



**ESCUELA SUPERIOR POLITÉCNICA DE CHIMBORAZO**

**FACULTAD DE CIENCIAS**

**CARRERA DE INGENIERÍA EN ESTADÍSTICA INFORMÁTICA**

**“COMPARACIÓN DE MODELOS LOGÍSTICOS Y  
ÁRBOLES DE DECISIÓN PARA IDENTIFICAR Y  
PREDECIR FACTORES ASOCIADOS A LA  
DESNUTRICIÓN CRÓNICA INFANTIL BASADOS EN LA  
ENCUESTA NACIONAL DE SALUD Y NUTRICIÓN –  
ENSANUT 2018-2019”**

**Trabajo de Titulación:**

Tipo: Proyecto de Investigación

Presentado para optar el grado académico de:

**INGENIERA EN ESTADÍSTICA INFORMÁTICA**

**AUTORA: GIORGIA NOHELIA CONGACHA ORTEGA**

**DIRECTOR: Ing. PABLO JAVIER FLORES MUÑOZ**

Riobamba – Ecuador

2020

**©2020, Giorgia Nohelia Congacha Ortega**

Se autoriza la reproducción total o parcial, con fines académicos, por cualquier medio o procedimiento, incluyendo la cita bibliográfica del documento, siempre y cuando se reconozca el Derecho de Autor.

Yo, Giorgia Nohelia Congacha Ortega, declaro que el presente trabajo de titulación es de mi autoría y los resultados del mismo son auténticos. Los textos en el documento que provienen de otras fuentes están debidamente citados y referenciados.

Como autora asumo la responsabilidad legal y académica de los contenidos de este trabajo de titulación; el patrimonio intelectual pertenece a la Escuela Superior Politécnica de Chimborazo.

Riobamba, viernes 11 de diciembre de 2020

**Giorgia Nohelia Congacha Ortega**

**060373438-5**

**ESCUELA SUPERIOR POLITÉCNICA DE CHIMBORAZO**

**FACULTAD DE CIENCIAS**

**CARRERA DE INGENIERÍA EN ESTADÍSTICA INFORMÁTICA**

El Tribunal del trabajo de titulación certifica que: El trabajo de investigación: Tipo: Proyecto de Investigación, **COMPARACIÓN DE MODELOS LOGÍSTICOS Y ÁRBOLES DE DECISIÓN PARA IDENTIFICAR Y PREDECIR FACTORES ASOCIADOS A LA DESNUTRICIÓN CRÓNICA INFANTIL BASADOS EN LA ENCUESTA NACIONAL DE SALUD Y NUTRICIÓN – ENSANUT 2018-2019**, realizado por la señorita: **GIORGIA NOHELIA CONGACHA ORTEGA**, ha sido minuciosamente revisado por los Miembros del Tribunal del trabajo de titulación. El mismo que cumple con los requisitos científicos, técnicos, legales, en tal virtud el Tribunal Autoriza su presentación.

FIRMA

FECHA

Ing. Johanna Enith Aguilar Reyes

\_\_\_\_\_

2020-12-11

**PRESIDENTE DEL TRIBUNAL**

Ing. Pablo Javier Flores Muñoz

\_\_\_\_\_

2020-12-11

**DIRECTOR DE TRABAJO DE  
TITULACIÓN**

Ing. Héctor Salomón Mullo Guaminga

\_\_\_\_\_

2020-12-11

**MIEMBRO DEL TRIBUNAL**

## **DEDICATORIA**

A mis padres Jorge y Miriam que siempre me han apoyado a lo largo de mi vida, han sido el pilar fundamental para continuar desarrollándome como profesional y siempre han sido mi ejemplo de perseverancia y lucha a pesar de los obstáculos que se presenten en la vida.

A mis hermanas Antonella y Sofía, por su amor y apoyo incondicional, por acompañarme cuando más lo he necesitado.

Giorgia

## **AGRADECIMIENTO**

Agradezco a la Escuela Superior Politécnica de Chimborazo por darme la oportunidad de formarme como profesional y poder ser de utilidad a la sociedad con los conocimientos adquiridos.

A mis maestros que con mucha dedicación y paciencia me han formado profesionalmente, sobre todo un agradecimiento especial al Ingeniero Pablo Flores director del trabajo de titulación y al Ingeniero Héctor Mullo quienes con sus conocimientos, tiempo y experiencia me han orientado para culminar con éxito la presente investigación.

A mis padres Jorge y Miriam que a lo largo de toda mi vida me han apoyado y motivado incondicionalmente para nunca rendirme en todo aspecto de mi vida.

