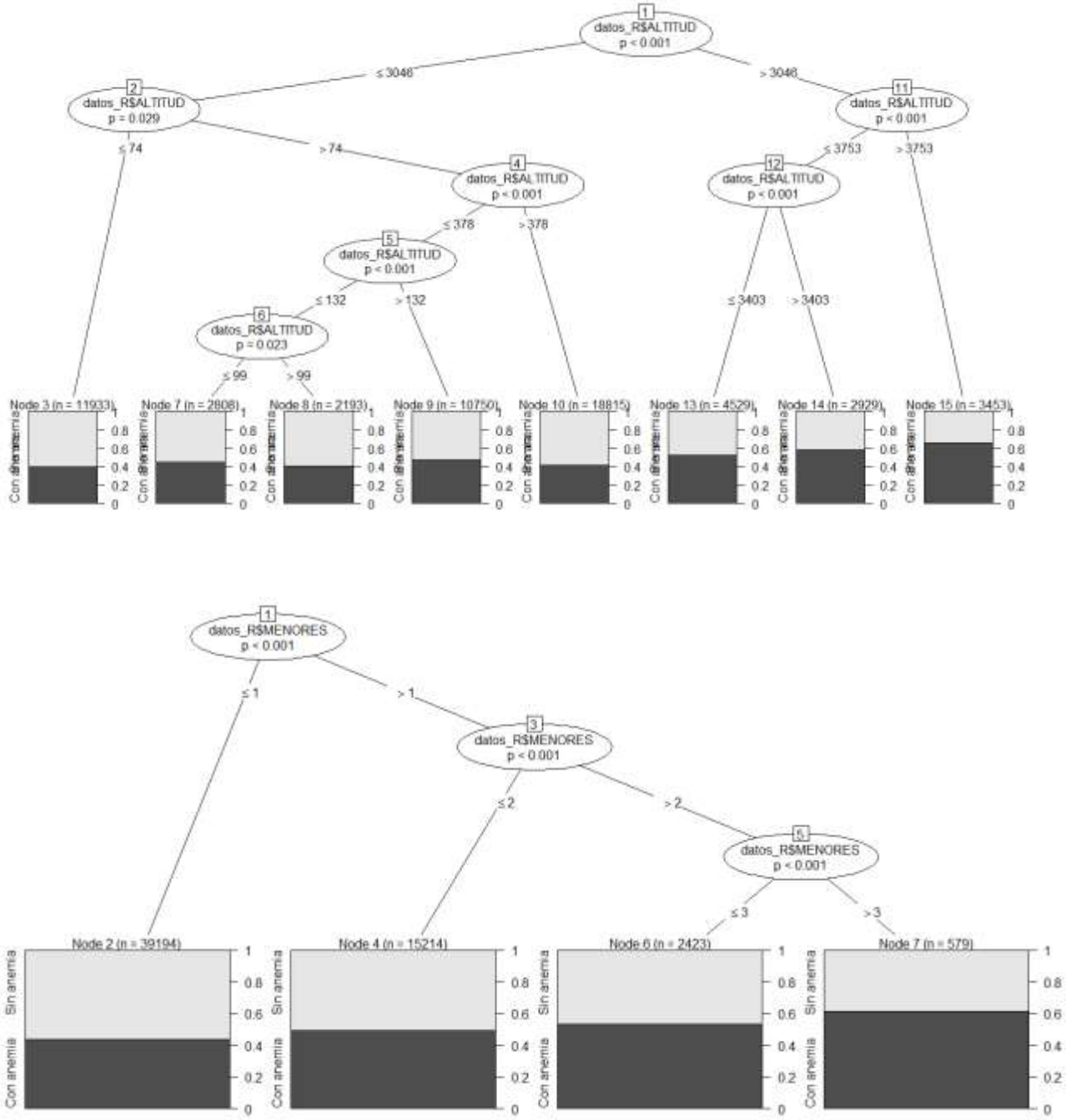


Figura B4. *Árbol de decisión CHAID correspondiente a la variable ALTITUD (altitud del conglomerado, arriba) y MENORES (número de niños menores de 5 años en el hogar, abajo)*

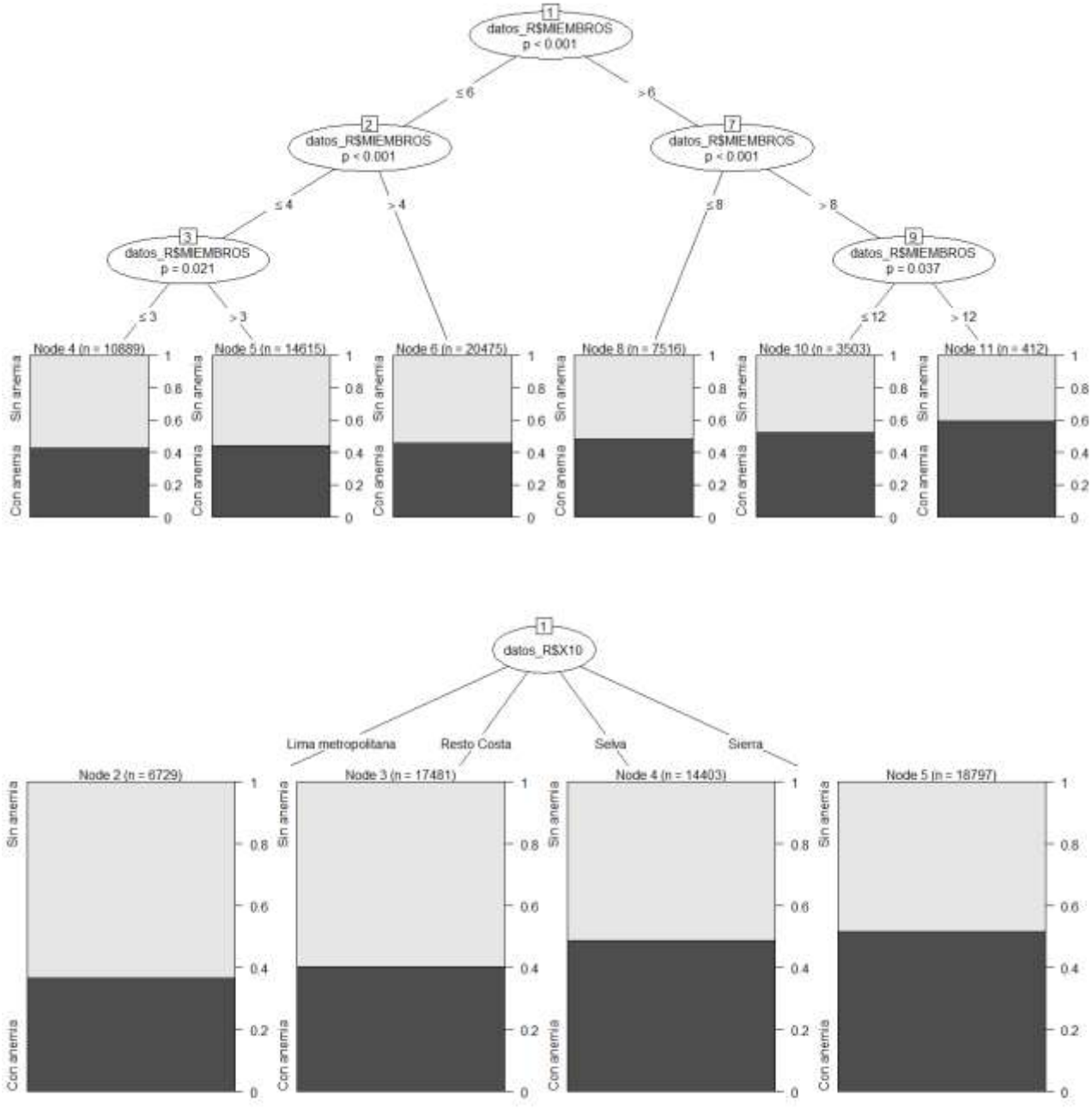


Figura B5. *Árbol de decisión CHAID correspondiente a la variable MIEMBROS (número de miembros del hogar, arriba) y X10 (región natural, abajo)*

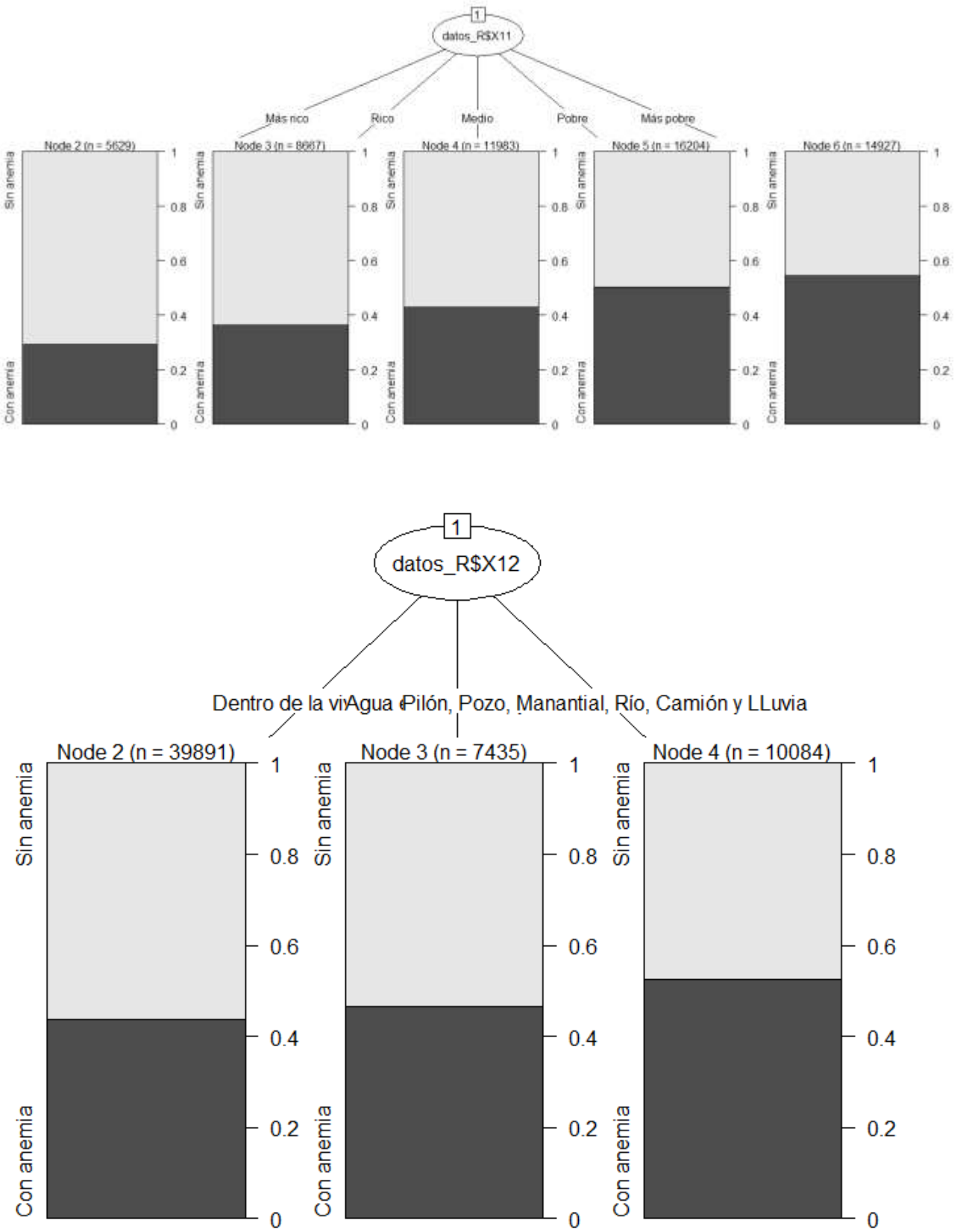


Figura B6. *Árbol de decisión CHAID correspondiente a la variable X11 (índice de riqueza, arriba) y X12 (fuente principal de abastecimiento de agua potable, abajo)*

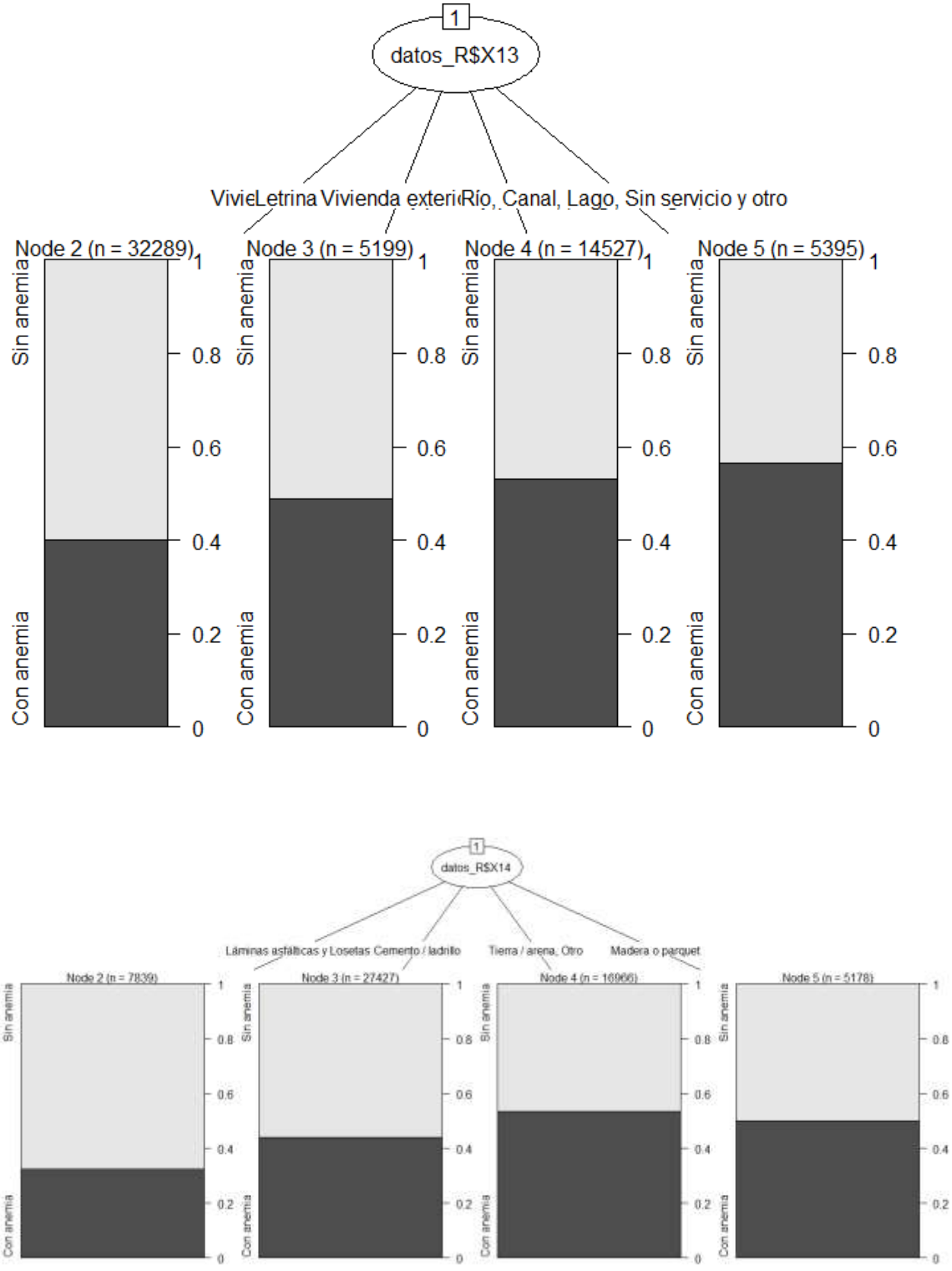


Figura B7. *Árbol de decisión CHAID correspondiente a la variable X13 (tipo de instalación sanitaria, arriba) y X14 (material predominante del piso de la vivienda, abajo)*

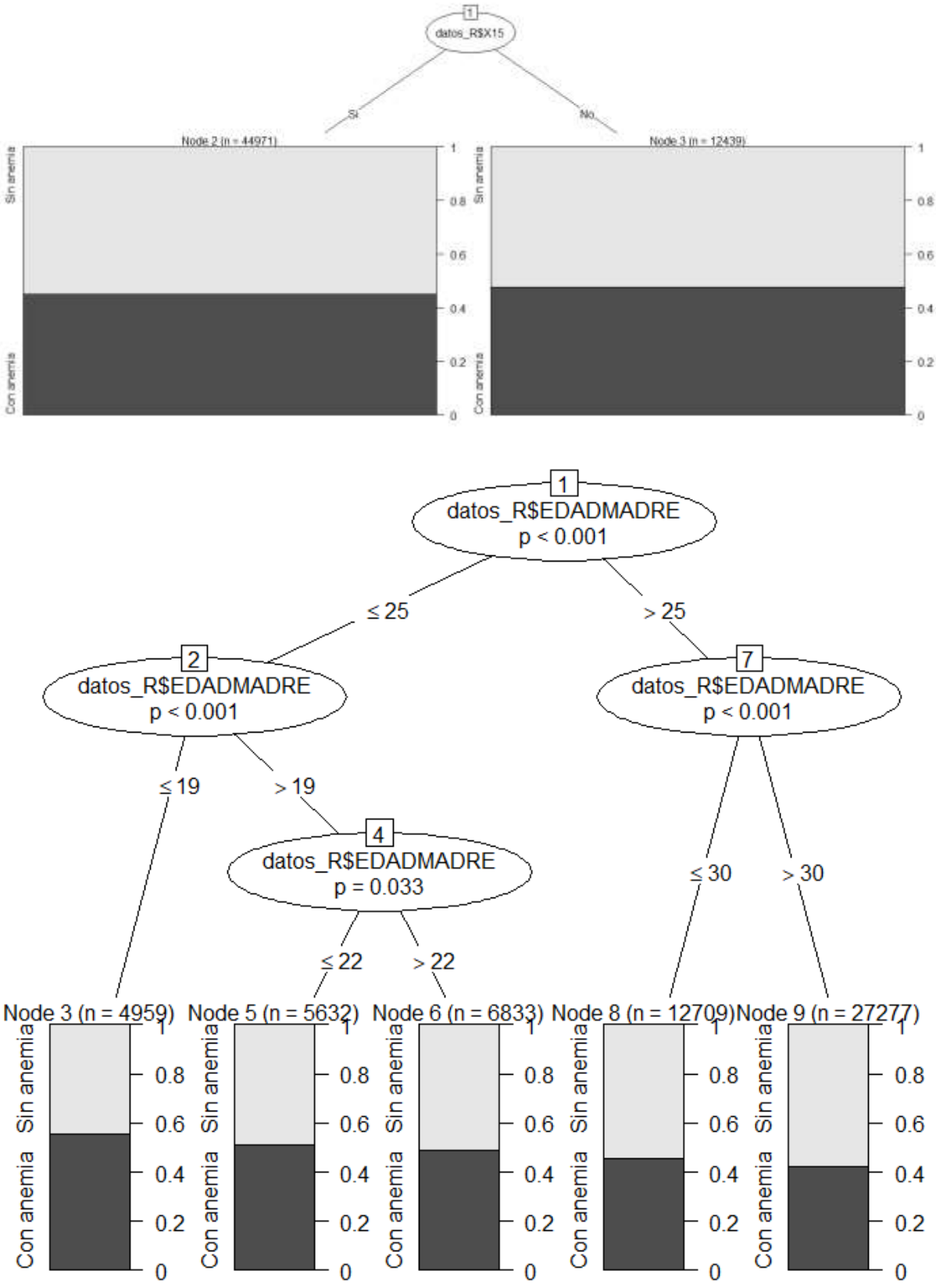


Figura B8. *Árbol de decisión CHAID correspondiente a la variable X15 (el agua usualmente es tratada: hervida, arriba) y EDADMADRE (edad de la madre, abajo)*

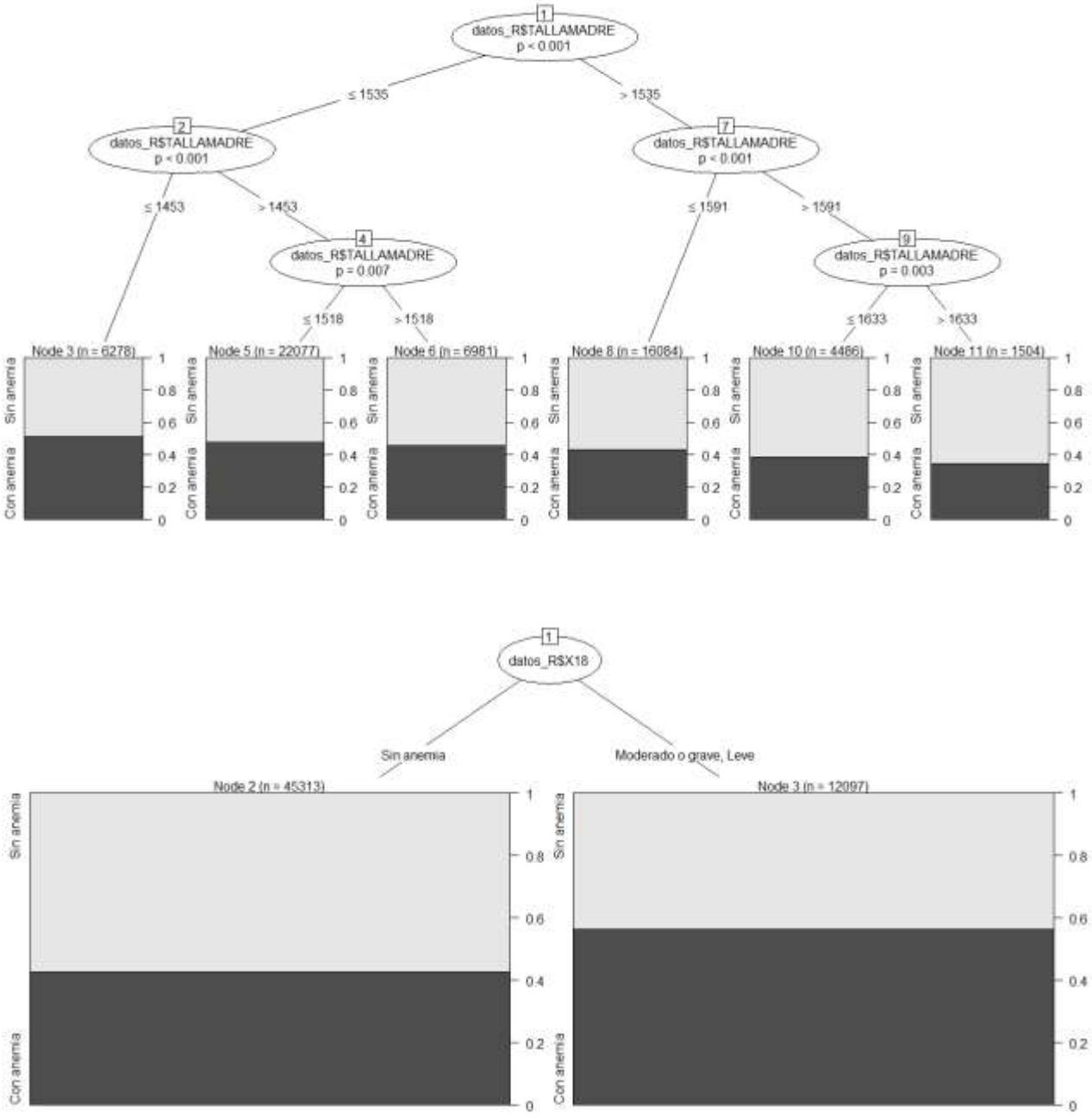


Figura B9. *Árbol de decisión CHAID correspondiente a la variable TALLAMADRE (talla de la madre, arriba) y X18 (nivel de anemia en la madre, abajo)*

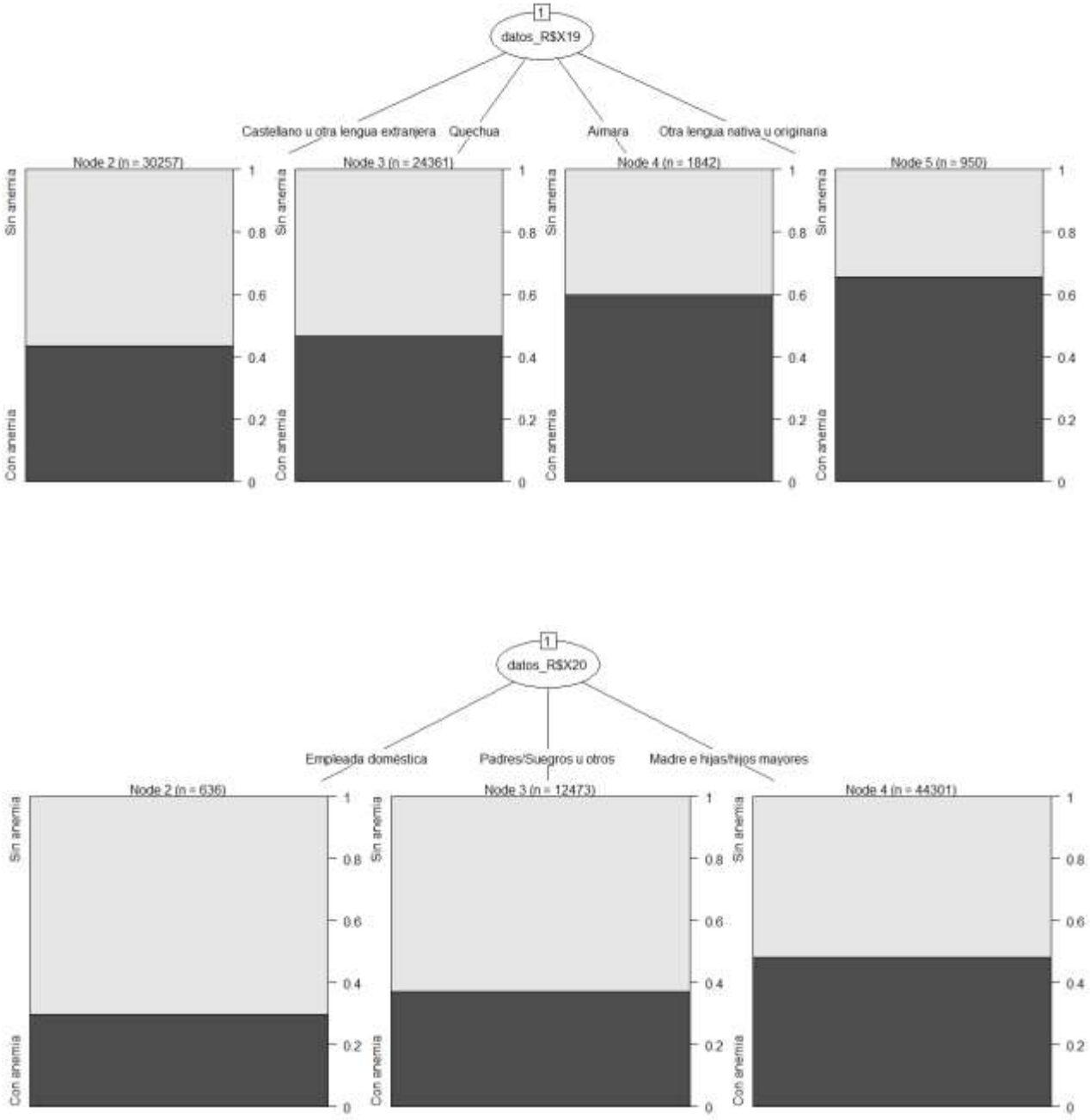


Figura B10. *Árbol de decisión CHAID correspondiente a la variable X19 (etnicidad de la madre, arriba) y X20 (persona que normalmente alimenta al niño, abajo)*

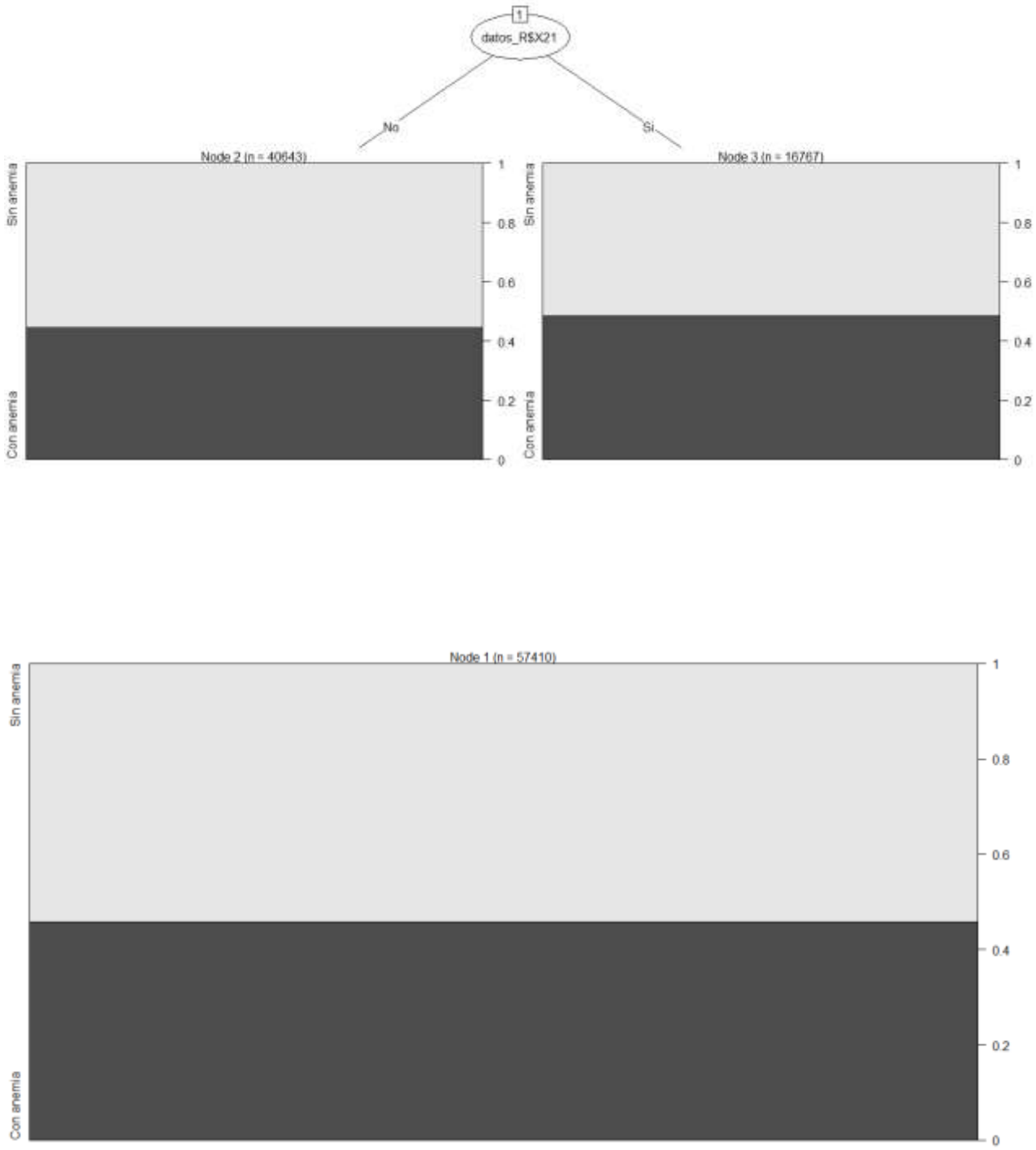


Figura B11. *Árbol de decisión CHAID correspondiente a la variable X21 (niño tomo hierro en jarabe, polvo, gotas u otra presentación, arriba) y X22 (le hicieron algún control de crecimiento y desarrollo al niño, abajo)*

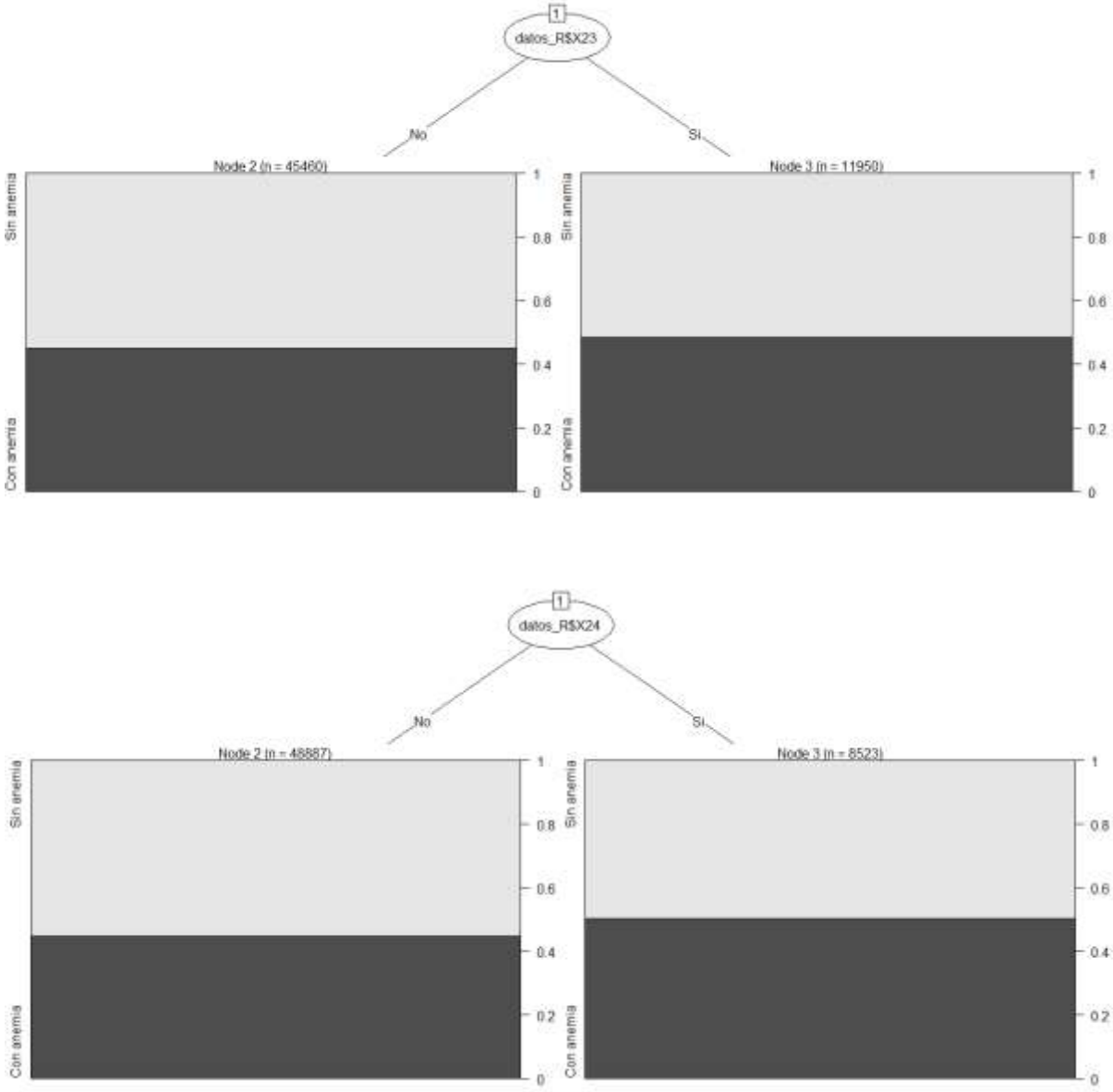


Figura B12. *Árbol de decisión CHAID correspondiente a la variable X23 (ha tenido fiebre en las últimas dos semanas, arriba) y X24 (en los últimos 14 días ha tenido diarrea la/el niña(o), abajo)*

Figura B13. *Árbol de decisión CHAID correspondiente a la variable X25 (ha tenido tos en las últimas dos semanas, arriba) y X26 (alguna vez recibió la dosis de vitamina A, abajo)*

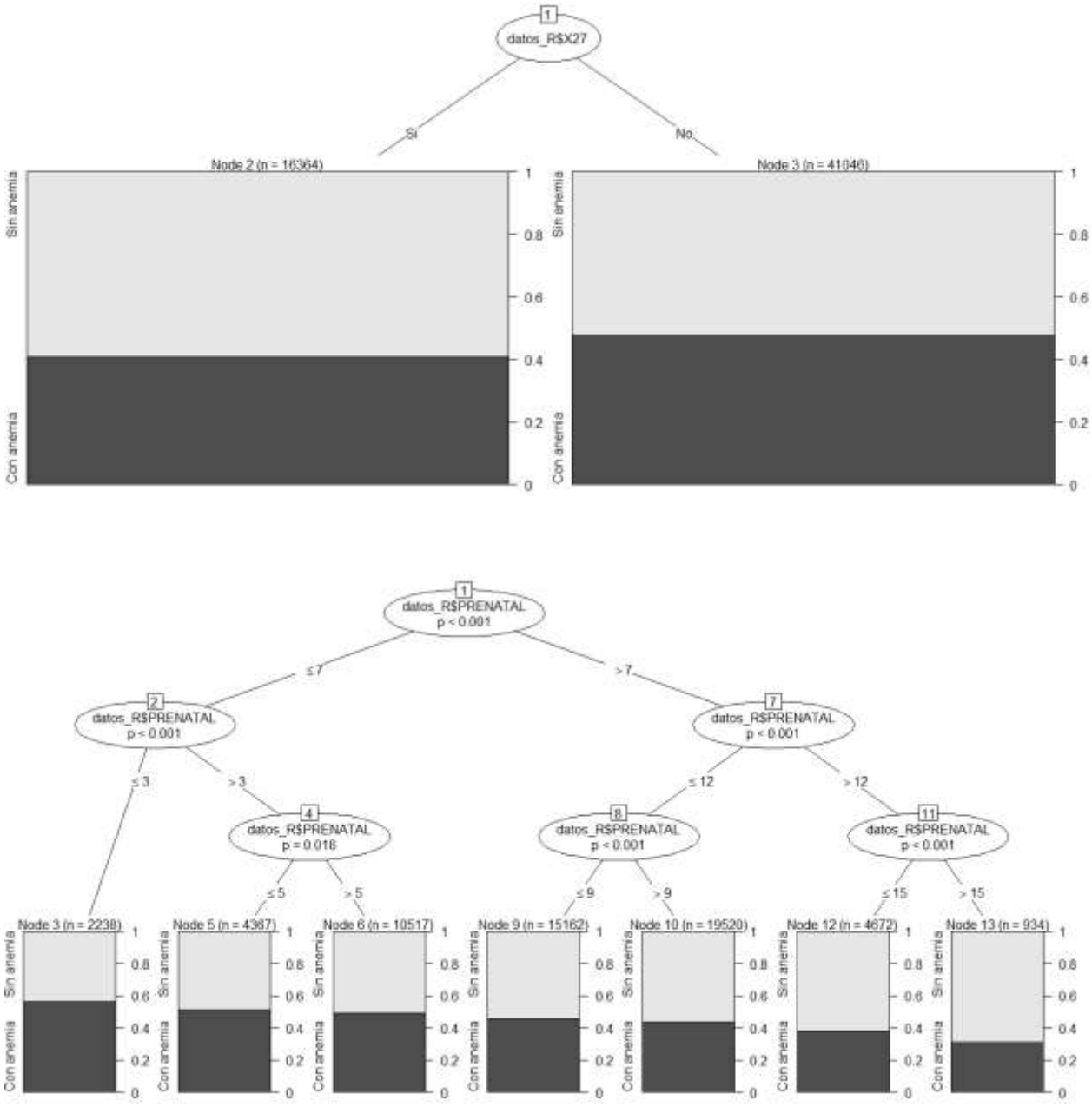


Figura B14. *Árbol de decisión CHAID correspondiente a la variable X27 (tomó medicamentos para parásitos intestinales en los últimos 6 meses, arriba) y PRENATAL (visitas prenatales por embarazo, abajo)*

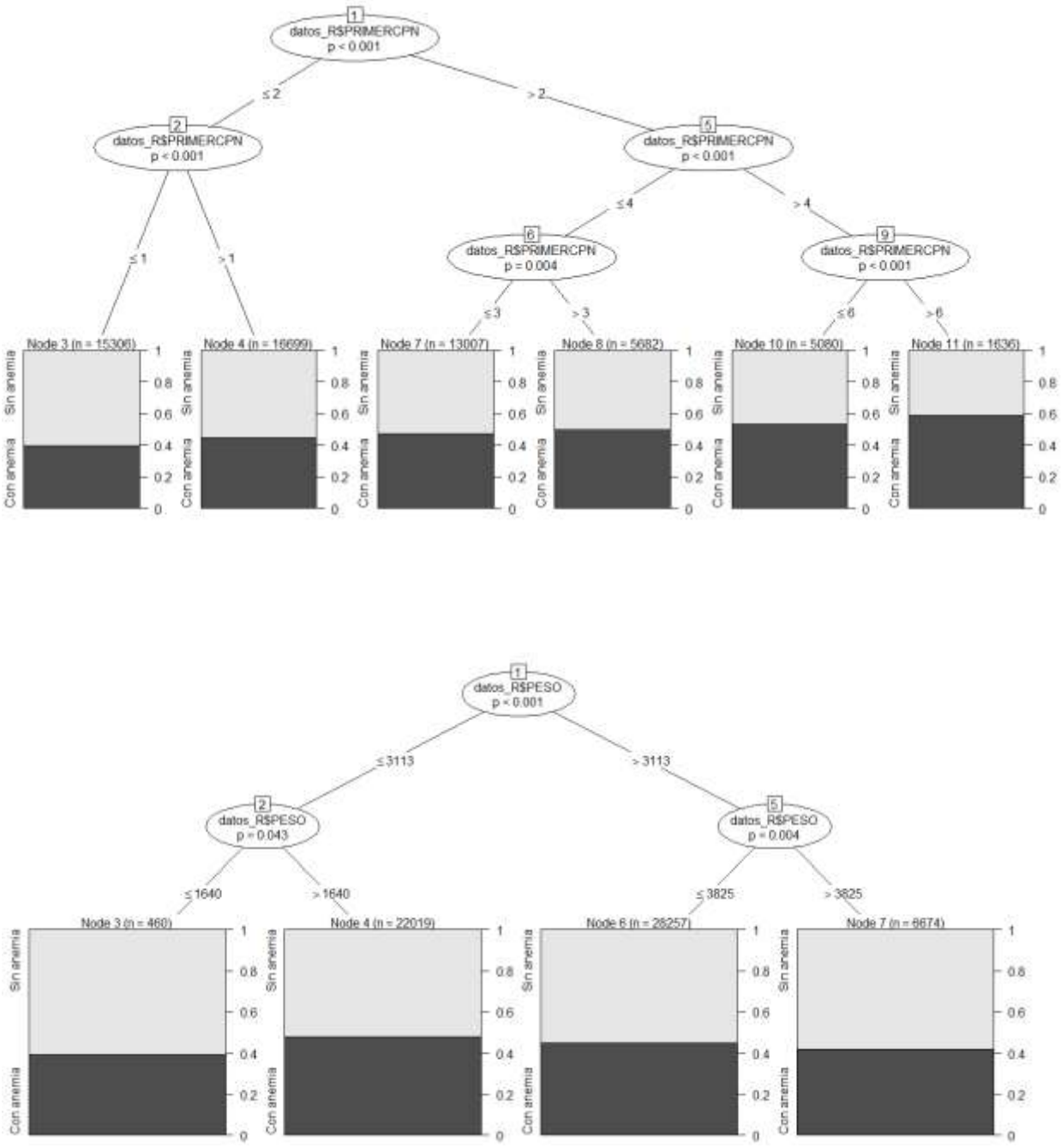


Figura B15. *Árbol de decisión CHAID correspondiente a la variable PRIMERCPN (momento del primer control prenatal, arriba) y PESO (peso del niño al nacer, abajo)*

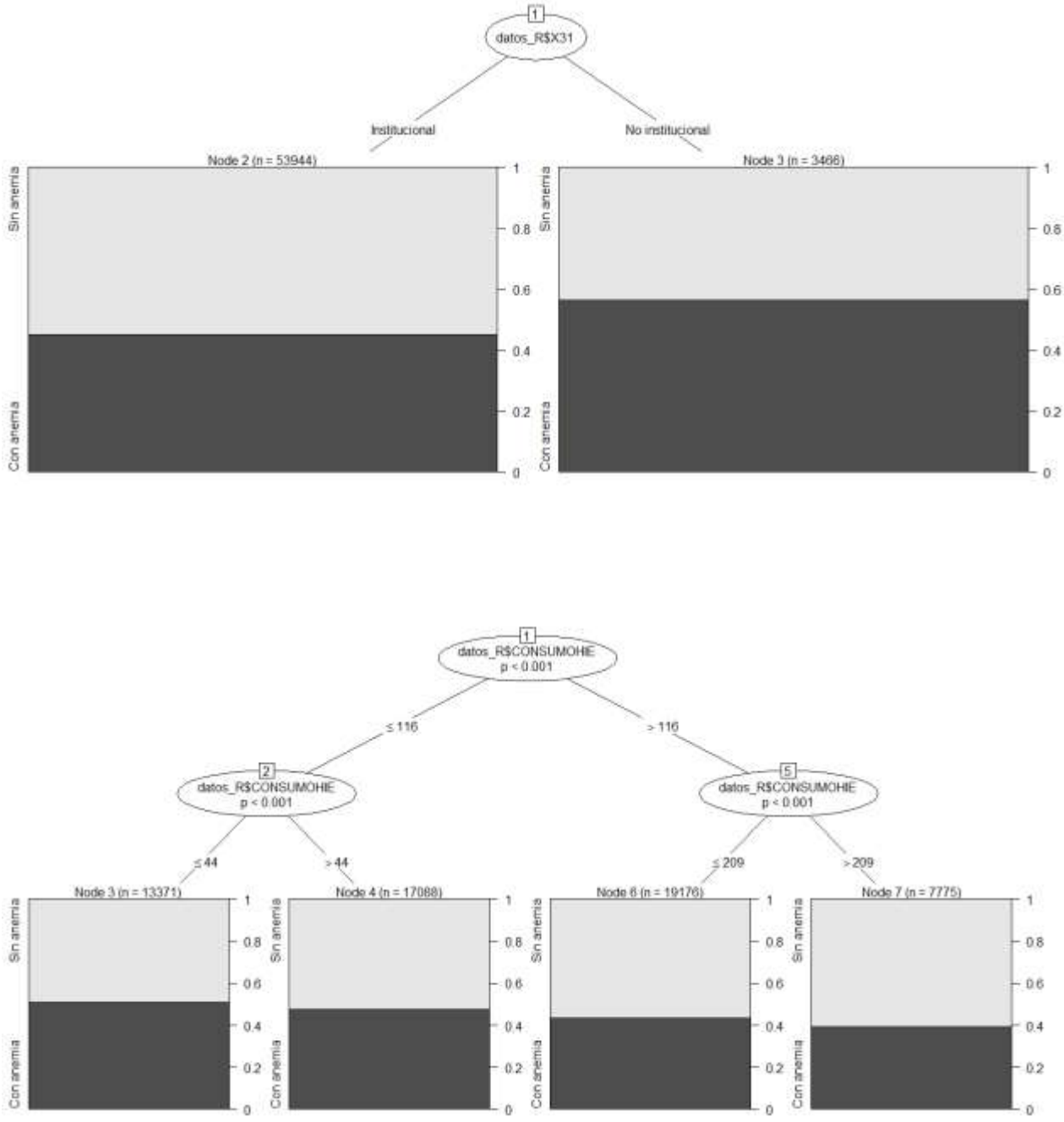


Figura B16. *Árbol de decisión CHAID correspondiente a la variable X31 (parto institucional, arriba) y CONSUMOHIE (por cuantos días tomo hierro y/o cuantas inyecciones recibió, abajo)*

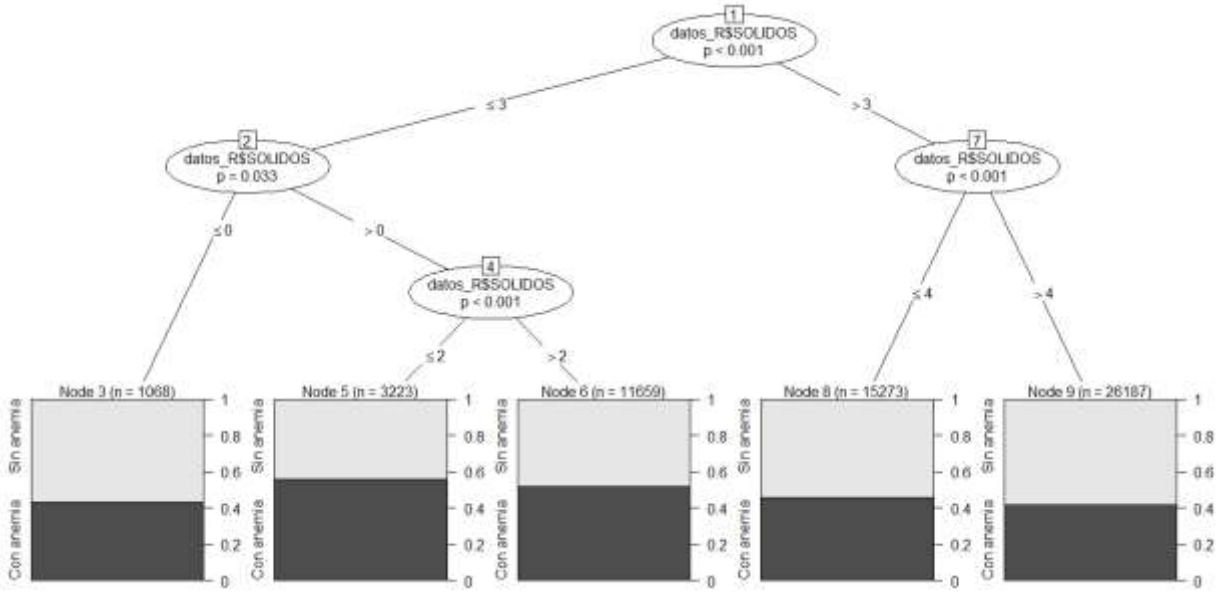


Figura B17. *Árbol de decisión CHAID correspondiente a la variable SOLIDOS (comidas sólidas o semisólidas)*

Anexo C

Tablas de clasificación de las categorías estimadas y la variable respuesta

Tabla C1. *Matriz de clasificación según procedimiento A*

Predicción	ANEMIA	Casos
0	0	6,719
0	1	3,260
1	0	2,630
1	1	4,613

Donde: 0: Sin anemia; 1: Con anemia

Tabla C2. *Matriz de clasificación según procedimiento B*

Predicción	ANEMIA	Casos
0	0	6,191
0	1	2,877
1	0	3,158
1	1	4,996

Donde: 0: Sin anemia; 1: Con anemia

Tabla C3. *Matriz de clasificación según procedimiento C*

Predicción	ANEMIA	Casos
0	0	6,664
0	1	3,279
1	0	2,685
1	1	4,594

Donde: 0: Sin anemia; 1: Con anemia

Tabla C4. *Matriz de clasificación según procedimiento D*

Predicción	ANEMIA	Casos
0	0	6,180
0	1	2,864
1	0	3,169
1	1	5,009

Donde: 0: Sin anemia; 1: Con anemia

Tabla C5. *Matriz de clasificación según procedimiento E*

Predicción	ANEMIA	Casos
0	0	6,662
0	1	3,288
1	0	2,687
1	1	4,585

Donde: 0: Sin anemia; 1: Con anemia

Tabla C6. *Matriz de clasificación según procedimiento F*

Predicción	ANEMIA	Casos
0	0	6,159
0	1	2,864
1	0	3,190
1	1	5,009

Donde: 0: Sin anemia; 1: Con anemia

Anexo D

Nivel de importancia de las variables en cada procedimiento alternativo

Tabla D1. *Resultados según procedimiento A*

Variables	Mean Decrease Gini
X01	417.87206
X02	343.01365
X03	2018.04028
X04	779.06551
X05	807.02232
X06	208.50334
X07	1162.6961
X08	398.07353
X09	967.79628
X10	624.56401
X11	729.80486
X12	419.84524
X13	558.0246
X14	556.00769
X15	241.60772
X16	888.88071
X17	1087.5784
X18	291.82548
X19	475.02227
X20	280.16645
X21	295.11881
X22	233.6044
X23	262.20359
X24	235.57368
X25	318.23531
X26	284.20151
X27	269.55542
X28	1123.07398
X29	1075.87498
X30	600.16401
X31	96.01715
X32	825.37059
X33	780.10359

Tabla D2. *Resultados según procedimiento B*

Variables	Mean Decrease Gini
X01	419.44517
X02	335.18989
X03	2065.33178
X04	788.46619
X05	830.25868
X06	204.47097
X07	1187.07911
X08	398.27096
X09	972.25175
X10	626.51262
X11	747.29169
X12	411.42457
X13	585.37482
X14	571.57943
X15	235.13309
X16	883.77915
X17	1072.55914
X18	296.63711
X19	470.50012
X20	283.39985
X21	291.34532
X22	230.59001
X23	258.87384
X24	233.32409
X25	314.22497
X26	275.95831
X27	263.76259
X28	1153.93868
X29	1106.53584
X30	593.73279
X31	93.84852
X32	852.221
X33	789.96926

Tabla D3. *Resultados según procedimiento C*

Variables	Mean Decrease Gini
X01	411.03601
X02	342.27813
X03	2007.57591
X04	778.72152
X05	807.21582
X06	207.76446
X07	1175.24297
X08	391.89442
X09	971.67669
X10	627.92274
X11	719.92471
X12	421.03931
X13	557.96844
X14	560.43298
X15	240.93784
X16	888.25731
X17	1086.8255
X18	290.27128
X19	473.62608
X20	281.75192
X21	294.63993
X22	236.08353
X23	261.92444
X24	234.27029
X25	319.59429
X26	282.80919
X27	267.19781
X28	1128.99499
X29	1078.04751
X30	600.36343
X31	96.00072
X32	826.15665
X33	789.80719

Tabla D4. *Resultados según procedimiento D*

Variables	Mean Decrease Gini
X01	420.08387
X02	339.34747
X03	2064.54208
X04	789.57733
X05	828.53339
X06	205.54318
X07	1182.95602
X08	394.47202
X09	972.24187
X10	620.5898
X11	756.18062
X12	414.85122
X13	579.73601
X14	575.56368
X15	239.2712
X16	885.10351
X17	1082.85751
X18	293.37562
X19	470.08296
X20	285.98068
X21	294.06399
X22	232.44936
X23	257.97356
X24	231.06829
X25	319.07662
X26	276.10185
X27	264.27353
X28	1149.08656
X29	1106.0932
X30	594.36247
X31	93.74052
X32	843.783
X33	782.29181

Tabla D5. *Resultados según procedimiento E*

Variables	Mean Decrease Gini
X01	414.7998
X02	342.50677
X03	2010.72675
X04	780.52339
X05	808.25783
X06	209.08034
X07	1168.38027
X08	392.75262
X09	970.78659
X10	625.32792
X11	724.33957
X12	418.31752
X13	555.89394
X14	562.66324
X15	241.90868
X16	884.60306
X17	1082.82128
X18	292.52546
X19	471.35992
X20	280.59091
X21	295.67379
X22	236.32635
X23	261.36378
X24	233.2585
X25	320.51454
X26	280.01877
X27	264.13307
X28	1131.2399
X29	1081.72206
X30	602.81817
X31	95.57022
X32	831.4
X33	782.99334

Tabla D6. *Resultados según procedimiento F*

Variables	Mean Decrease Gini
X01	421.12388
X02	338.60955
X03	2062.11488
X04	776.47265
X05	823.97489
X06	206.17056
X07	1184.6701
X08	395.40368
X09	973.30731
X10	626.03822
X11	747.1357
X12	414.93898
X13	585.41032
X14	565.05443
X15	237.81372
X16	879.97646
X17	1078.62673
X18	296.61653
X19	469.51308
X20	288.87724
X21	294.44802
X22	232.3727
X23	259.11775
X24	230.31564
X25	315.04842
X26	276.25952
X27	266.26444
X28	1158.65397
X29	1112.65744
X30	595.59924
X31	93.25628
X32	847.70419
X33	787.76126

Anexo E

Sintaxis en IBM SPSS para la construcción de la base de datos y la creación de variables

Encoding: UTF-8.

```
RECODE ID1 (SYSMIS=2019).
RECODE ID1 (SYSMIS=2018).
RECODE ID1 (SYSMIS=2017).
RECODE ID1 (SYSMIS=2016).
RECODE ID1 (SYSMIS=2015).

*****Despues de fundir y apilar las bases de datos de los modulos: RECH6,
REC0111, ******RECH0, REC91, RECH23, RECH5, REC95, REC43 y REC41.
****** en una sola base de datos para los años 2015 al 2019 se crean las variables.
FREQ HW57.
FREQ ID1.
DATASET ACTIVATE ConjuntoDatos1.
ADD FILES /FILE=*
  /FILE='ConjuntoDatos2'.
EXECUTE.

DATASET ACTIVATE ConjuntoDatos1.
ADD FILES /FILE=*
  /FILE='ConjuntoDatos5'.
EXECUTE.

FREQ M39.
FREQ H34 H22 H11 H31 H41A H41B H42 H43.
FREQ S465DD_DC S465EA.
FREQ S493I S493N.
FREQ S465EA S465EB S465EC S465ED.
FREQ S465DB_A S465DB_B S465DB_C S465DB_D .
FREQ S466 S466B S466C.
FREQ ID1.
FREQ HC13 HC55 ID1 HC57.
FREQ HA57 HC57.
FREQ HC52 HC55 HC57.

CROSSTABS
  /TABLES=HA57 BY HC57
  /FORMAT=AVALUE TABLES
  /CELLS=COUNT
  /COUNT ROUND CELL.

CROSSTABS
  /TABLES=HA57 BY SH64
  /FORMAT=AVALUE TABLES
```

```
  /CELLS=COUNT
  /COUNT ROUND CELL.


COMPUTE ddd=CHAR.SUBSTR(CASEID,1,16).
EXECUTE.

FREQ M3A.
COMPUTE PARTO1=M3A + M3B + M3C.
IF PARTO1 >=1 & PARTO1<=3 PARTO = 1.
IF PARTO1=0 PARTO=0.

FREQ PARTO1 PARTO.
SAVE OUTFILE='C:\Users\Home\Downloads\Tesis UNS\ENDES '+
    '2015-2019\REC41-43-95-H6-H5-91-0111-H0-H6_2015-2019.SAV'
  /COMPRESSED.
VALUE LABELS
/PARTO
0 'No institucional'
1 'Institucional'.

DELETE VARIABLES PARTO1.
DELETE VARIABLES M3A M3B M3C.


IF (S465EA>=2 OR S465EB>=2 OR S465EC>=2 OR S465ED>=2)  HIERRO=0.
IF (S465EA=1 OR S465EB=1 OR S465EC=1 OR S465ED=1)  HIERRO=1.
VALUE LABELS
/HIERRO
0 'No'
1 'Si'.

FREQ M15 HIERRO.
DELETE VARIABLES S465EA S465EB S465EC S465ED.
freq v101 v102 v139 v140 v141.
freq s119 v131.
FREQ HA1 V012.

FILTER OFF.
USE ALL.
SELECT IF (HC55 >= 0 & HC55 <= 6).
EXECUTE.

SAVE OUTFILE='C:\Users\Home\Downloads\Tesis UNS\ENDES 2015-
2019\Base_anemia_niños_2015-2019.SAV'
  /COMPRESSED.

GET
  FILE='C:\Users\Home\Downloads\Tesis UNS\ENDES 2015-
2019\Base_FINAL_anemia_niños_2015-2019.SAV'.
```

```
FREQ HC57 HC61 HC27 HC1 HC64 HC63 V102 V040 HV014 V136 SREGION
V190 V113 V116 HV213 HV237A HA1 HA3 HA57 V131 S496 HIERRO S466 H22
H11 H31 H41B H43 M14 M13 M19 PARTO M46 M39.
EXECUTE.

COMPUTE ANEMIA=.
IF (HC57>=1 & HC57<=3) ANEMIA=1.
IF HC57=4 ANEMIA=0.
VALUE LABELS
/ANEMIA
1 'Con anemia'
0 'Sin anemia'.
VARIABLE LABELS ANEMIA 'Prevalencia de Anemia'.

NUMERIC X01 (F8.0).
EXECUTE.
IF HC61=3 X01=1.
IF HC61=2 X01=2.
IF HC61=1 X01=3.
IF HC61=0 X01=4.
VALUE LABELS
/X01
1 'Superior'
2 'Secundaria'
3 'Primaria'
4 'Sin educación'.
VARIABLE LABELS X01 'Nivel educativo más alto de la madre'.

NUMERIC X02 (F8.0).
EXECUTE.
IF HC27=1 X02=2.
IF HC27=2 X02=1.
VALUE LABELS
/X02
1 'Mujer'
2 'Hombre'.
VARIABLE LABELS X02 'Sexo'.

NUMERIC EDAD (F8.0).
EXECUTE.
COMPUTE EDAD=HC1.
VARIABLE LABELS EDAD 'Edad (En meses)'.

NUMERIC ORDEN (F8.0).
EXECUTE.
COMPUTE ORDEN=HC64.
VARIABLE LABELS ORDEN 'Orden de nacimiento del niño'.

NUMERIC INTERVALO (F8.0).
```

```
EXECUTE.
COMPUTE INTERVALO=HC63.
VARIABLE LABELS INTERVALO 'Intervalo entre nacimientos anteriores al niño'.

NUMERIC X06 (F8.0).
EXECUTE.
IF V102=1 X06=1.
IF V102=2 X06=2.
VALUE LABELS
/X06
1 'Urbana'
2 'Rural'.
VARIABLE LABELS X06 'Lugar de residencia'.

NUMERIC ALTITUD (F8.0).
EXECUTE.
COMPUTE ALTITUD=V040.
VARIABLE LABELS ALTITUD 'Altitud del conglomerado (en metros)'.

NUMERIC MENORES (F8.0).
EXECUTE.
COMPUTE MENORES=HV014.
VARIABLE LABELS MENORES 'Número de niños menores de 5 años'.

NUMERIC MIEMBROS (F8.0).
EXECUTE.
COMPUTE MIEMBROS=V136.
VARIABLE LABELS MIEMBROS 'Número de miembros del hogar'.

NUMERIC X10 (F8.0).
EXECUTE.
IF SREGION=1 X10=1.
IF SREGION=2 X10=2.
IF SREGION=4 X10=3.
IF SREGION=3 X10=4.
VALUE LABELS
/X10
1 'Lima metropolitana'
2 'Resto Costa'
3 'Selva'
4 'Sierra'.
VARIABLE LABELS X10 'Región natural'.

NUMERIC X11 (F8.0).
EXECUTE.
IF V190=1 X11=5.
IF V190=2 X11=4.
IF V190=3 X11=3.
IF V190=4 X11=2.
IF V190=5 X11=1.
```

```
VALUE LABELS
/X11
1 'Más rico'
2 'Rico'
3 'Medio'
4 'Pobre'
5 'Más pobre'.
VARIABLE LABELS X11 'Índice de riqueza'.

NUMERIC X12 (F8.0).
EXECUTE.
IF V113=11 X12=1.
IF (V113>=71 & V113<=96) X12=2.
IF (V113>=12 & V113<=61) X12=3.
VALUE LABELS
/X12
1 'Dentro de la vivienda'
2 'Agua embotellada y otro'
3 'Pilón, Pozo, Manantial, Río, Camión y LLuvia'.
VARIABLE LABELS X12 'Fuente principal de abastecimiento de agua potable que
utilizan en su hogar para tomar o beber'.

NUMERIC X13 (F8.0).
EXECUTE.
IF V116=11 X13=1.
IF (V116>=21 & V116<=22) X13=2.
IF V116=12 X13=3.
IF V116=23 X13=4.
IF (V116>=24 & V116<=96) X13=5.
VALUE LABELS
/X13
1 'Vivienda interior'
2 'Letrina ventilada y pozo séptico'
3 'Vivienda exterior'
4 'Latrina (ciego o negro)'
5 'Río, Canal, Lago, Sin servicio y otro'.
VARIABLE LABELS X13 'Tipo de instalación sanitaria'.

NUMERIC X14 (F8.0).
EXECUTE.
IF (HV213>=32 & HV213<=33) X14=1.
IF HV213=34 X14=2.
IF HV213=11 X14=3.
IF (HV213>=12 & HV213<=31) X14=4.
IF (HV213>=35 & HV213<=96) X14=5.
VALUE LABELS
/X14
1 'Láminas asfálticas y Losetas'
2 'Cemento / ladrillo'
3 'Tierra / arena'
```

```
4 'Madera o parquet'
5 'Otro'.
VARIABLE LABELS X14 'Material predominante del piso de la vivienda'.

NUMERIC X15 (F8.0).
EXECUTE.
IF HV237A=1 X15=1.
IF HV237A=0 X15=2.
VALUE LABELS
/X15
1 'Si'
2 'No'.
VARIABLE LABELS X15 'El agua usualmente es tratada por: hervida'.

NUMERIC EDADMADRE (F8.0).
EXECUTE.
COMPUTE EDADMADRE=HA1.
VARIABLE LABELS EDADMADRE 'Edad de la mujer (en años)'.

NUMERIC TALLAMADRE (F8.0).
EXECUTE.
COMPUTE TALLAMADRE=HA3.
VARIABLE LABELS TALLAMADRE 'Talla en centímetros (1 decimal)'.

NUMERIC X18 (F8.0).
EXECUTE.
IF HA57=4 X18=1.
IF (HA57=1 OR HA57=2) X18=2.
IF HA57=3 X18=3.
VALUE LABELS
/X18
1 'Sin anemia'
2 'Moderado o grave'
3 'Leve'.
VARIABLE LABELS X18 'Nivel de anemia en la madre'.

NUMERIC X19 (F8.0).
EXECUTE.
IF (V131>=10 & V131<=12) X19=1.
IF V131=1 X19=2.
IF V131=2 X19=3.
IF (V131>=3 & V131<=9) X19=4.
VALUE LABELS
/X19
1 'Castellano u otra lengua extranjera'
2 'Quechua'
3 'Aimara'
4 'Otra lengua nativa u originaria'.

VARIABLE LABELS X19 'Etnicidad'.
```

```
NUMERIC X20 (F8.0).
EXECUTE.
IF S496=8 X20=1.
IF S496=2 X20=2.
IF (S496>=4 & S496<=7) X20=2.
IF (S496>=9 & S496<=96) X20=2.
IF (S496=1 OR S496=3) X20=3.
VALUE LABELS
/X20
1 'Empleada doméstica'
2 'Padres/Suegros u otros'
3 'Madre e hijas/hijos mayores'.
VARIABLE LABELS X20 'Persona que normalmente alimenta al niño'.

NUMERIC X21 (F8.0).
EXECUTE.
IF HIERRO=0 X21=1.
IF HIERRO=1 X21=2.
VALUE LABELS
/X21
1 'No'
2 'Si'.
VARIABLE LABELS X21 'Niño tomó hierro en jarabe, polvo, gotas u otra
presentación'.

NUMERIC X22 (F8.0).
EXECUTE.
IF S466=0 X22=1.
IF S466=1 X22=2.
VALUE LABELS
/X22
1 'No'
2 'Si'.
VARIABLE LABELS X22 'Le hicieron algún control de crecimiento y desarrollo'.

NUMERIC X23 (F8.0).
EXECUTE.
IF H22=0 X23=1.
IF H22=1 X23=2.
VALUE LABELS
/X23
1 'No'
2 'Si'.
VARIABLE LABELS X23 'Ha tenido fiebre en las últimas dos semanas'.

NUMERIC X24 (F8.0).
EXECUTE.
IF H11=0 X24=1.
IF H11=2 X24=2.
```

```
VALUE LABELS
/X24
1 'No'
2 'Si'.
VARIABLE LABELS X24 'En los últimos 14 dìas, ha tenido diarrea la niña(o)'.

NUMERIC X25 (F8.0).
EXECUTE.
IF H31=0 X25=1.
IF H31=2 X25=2.
VALUE LABELS
/X25
1 'No'
2 'Si'.
VARIABLE LABELS X25 'Ha tenido tos en las últimas dos semanas'.

NUMERIC X26 (F8.0).
EXECUTE.
IF H41B=0 X26=1.
IF H41B=1 X26=2.
VALUE LABELS
/X26
1 'No'
2 'Si'.
VARIABLE LABELS X26 'Alguna vez recibió la dosis de vitamina A'.

NUMERIC X27 (F8.0).
EXECUTE.
IF H43=1 X27=1.
IF H43=0 X27=2.
VALUE LABELS
/X27
1 'Si'
2 'No'.
VARIABLE LABELS X27 'Medicamentos para parásitos intestinales en los últimos 6
meses'.

NUMERIC PRENATAL (F8.0).
EXECUTE.
COMPUTE PRENATAL=M14.
VARIABLE LABELS PRENATAL 'Visitas prenatales por embarazo'.
RECODE PRENATAL (98=SYSMIS).

NUMERIC PRIMERCPN (F8.0).
EXECUTE.
COMPUTE PRIMERCPN=M13.
VARIABLE LABELS PRIMERCPN 'Momento del primer control prenatal'.
RECODE PRIMERCPN (98=SYSMIS).

NUMERIC PESO (F8.0).
```

```
EXECUTE.
COMPUTE PESO=M19.
VARIABLE LABELS PESO 'Peso del niño al nacer  (kilos - 3 dec.)'.
RECODE PESO (9996=SYSMIS).
RECODE PESO (9998=SYSMIS).
RECODE PESO (9999=SYSMIS).

RECODE PARTO (0=2) (1=1).
EXECUTE.
VALUE LABELS
/PARTO
1 'Institucional'
2 'No institucional'.

NUMERIC CONSUMOHIE (F8.0).
EXECUTE.
COMPUTE CONSUMOHIE=M46.
VARIABLE LABELS CONSUMOHIE 'Por cuantos días tomó hierro y/o cuantas
inyecciones recibió'.
RECODE CONSUMOHIE (998=SYSMIS).
RECODE CONSUMOHIE (999=SYSMIS).

NUMERIC SOLIDOS (F8.0).
EXECUTE.
COMPUTE SOLIDOS=M39.
VARIABLE LABELS SOLIDOS 'El día de ayer o durante el dìa o la noche cuantas
veces le dio comida sólidas o semisólidas'.
RECODE SOLIDOS (8=SYSMIS).
FREQ M39 SOLIDOS.

SAVE OUTFILE='C:\Users\Home\Downloads\Tesis UNS\ENDES 2015-
2019\Base_FINAL_anemia_niños_2015-2019.SAV'
  /COMPRESSED.

*******PREPARANDO EL NUEVO ARCHIVO.
*FILTROS.
*ELIMINAR MISSING DE VARIABLE ANEMIA.
*EDAD DEL NIÑO DE 6 A 35 MESES DE EDAD.

FREQ ANEMIA.

FILTER OFF.
USE ALL.
SELECT IF (ANEMIA <=1).
EXECUTE.

FILTER OFF.
USE ALL.
SELECT IF (EDAD >= 6 & EDAD <= 35).
EXECUTE.
```

```
SAVE OUTFILE='C:\Users\Home\Downloads\Tesis UNS\ENDES 2015-
2019\Base_FINAL2_anemia_niños_2015-2019.SAV'
  /COMPRESSED.

DELETE VARIABLES ID1 HC1 HC27 HC56 HC57 HC61 HC63 HC64 M13 M14 M18
M19 M39 M46 H11 H22 H31 H41B H43 S466 HA1 HA3 HA56 HA57 SREGION
S490.
DELETE VARIABLES S496 V026 V040 V101 V102 V113 V116 V119 V127 V131
V136 V190 HV014 HV213 HV237 HV237A HIERRO.

SAVE OUTFILE='C:\Users\Home\Downloads\Tesis UNS\ENDES 2015-
2019\Base_FINAL3_anemia_niños_2015-2019.SAV'
  /COMPRESSED.

IF ORDEN=1 INTERVALO=0.
EXECUTE.

NUMERIC X31 (F8.0).
EXECUTE.
COMPUTE X31=PARTO.
VARIABLE LABELS X31 'Parto institucional'.
EXECUTE.
VALUE LABELS
/X31
1 'Institucional'
2 'No institucional'.

FREQ X31 PARTO.
DELETE VARIABLES HHID PARTO.

SAVE OUTFILE='C:\Users\Home\Downloads\Tesis
UNS\Base_FINAL4_anemia_niños_2015-2019.SAV'
  /COMPRESSED.

*****SE DISCRETIZAN LAS VARIABLES DE ACUERDO A LOS RESULTADOS DE
*****LOS ARBOLES CHAID.

IF X01=4 X01=3.
EXECUTE.
FREQ X01.

NUMERIC X03 (F8.0).
EXECUTE.
IF (EDAD<=16) X03=8.
IF (EDAD>16 & EDAD<=18) X03=7.
IF (EDAD>18 & EDAD<=19) X03=6.
IF (EDAD>19 & EDAD<=20) X03=5.
IF (EDAD>20 & EDAD<=22) X03=4.
IF (EDAD>22 & EDAD<=25) X03=3.
```

```
IF (EDAD>25 & EDAD<=30) X03=2.
IF (EDAD>30) X03=1.
VALUE LABELS
/X03
1 '>30'
2 '<25 - 30]'
3 '<22 - 25]'
4 '<20 - 22]'
5 '<19 - 20]'
6 '<18 - 19]'
7 '<16 - 18]'
8 '<=16]'.
VARIABLE LABELS X03 'Edad (En meses)'.

NUMERIC X04 (F8.0).
EXECUTE.
IF (ORDEN<=1) X04=1.
IF (ORDEN>1 & ORDEN<=2) X04=2.
IF (ORDEN>2 & ORDEN<=3) X04=3.
IF (ORDEN>3 & ORDEN<=5) X04=4.
IF (ORDEN>5) X04=5.
VALUE LABELS
/X04
1 '<=1'
2 '<1 - 2]'
3 '<2 - 3]'
4 '<3 - 5]'
5 '>5]'.
VARIABLE LABELS X04 'Orden de nacimiento del niño'.

NUMERIC X05 (F8.0).
EXECUTE.
IF (INTERVALO>109) X05=1.
IF (INTERVALO<=11) X05=2.
IF (INTERVALO>56 & INTERVALO<=109) X05=3.
IF (INTERVALO>41 & INTERVALO<=56) X05=4.
IF (INTERVALO>11 & INTERVALO<=41) X05=5.
VALUE LABELS
/X05
1 '>109'
2 '<=11]'
3 '<56 - 109]'
4 '<41 - 56]'
5 '<11 - 41]'.
VARIABLE LABELS X05 'Intervalo entre nacimientos anteriores al niño'.
EXECUTE.

NUMERIC X07 (F8.0).
EXECUTE.
IF (ALTITUD<=74) X07=1.
```

```
IF (ALTITUD>99 & ALTITUD<=132) X07=2.
IF (ALTITUD>378 & ALTITUD<=3046) X07=3.
IF (ALTITUD>74 & ALTITUD<=99) X07=4.
IF (ALTITUD>132 & ALTITUD<=378) X07=5.
IF (ALTITUD>3046 & ALTITUD<=3403) X07=6.
IF (ALTITUD>3403 & ALTITUD<=3753) X07=7.
IF (ALTITUD>3753) X07=8.
VALUE LABELS
/X07
1 '<=74'
2 '<99 - 132]'
3 '<378 - 3046]'
4 '<74 - 99]'
5 '<132 - 378]'
6 '<3046 - 3403]'
7 '<3403 - 3753]'
8 '>3753'.
VARIABLE LABELS X07 'Altitud del conglomerado (en metros)'.

NUMERIC X08 (F8.0).
EXECUTE.
IF (MENORES<=1) X08=1.
IF (MENORES=2) X08=2.
IF (MENORES=3) X08=3.
IF (MENORES>3) X08=4.
VALUE LABELS
/X08
1 '<=1'
2 '2'
3 '3'
4 '>3'.
VARIABLE LABELS X08 'Número de niños menores de 5 años'.
EXECUTE.

NUMERIC X09 (F8.0).
EXECUTE.
IF (MIEMBROS<=3) X09=1.
IF (MIEMBROS=4) X09=2.
IF (MIEMBROS>4 & MIEMBROS<=6) X09=3.
IF (MIEMBROS>6 & MIEMBROS<=8) X09=4.
IF (MIEMBROS>8 & MIEMBROS<=12) X09=5.
IF (MIEMBROS>12) X09=6.
VALUE LABELS
/X09
1 '<=3'
2 '4'
3 '<4 - 6]'
4 '<6 - 8]'
5 '<8 - 12]'
6 '>12'.
```

```
VARIABLE LABELS X09 'Número de miembros del hogar'.
EXECUTE.

IF X13=4 X13=3.
IF X13=5 X13=4.
VALUE LABELS
/X13
1 'Vivienda interior'
2 'Letrina ventilada y pozo séptico'
3 'Vivienda exterior y Latrina (ciego o negro)'
4 'Río, Canal, Lago, Sin servicio y otro'.

IF X14=4 X14=13.
IF X14=5 X14=4.
IF X14=3 X14=4.
IF X14=13 X14=3.
VALUE LABELS
/X14
1 'Láminas ásfalticas y Losetas'
2 'Cemento / Ladrillo'
3 'Madera o parquet'
4 'Tierra / Arena, Otro'.

NUMERIC X16 (F8.0).
EXECUTE.
IF (EDADMADRE>30) X16=1.
IF (EDADMADRE>25 & EDADMADRE<=30) X16=2.
IF (EDADMADRE>22 &EDADMADRE<=25) X16=3.
IF (EDADMADRE>19 & EDADMADRE<=22) X16=4.
IF (EDADMADRE<=19) X16=5.
VALUE LABELS
/X16
1 '>30'
2 '<25 - 30]'
3 '<22 - 25]'
4 '<19 - 22]'
5 '<=19'.
VARIABLE LABELS X16 'Edad de la mujer (en años)'.
EXECUTE.

NUMERIC X17 (F8.0).
EXECUTE.
IF (TALLAMADRE>1633) X17=1.
IF (TALLAMADRE>1591 & TALLAMADRE<=1633) X17=2.
IF (TALLAMADRE>1535 &TALLAMADRE<=1591) X17=3.
IF (TALLAMADRE>1518 & TALLAMADRE<=1535) X17=4.
IF (TALLAMADRE>1453 & TALLAMADRE<=1518) X17=5.
IF (TALLAMADRE<=1453) X17=6.
VALUE LABELS
/X17
```

```
1 '>1633'
2 '<1591 - 1633]'
3 '<1535 - 1591]'
4 '<1518 - 1535]'
5 '<1453 - 1518]'
6 '<=1453'.
VARIABLE LABELS X17 'Talla en centímetros (1 decimal)'.
EXECUTE.

IF X18=3 X18=2.
VALUE LABELS
/X18
1 'Sin anemia'
2 'Moderado, grave o leve'.

NUMERIC X28 (F8.0).
EXECUTE.
IF (PRENATAL>15) X28=1.
IF (PRENATAL>12 & PRENATAL<=15) X28=2.
IF (PRENATAL>9 & PRENATAL<=12) X28=3.
IF (PRENATAL>7 & PRENATAL<=9) X28=4.
IF (PRENATAL>5 & PRENATAL<=7) X28=5.
IF (PRENATAL>3 & PRENATAL<=5) X28=6.
IF (PRENATAL<=3) X28=7.
VALUE LABELS
/X28
1 '>15'
2 '<12 - 15]'
3 '<9 - 12]'
4 '<7 - 9]'
5 '<5 - 7]'
6 '<3 - 5]'
7 '<=3'.
VARIABLE LABELS X28 'Visitas prenatales por embarazo'.
EXECUTE.

NUMERIC X29 (F8.0).
EXECUTE.
IF (PRIMERCPN <= 1) X29=1.
IF (PRIMERCPN>1 & PRIMERCPN<=2) X29=2.
IF (PRIMERCPN>2 & PRIMERCPN<=3) X29=3.
IF (PRIMERCPN>3 & PRIMERCPN<=4) X29=4.
IF (PRIMERCPN>4 & PRIMERCPN<=6) X29=5.
IF (PRIMERCPN>6) X29=6.
VALUE LABELS
/X29
1 '<=1'
2 '<1 - 2]'
3 '<2 - 3]'
4 '<3 - 4]'
```

```
5 '<4 - 6]'
6 '>6'.
VARIABLE LABELS X29 'Momento del primer control prenatal'.
EXECUTE.

NUMERIC X30 (F8.0).
EXECUTE.
IF (PESO <= 1640) X30=1.
IF (PESO>3825) X30=2.
IF (PESO>3113 & PESO<=3825) X30=3.
IF (PESO>1640 & PESO<=3113) X30=4.
VALUE LABELS
/X30
1 '<= 1640'
2 '> 3825'
3 '<3113 - 3825]'
4 '<1640 - 3113]'.
VARIABLE LABELS X30 'Peso del niño al nacer (kilos - 3 dec.)'.
EXECUTE.

NUMERIC X32 (F8.0).
EXECUTE.
IF (CONSUMOHIE>209) X32=1.
IF (CONSUMOHIE>116 & CONSUMOHIE<=209) X32=2.
IF (CONSUMOHIE>44 & CONSUMOHIE<=116) X32=3.
IF (CONSUMOHIE <= 44) X32=4.
VALUE LABELS
/X32
1 '> 209'
2 '<116 - 209]'
3 '<44 - 116]'
4 '<= 44'.
VARIABLE LABELS X32 'Por cuantos días tomó hierro y/o cuantas inyecciones
recibió'.
EXECUTE.

NUMERIC X33 (F8.0).
EXECUTE.
IF (SOLIDOS=0) X33=1.
IF (SOLIDOS > 4) X33=2.
IF (SOLIDOS>3 & SOLIDOS<=4) X33=3.
IF (SOLIDOS>2 & SOLIDOS<=3) X33=4.
IF (SOLIDOS>0 & SOLIDOS<=2) X33=5.
VALUE LABELS
/X33
1 '0'
2 '> 4'
3 '<3 - 4]'
4 '<2 - 3]'.
5 '<0 - 2]'.
```

```
VARIABLE LABELS X33 'El día de ayer o durante el día o la noche cuantas veces le
dio comida sólidas o semisólidas'.
EXECUTE.

IF ANEMIA=1 ANEMIA=0.
IF ANEMIA=2 ANEMIA=1.
VALUE LABELS
/ANEMIA
0 'Sin anemia'
1 'Con anemia'.

FREQ ANEMIA.
EXECUTE.

DELETE VARIABLES EDAD ORDEN INTERVALO ALTITUD MENORES
MIEMBROS EDADMADRE TALLAMADRE PRENATAL PRIMERCPN PESO
CONSUMOHIE SOLIDOS.

SAVE OUTFILE='C:\Users\Home\Downloads\Tesis
UNS\Base_FINAL5_anemia_niños_2015-2019.SAV'
  /COMPRESSED.
```

Anexo F

Código en R sobre los procedimientos alternativos con el algoritmo “Random Forest”

```
rm(list=ls()) ## eliminar archivos guardados en memoria
dev.off()   ##elimina gráficos de memoria
options(scipen=999) ##Evita que los datos salgan con notacion cientifica
##Cambiar el directorio de trabajo
setwd(dirname(rstudioapi::getActiveDocumentContext()$path))
#setwd("C:/Users/Home/Downloads/Tesis UNS")
getwd()
.libPaths("c:/Rlib")  ## Cambia de directorio cuando sale acceso denegado
library(ggvis)
library(party)
library(Boruta)
library(pROC)
library(randomForest)
library(e1071)
library(caret)
library(glmnet)
library(mboost)
library(adabag)
library(xgboost)
library(ROCR)
library(mlr)
library(lattice)
library(gmodels)
library(gplots)
library(DMwR)
library(rminer)
library(polycor)
library(class)
library(neuralnet)

#########################################################################
########### -- TESIS DE DOCTORADO -- ###################################
#########################################################################
####### Autor: Mg. Bernardo Céspedes Panduro ############################
#########################################################################

### -- 0) Direccionar a la base de datos
### -- 1) Librerías a usar ####
#if (!is.null(getOption("sqldf.connection"))) sqldf()
#library(MASS)
#install.packages("sqldf")
library(sqldf) ## librería para usar comandos sql
#install.packages("tidyverse")
library(tidyverse) ## librería para tratamiento de data
library(ggplot2) ## librería para gráficos
#install.packages("mlr")
```

```
library(mlr)
library(foreign)  ##importar la data del spss
library(rpart)
#install.packages("rpart.plot")
library(rpart.plot)  ##permite hacer mejores graficos en arboles
library(rattle)
#install.packages("CHAID")
#require(CHAID)
install.packages("partykit")
library(partykit)
install.packages("CHAID", repos="http://R-Forge.R-project.org")
library(CHAID)
# Carga el paquete específico del Árbol de clasificación C5.0
# Útil para arboles de decisión cuando la variable dependiente es no métrica
# Variables independientes son métricas
install.packages("C50", dependencies = TRUE)
library(C50)

### -- 2) Datos a Utilizar ####

endes<-read.spss("Base_FINAL4_anemia_niños_2015-
2019.SAV",use.value.labels=TRUE,to.data.frame=TRUE)
View(endes)
attr(endes,"variable.labels")<-NULL   #Elimina las etiquetas de las variables
contrasts(endes$ANEMIA)  ##para saber la categoria de referencia
attach(endes)
table(endes$ANEMIA)
prop.table(table(endes$ANEMIA))

#imputación casos perdidos.
install.packages("missForest")
library(missForest)

#cargamos datos
head(endes)
df<-as.data.frame(endes)
View(df)

sapply(df, class)  #Para ver el tipo de variable
#si hubiera que hacer cambios Inv <- lapply(Inv, as.ordered)

colSums(is.na(df)) ##Número de registros faltantes usando la función
colSums(is.na(data))

## Número de registros faltantes usando la función missForest::missForest
#hacemos la imputación
imp<-missForest(df)
imp <- missForest(df, verbose = TRUE, variablewise = FALSE)
imp$OOBerror
imp <- missForest(df, verbose = TRUE, variablewise = TRUE)
```

```
imp$OOBerror
sapply(df, class)

dflimpio<-as.data.frame(imp$ximp)
View(dflimpio)

comparacion<-cbind(df$body_mass_g,df$sex,dflimpio$sex)
View(comparacion)

datos<-as.data.frame(imp$ximp)

## Imputación de datos #####
install.packages("Hmisc")
library(Hmisc)
library(survival)
#impute_arg<-
aregImpute(~ANEMIA+X01+X02+EDAD+ORDEN+INTERVALO+X06+ALTITUD+ME
NORES+MIEMBROS+X10+X11+X12+X13+X14+X15+EDADMADRE+TALLAMADR
E+X18+X19+X20+X21+X22+X23+X24+X25+X26+X27+PRENATAL+PRIMERCPN+
PESO+X31+CONSUMOHIE+SOLIDOS,data=df,match='closest',n.impute=10,nk=4)
#impute_arg<-
aregImpute(~ANEMIA+X01+X02+EDAD+ORDEN+INTERVALO+X06+ALTITUD+MIE
MBROS+X10+X11+X12+X13+X14+X15+EDADMADRE+TALLAMADRE+X18+X19+
X20+X21+X22+X23+X24+X25+X26+X27+PRENATAL+PRIMERCPN+PESO+X31+C
ONSUMOHIE+SOLIDOS,data=df,match='closest',n.impute=10,nk=4)
# Find value of nk that yields best validating imputation models
# tlinear=FALSE means to not force the target variable to be linear
#impute_arg<-
aregImpute(~ANEMIA+X01+X02+EDAD+ORDEN+INTERVALO+X06+ALTITUD+ME
NORES+MIEMBROS+X10+X11+X12+X13+X14+X15+EDADMADRE+TALLAMADR
E+X18+X19+X20+X21+X22+X23+X24+X25+X26+X27+PRENATAL+PRIMERCPN+
PESO+X31+CONSUMOHIE+SOLIDOS,data=df,match='closest',n.impute=10,nk=c(0,
3:5),tlinear=FALSE,B=75)
impute_arg<-
aregImpute(~ANEMIA+X01+X02+EDAD+ORDEN+INTERVALO+X06+ALTITUD+ME
NORES+MIEMBROS+X10+X11+X12+X13+X14+X15+EDADMADRE+TALLAMADR
E+X18+X19+X20+X21+X22+X23+X24+X25+X26+X27+PRENATAL+PRIMERCPN+
PESO+X31+CONSUMOHIE+SOLIDOS,data=df,match='closest',n.impute=10,nk=0,tli
near=FALSE,B=75)
impute_arg
View(impute_arg)
#Creamos un nuevo archivo con todas las variables imputadas
fill_data <- function(impute = impute_arg, data = df, im = 1) {
  cbind.data.frame(impute.transcan(x = impute,
                                   imputation = im,
                                   data = data,
                                   list.out = TRUE,
                                   pr = FALSE))
}
full_dat1 <- fill_data(im = 1)
```

```
full_dat2 <- fill_data(im = 2)
View(full_dat2)

imp=data.frame(full_dat2)
imp
View(imp)
write.table(imp,"imp.txt")
write.csv(x=imp,"imp.csv")
write.foreign(x=imp,"imp.txt","imp.sps",package = "SPSS")

install.packages("writexl")
library("writexl")
write_xlsx(imp,"imp.xlsx")
tempfile(fileext = "imp.xlsx")
#Comprobar la variable imputada X31
impute_arg$imputed$X01

head(impute_arg)

comparacion<-cbind(df$MIEMBROS,df$MENORES,imp$MENORES)
View(comparacion)

#Grabamos el archivo imputado con extensión R
saveRDS(imp, file = "imp.rds")
#Abrimos el archivo con extensión R
imp <- readRDS("C:/Users/Home/Downloads/Tesis UNS/imp.rds")
View(imp)
sapply(imp,class)

datos_R<-data.frame(imp)
View(datos_R)

str(datos_R)
# colocar los tipos de datos a las categóricas

datos_R$ANEMIA <- factor (datos_R$ANEMIA)
datos_R$X01 <- factor (datos_R$X01)
datos_R$X02 <- factor (datos_R$X02)
datos_R$EDAD <- as.integer(datos_R$EDAD)
datos_R$EDAD <- as.numeric(datos_R$EDAD)
datos_R$ORDEN <- as.integer(datos_R$ORDEN)
datos_R$INTERVALO <- as.integer(datos_R$INTERVALO)
datos_R$X06 <- factor (datos_R$X06)
datos_R$ALTITUD <- as.integer(datos_R$ALTITUD)
datos_R$MENORES <- as.integer(datos_R$MENORES)
datos_R$MIEMBROS <- as.integer(datos_R$MIEMBROS)
datos_R$X10 <- factor (datos_R$X10)
datos_R$X11 <- factor (datos_R$X11)
datos_R$X12 <- factor (datos_R$X12)
datos_R$X13 <- factor (datos_R$X13)
```

```
datos_R$X14 <- factor (datos_R$X14)
datos_R$X15 <- factor (datos_R$X15)
datos_R$EDADMADRE <- as.integer(datos_R$EDADMADRE)
datos_R$TALLAMADRE <- as.integer(datos_R$TALLAMADRE)
datos_R$X18 <- factor (datos_R$X18)
datos_R$X19 <- factor (datos_R$X19)
datos_R$X20 <- factor (datos_R$X20)
datos_R$X21 <- factor (datos_R$X21)
datos_R$X22 <- factor (datos_R$X22)
datos_R$X23 <- factor (datos_R$X23)
datos_R$X24 <- factor (datos_R$X24)
datos_R$X25 <- factor (datos_R$X25)
datos_R$X26 <- factor (datos_R$X26)
datos_R$X27 <- factor (datos_R$X27)
datos_R$PRENATAL <- as.integer (datos_R$PRENATAL)
datos_R$PRIMERCPN <- as.integer (datos_R$PRIMERCPN)
datos_R$PESO <- as.integer(datos_R$PESO)
datos_R$X31 <- factor (datos_R$X31)
datos_R$CONSUMOHIE <- as.integer (datos_R$CONSUMOHIE)
datos_R$SOLIDOS <- as.integer (datos_R$SOLIDOS)

datos_R$ANEMIA <- factor(datos_R$ANEMIA)

#levels(datos_R$ANEMIA) <- c("Sin anemia","Con anemia")
datos_R<-data.frame(datos_R)
# estadísticas de las variables principales
estadísticas <- summarizeColumns(datos_R)
estadísticas
sapply(datos_R, class)
colSums(is.na(datos_R))
##Número de registros faltantes usando la función colSums(is.na(data))
##Después de organizar las variables por su tipo se graban en un nuevo archivo
## datos_R.sav y datos_R.xlsx
write.foreign(x=datos_R,"c:/Rlib/datos_R.txt","c:/Rlib/datos_R.sps",package="SPSS")
write_xlsx(datos_R,"c:/Rlib/datos_R.xlsx")

install.packages("haven")
library(haven)
write_sav(datos_R,"c:/Rlib/datos_R.sav")

#Se crean los arboles CHAID
arbol_1 <- chaid(datos_R$ANEMIA~datos_R$X01,data = datos_R)
plot(arbol_1)
summary(arbol_1)
arbol_2 <- chaid(datos_R$ANEMIA ~ datos_R$X02, data = datos_R)
plot(arbol_2)
arbol_3 <- ctree(datos_R$ANEMIA ~ datos_R$EDAD,data = datos_R)
plot(arbol_3) # Gráfico
arbol_4 <- ctree(datos_R$ANEMIA ~ datos_R$ORDEN,data = datos_R)
plot(arbol_4) # Gráfico
```

```
arbol_5 <- ctree(datos_R$ANEMIA ~ datos_R$INTERVALO,data = datos_R)
plot(arbol_5) # Gráfico
arbol_6 <- chaid(datos_R$ANEMIA ~ datos_R$X06, data = datos_R)
plot(arbol_6)
arbol_7 <- ctree(datos_R$ANEMIA ~ datos_R$ALTITUD,data = datos_R)
plot(arbol_7) # Gráfico
arbol_8 <- ctree(datos_R$ANEMIA ~ datos_R$MENORES,data = datos_R)
plot(arbol_8) # Gráfico
arbol_9 <- ctree(datos_R$ANEMIA ~ datos_R$MIEMBROS,data = datos_R)
plot(arbol_9) # Gráfico
arbol_10 <- chaid(datos_R$ANEMIA ~ datos_R$X10, data = datos_R)
plot(arbol_10)
arbol_11 <- chaid(datos_R$ANEMIA ~ datos_R$X11, data = datos_R)
plot(arbol_11)
arbol_12 <- chaid(datos_R$ANEMIA ~ datos_R$X12, data = datos_R)
plot(arbol_12)
arbol_13 <- chaid(datos_R$ANEMIA ~ datos_R$X13, data = datos_R)
plot(arbol_13)
arbol_14 <- chaid(datos_R$ANEMIA ~ datos_R$X14, data = datos_R)
plot(arbol_14)
arbol_15 <- chaid(datos_R$ANEMIA ~ datos_R$X15, data = datos_R)
plot(arbol_15)
arbol_16 <- ctree(datos_R$ANEMIA ~ datos_R$EDADMADRE,data = datos_R)
plot(arbol_16) # Gráfico
arbol_17 <- ctree(datos_R$ANEMIA ~ datos_R$TALLAMADRE,data = datos_R)
plot(arbol_17) # Gráfico
arbol_18 <- chaid(datos_R$ANEMIA ~ datos_R$X18, data = datos_R)
plot(arbol_18)
arbol_19 <- chaid(datos_R$ANEMIA ~ datos_R$X19, data = datos_R)
plot(arbol_19)
arbol_20 <- chaid(datos_R$ANEMIA ~ datos_R$X20, data = datos_R)
plot(arbol_20)
arbol_21 <- chaid(datos_R$ANEMIA ~ datos_R$X21, data = datos_R)
plot(arbol_21)
arbol_22 <- chaid(datos_R$ANEMIA ~ datos_R$X22, data = datos_R)
plot(arbol_22)
arbol_23 <- chaid(datos_R$ANEMIA ~ datos_R$X23, data = datos_R)
plot(arbol_23)
arbol_24 <- chaid(datos_R$ANEMIA ~ datos_R$X24, data = datos_R)
plot(arbol_24)
arbol_25 <- chaid(datos_R$ANEMIA ~ datos_R$X25, data = datos_R)
plot(arbol_25)
arbol_26 <- chaid(datos_R$ANEMIA ~ datos_R$X26, data = datos_R)
plot(arbol_26)
arbol_27 <- chaid(datos_R$ANEMIA ~ datos_R$X27, data = datos_R)
plot(arbol_27)
arbol_28<-ctree(datos_R$ANEMIA ~ datos_R$PRENATAL,data = datos_R)
plot(arbol_28)
arbol_29 <- ctree(datos_R$ANEMIA ~ datos_R$PRIMERCPN,data = datos_R)
plot(arbol_29) # Gráfico
```

```
arbol_30 <- ctree(datos_R$ANEMIA ~ datos_R$PESO,data = datos_R)
plot(arbol_30) # Gráfico
arbol_31 <- chaid(datos_R$ANEMIA ~ datos_R$X31, data = datos_R)
plot(arbol_31)
arbol_32 <- ctree(datos_R$ANEMIA ~ datos_R$CONSUMOHIE,data = datos_R)
plot(arbol_32) # Gráfico
arbol_33 <- ctree(datos_R$ANEMIA ~ datos_R$SOLIDOS,data = datos_R)
plot(arbol_33) # Gráfico

#ANEMIA+X01+X02+EDAD+ORDEN+INTERVALO+X06+ALTITUD+MENORES+
#MIEMBROS+X10+X11+X12+X13+X14+X15+EDADMADRE+TALLAMADRE+X18+
#X19+X20+X21+X22+X23+X24+X25+X26+X27+PRENATAL+PRIMERCPN+PESO+
#X31+CONSUMOHIE+SOLIDOS
## se discretiza las variables mediante arboles de decision Chaid
#La discretización se realiza en el software SPSS.
#Después de la discretización se graban en un archivo con título
#Base_FINAL5_anemia_niños_2015-2019.SAV

##En R llamamos al archivo Base_FINAL5_anemia_niños_2015-2019.SAV

datos_R<-read.spss("Base_FINAL5_anemia_niños_2015-
2019.SAV",use.value.labels=TRUE,to.data.frame=TRUE)
View(datos_R)
attr(datos_R,"variable.labels")<-NULL   #Elimina las etiquetas de las variables
contrasts(datos_R$ANEMIA)  ##para saber la categoria de referencia
attach(datos_R)
table(datos_R$ANEMIA)
prop.table(table(datos_R$ANEMIA))
datos_R$ANEMIA <- factor(datos_R$ANEMIA)
datos_R$X01 <- factor (datos_R$X01)
datos_R$X02 <- factor (datos_R$X02)
datos_R$X03 <- factor (datos_R$X03)
datos_R$X04 <- factor (datos_R$X04)
datos_R$X05 <- factor (datos_R$X05)
datos_R$X06 <- factor (datos_R$X06)
datos_R$X07 <- factor (datos_R$X07)
datos_R$X08 <- factor (datos_R$X08)
datos_R$X09 <- factor (datos_R$X09)
datos_R$X10 <- factor (datos_R$X10)
datos_R$X11 <- factor (datos_R$X11)
datos_R$X12 <- factor (datos_R$X12)
datos_R$X13 <- factor (datos_R$X13)
datos_R$X14 <- factor (datos_R$X14)
datos_R$X15 <- factor (datos_R$X15)
datos_R$X16 <- factor (datos_R$X16)
datos_R$X17 <- factor (datos_R$X17)
datos_R$X18 <- factor (datos_R$X18)
datos_R$X19 <- factor (datos_R$X19)
datos_R$X20 <- factor (datos_R$X20)
datos_R$X21 <- factor (datos_R$X21)
```

```
datos_R$X22 <- factor (datos_R$X22)
datos_R$X23 <- factor (datos_R$X23)
datos_R$X24 <- factor (datos_R$X24)
datos_R$X25 <- factor (datos_R$X25)
datos_R$X26 <- factor (datos_R$X26)
datos_R$X27 <- factor (datos_R$X27)
datos_R$X28 <- factor (datos_R$X28)
datos_R$X29 <- factor (datos_R$X29)
datos_R$X30 <- factor (datos_R$X30)
datos_R$X31 <- factor (datos_R$X31)
datos_R$X32 <- factor (datos_R$X32)
datos_R$X33 <- factor (datos_R$X33)

sapply(datos_R, class)
#write.csv(datos_R,"data_recodificadaf.csv",row.names = F)

input_name <- 'data_recod'

## Categorizacion de las variables ##

datos_R[,1:ncol(datos_R)] <-
lapply(datos_R[,1:ncol(datos_R)],as.factor)

str(datos_R)

contrasts(datos_R$ANEMIA)

library(caret)
particion <- createDataPartition(y = datos_R$ANEMIA, p = 0.70, list = FALSE, times
= 1)
train <- datos_R[particion, ]
test  <- datos_R[-particion, ]
dim(train)
dim(test)

################
#Balanceo de datos
#both sampling
library(ROSE)
data.bal<- ovun.sample(ANEMIA ~ ., data = train, method = "both", p=0.5)$data
table(data.bal$ANEMIA)

########
#Modelo 1
library(randomForest)
set.seed(123)
modelo1 <- randomForest(ANEMIA~., data=train)

modelo1
varImpPlot(modelo1)
```

```
plot(modelo1)
modelo1$importance
predicciones1 <- predict(modelo1, test)

#Calcular la matriz de confusión
table(predicciones1,test$ANEMIA)
#Indicadores
library(caret)
indicadores=confusionMatrix(predicciones1, test$ANEMIA)
indicadores

#curva roc
pred_prob1<-predict(modelo1, test, type = "prob")[,1]
library(ROCR)
predR1 <- prediction(pred_prob1, test$ANEMIA)
predR1.1<-performance(predR1, "tpr", "fpr")
plot(predR1.1, colorize = T)
lines(x=c(0, 1), y=c(0, 1), col=" blue", lwd=1, lty=3);
lines(x=c(1, 0), y=c(0, 1), col="red", lwd=1, lty=4)

#auc
auc1<-performance(predR1, measure = "auc")@y.values[[1]]*100
auc1
#Indice GINI
ROCRN1 <- round(performance(predR1, measure = "auc")@y.values[[1]]*100, 2)
giniRN1 <- (2*ROCRN1 - 100)
giniRN1
##########################################

#Modelo 2
library(randomForest)
set.seed(123)
modelo2 <- randomForest(ANEMIA~., data=data.bal)

modelo2
plot(modelo2)
modelo2$importance
predicciones2 <- predict(modelo2, test)

#Calcular la matriz de confusión
table(predicciones2,test$ANEMIA)
#Indicadores
library(caret)
indicadores2=confusionMatrix(predicciones2, test$ANEMIA)
indicadores2

#curva roc
pred_prob2<-predict(modelo2, test, type = "prob")[,1]
library(ROCR)
predR2 <- prediction(pred_prob2, test$ANEMIA)
```

```
predR2.1<-performance(predR2, "tpr", "fpr")
plot(predR2.1, colorize = T)
lines(x=c(0, 1), y=c(0, 1), col=" blue", lwd=1, lty=3);
lines(x=c(1, 0), y=c(0, 1), col="red", lwd=1, lty=4)

#auc
auc2<-performance(predR2, measure = "auc")@y.values[[1]]*100
auc2
#Indice GINI
ROCRN2 <- round(performance(predR2, measure = "auc")@y.values[[1]]*100, 2)
giniRN2 <- (2*ROCRN2 - 100)
giniRN2
##############################################

#Modelo 3
#Modelo tuneado no balanceado

set.seed(123)
t <- tuneRF(train[,-34], train$ANEMIA,
            stepFactor = 0.5,
            plot = TRUE,
            ntreeTry = 300,
            trace = TRUE,
            improve = 0.01)

best.m <- t[t[, 2] == min(t[, 2]), 1]
print(t)
print(best.m)

#hist(treesize(modelo2),
#main = "Número de nodos para los árboles",
#col = "red")

#######################################
set.seed(123)
modelo3 <-randomForest(ANEMIA~.,data=train,
                       mtry=best.m,
                       importance=TRUE,
                       ntree=300)

modelo3
importance(modelo3)
varImpPlot(modelo3)

predicciones3 <- predict(modelo3, test)

#Calcular la matriz de confusión
table(predicciones3,test$ANEMIA)
#Indicadores
library(caret)
```

```
indicadores3=confusionMatrix(predicciones3, test$ANEMIA)
indicadores3

#curva roc
pred_prob3<-predict(modelo3, test, type = "prob")[,1]
library(ROCR)
predR3 <- prediction(pred_prob3, test$ANEMIA)
predR3.1<-performance(predR3, "tpr", "fpr")
plot(predR3.1, colorize = T)
lines(x=c(0, 1), y=c(0, 1), col=" blue", lwd=1, lty=3);
lines(x=c(1, 0), y=c(0, 1), col="red", lwd=1, lty=4)

#auc
auc3<-performance(predR3, measure = "auc")@y.values[[1]]*100
auc3
#Indice GINI
ROCRN3 <- round(performance(predR3, measure = "auc")@y.values[[1]]*100, 2)
giniRN3 <- (2*ROCRN3 - 100)
giniRN3
#####################################3
#Modelo 4
#Modelo tuneado balanceado
library(randomForest)
set.seed(123)
t_bal <- tuneRF(data.bal[,-34], data.bal$ANEMIA,
                stepFactor = 0.5,
                plot = TRUE,
                ntreeTry = 300,
                trace = TRUE,
                improve = 0.01)

best.m_bal <- t_bal[t_bal[, 2] == min(t_bal[, 2]), 1]
print(t_bal)
print(best.m_bal)

#hist(treesize(modelo2),
#main = "Número de nodos para los árboles",
#col = "red")

#####################################
set.seed(123)
modelo3.bal <-randomForest(ANEMIA~.,data=data.bal,
                           mtry=best.m_bal,
                           importance=TRUE,
                           ntree=300)

modelo3.bal
importance(modelo3.bal)
varImpPlot(modelo3.bal)
```

```
predicciones3.bal <- predict(modelo3.bal, test)
#Calcular la matriz de confusión
table(predicciones3.bal,test$ANEMIA)
#Indicadores
library(caret)
indicadores3=confusionMatrix(predicciones3.bal, test$ANEMIA)
indicadores3

#curva roc
pred_prob3.bal<-predict(modelo3.bal, test, type = "prob")[,1]
library(ROCR)
predR3.bal <- prediction(pred_prob3.bal, test$ANEMIA)
predR3.1.bal<-performance(predR3.bal, "tpr", "fpr")
plot(predR3.1, colorize = T)
lines(x=c(0, 1), y=c(0, 1), col=" blue", lwd=1, lty=3);
lines(x=c(1, 0), y=c(0, 1), col="red", lwd=1, lty=4)

#auc
auc3.bal<-performance(predR3.bal, measure = "auc")@y.values[[1]]*100
auc3.bal
#Indice GINI
ROCRN3.bal <- round(performance(predR3.bal, measure =
"auc")@y.values[[1]]*100, 2)
giniRN3.bal <- (2*ROCRN3 - 100)
giniRN3.bal
################################################
#Modelo 4
#Modelos con selección de variables NO BALANCEADO
modelo4.sin.bal <- randomForest(ANEMIA~.,
                                data=train,
                                importance=T)

modelo4.sin.bal
importance(modelo4.sin.bal)
varImpPlot(modelo4.sin.bal)

predicciones4.sin.bal <- predict(modelo4.sin.bal, test)

#Calcular la matriz de confusión
table(predicciones4.sin.bal,test$ANEMIA)
#Indicadores
library(caret)
indicadores4.sin.bal=confusionMatrix(predicciones4.sin.bal, test$ANEMIA)
indicadores4.sin.bal

#curva roc
pred_prob4.sin.bal<-predict(modelo4.sin.bal, test, type = "prob")[,1]
library(ROCR)
predR4.sin.bal <- prediction(pred_prob4.sin.bal, test$ANEMIA)
predR4.1.sin.bal <-performance(predR4.sin.bal, "tpr", "fpr")
```

```
plot(predR4.sin.bal, colorize = T)
lines(x=c(0, 1), y=c(0, 1), col=" blue", lwd=1, lty=3);
lines(x=c(1, 0), y=c(0, 1), col="red", lwd=1, lty=4)

#auc
auc4.sin.bal<-performance(predR4.sin.bal, measure = "auc")@y.values[[1]]*100
auc4.sin.bal
#Indice GINI
ROCRN4.sin.bal <- round(performance(predR4.sin.bal, measure = "auc")@y.values[[1]]*100, 2)
giniRN4.sin.bal <- (2*ROCRN4.sin.bal - 100)
giniRN4.sin.bal
########################################

#Modelo 4
#Modelos con selección de variables BALANCEADO
modelo4.bal <- randomForest(ANEMIA~.,
                            data=data.bal,
                            importance=T)

modelo4.bal
importance(modelo4.bal)
varImpPlot(modelo4.bal)

predicciones4.bal <- predict(modelo4.bal, test)

#Calcular la matriz de confusión
table(predicciones4.bal,test$ANEMIA)
#Indicadores
library(caret)
indicadores4.bal=confusionMatrix(predicciones4.bal, test$ANEMIA)
indicadores4.bal

#curva roc
pred_prob4.bal<-predict(modelo4.bal, test, type = "prob")[,1]
library(ROCR)
predR4.bal <- prediction(pred_prob4.bal, test$ANEMIA)
predR4.1.bal<-performance(predR4.bal, "tpr", "fpr")
plot(predR4.1.bal, colorize = T)
lines(x=c(0, 1), y=c(0, 1), col=" blue", lwd=1, lty=3);
lines(x=c(1, 0), y=c(0, 1), col="red", lwd=1, lty=4)

#auc
auc4.bal<-performance(predR4.bal, measure = "auc")@y.values[[1]]*100
auc4.bal
#Indice GINI
ROCRN4.bal <- round(performance(predR4.bal, measure = "auc")@y.values[[1]]*100, 2)
giniRN4.bal <- (2*ROCRN4.bal - 100)
giniRN4.bal
```

```
##############################################

# Se crea la base de datos Base_FINAL4_anemia_niños_2015-2019.SAV
#Es la Base de datos sin imputar
#Luego la base de datos imp.rds
#Base de datos imputada en R.
##Después de organizar las variables por su tipo se graban en un nuevo archivo
## datos_R.sav y datos_R.xlsx
## se discretiza las variables mediante arboles de decision Chaid
#La discretización se realiza en el software SPSS.
#Después de la discretización se graban en un archivo con título
#Base_FINAL5_anemia_niños_2015-2019.SAV

##En R llamamos al archivo Base_FINAL5_anemia_niños_2015-2019.SAV
##En esta base de datos se desarrollan todos los modelos
```

APLICACIÓN DEL ALGORITMO "RANDOM FOREST" PARA UN MODELO DE CLASIFICACIÓN SOBRE LA TENENCIA DE ANEMIA DE NIÑOS DEL PERÚ

INFORME DE ORIGINALIDAD

19%	19%	4%	4%
INDICE DE SIMILITUD	FUENTES DE INTERNET	PUBLICACIONES	TRABAJOS DEL ESTUDIANTE

FUENTES PRIMARIAS

1	repositorio.urp.edu.pe Fuente de Internet	6%
2	revistas.um.es Fuente de Internet	2%
3	repositorio.uns.edu.pe Fuente de Internet	2%
4	Submitted to Pontificia Universidad Catolica del Peru Trabajo del estudiante	2%
5	1library.co Fuente de Internet	1%
6	redi.unjbg.edu.pe Fuente de Internet	1%
7	comisiones.ipgh.org Fuente de Internet	1%
8	www.scribd.com Fuente de Internet	1%

9	webinei.inei.gob.pe Fuente de Internet	1 %
10	repositorio.uap.edu.pe Fuente de Internet	1 %
11	rde.inegi.org.mx Fuente de Internet	1 %
12	repositorio.unasam.edu.pe Fuente de Internet	1 %
13	repositorio.lamolina.edu.pe Fuente de Internet	<1 %
14	repositorio.uwiener.edu.pe Fuente de Internet	<1 %