A mi abuelita Mamita Vicha por ser mi ejemplo de perseverancia, por confiar en mí y siempre alentarme para seguir adelante.

Giorgia

## TABLA DE CONTENIDO

ÍNDICE DE TABLAS .....	ix
ÍNDICE DE FIGURAS .....	x
ÍNDICE DE GRÁFICOS .....	xi
ÍNDICE DE ANEXOS .....	xii
ÍNDICE DE ABREVIATURAS .....	xiii
RESUMEN .....	xiv
ABSTRACT .....	xv
INTRODUCCIÓN .....	1
CAPÍTULO I	
1 MARCO TEÓRICO REFERENCIAL .....	11
1.1 Conceptos generales de desnutrición .....	11
1.1.1 <i>Desnutrición</i> .....	11
1.1.2 Clasificación de desnutrición infantil .....	12
1.1.3 Factores asociados a la desnutrición infantil .....	13
1.1.4 Encuesta Nacional de Salud y Nutrición - ENSANUT .....	14
1.1.4.1 Estructura de la ENSANUT .....	14
1.1.4.2 Historia de la ENSANUT .....	16
1.1.4.3 Objetivo ENSANUT .....	17
1.1.4.4 Justificación ENSANUT .....	17
1.2 Teoría Estadística .....	17
1.2.1 Análisis exploratorio de datos .....	17
1.2.2 Imputación de datos faltantes .....	17
1.2.2.1 Técnicas para el tratamiento de valores faltantes .....	18
1.2.3 <i>Balanceo de clases</i> .....	19
1.2.4 <i>Regresión logística</i> .....	19
1.2.4.1 Modelo logístico binario .....	20
1.2.4.2 Estimación de los parámetros del modelo .....	21

1.2.4.3	Contrastes o pruebas de significancia del modelo .....	22
1.2.4.	Pseudo estadísticas $R^2$ .....	22
1.2.4.5	Evaluación de la bondad de ajuste del modelo .....	23
1.2.4.6	Ventajas y desventajas del modelo de regresión logística .....	24
1.2.5	<i>Machine Learning</i> .....	24
1.2.6	Árboles de decisión individuales.....	25
1.2.6.1	Definición.....	25
1.2.6.2	Estructura básica de un árbol de decisión .....	25
1.2.6.3	Generación de un árbol de decisión .....	26
1.2.6.4	Ventajas e inconvenientes de los árboles de decisión .....	27
1.2.6.5	Técnicas de ensemble .....	28
1.2.6.6	Overfitting y Underfitting.....	30
1.2.6.7	Métricas de error en problemas de clasificación.....	30
 CAPITULO II		
2	MARCO METODOLÓGICO .....	34
2.1	Tipo y Diseño de investigación .....	34
2.1.1	<i>Localización del estudio</i> .....	34
2.1.2	<i>Población de estudio</i> .....	34
2.1.3	<i>Tamaño de la muestra</i> .....	34
2.1.4	<i>Método de muestreo</i> .....	35
2.1.5	<i>Recolección de información</i> .....	35
2.2	Variables en estudio .....	35
2.2.1	Operacionalización de variables.....	35
2.3	Análisis estadístico .....	42
2.3.1	Instrumentos de procesamiento y análisis de información .....	42
2.3.2	Análisis exploratorio de datos.....	42
2.3.3	<i>Preprocesado de datos</i> .....	43
2.3.3.1	Análisis de datos faltantes .....	43
2.3.3.2	Centrado y escalado de variables numéricas.....	44



2.3.3.3	Variables Dummies .....	44
2.3.3.4	Variables con varianza cero .....	44
2.3.4	Construcción de modelos de clasificación .....	46
2.3.4.1	Árboles de decisión .....	46
2.3.4.2	Regresión logística binaria .....	47
2.3.4.3	Evaluación de modelos .....	47
2.3.5	Construcción de la aplicación web interactiva .....	47
<b>CAPÍTULO III</b>		
3	MARCO DE RESULTADOS Y DISCUSIÓN DE LOS RESULTADOS .....	49
3.1	Análisis exploratorio univariado .....	49
3.2	Modelos de Clasificación: Regresión logística .....	56
3.2.1	<i>Significatividad del Modelo</i> .....	58
3.2.2	<i>Matriz de confusión</i> .....	58
3.3	Modelo de clasificación: Árboles de decisión .....	59
3.3.1	<i>Matriz de confusión</i> .....	62
3.3.2	<i>Modelo Gradient Boosting</i> .....	62
3.4	Comparativa de los modelos: Regresión logística, Árboles de decisión y Gradient Boosting .....	63
CONCLUSIONES .....		66
RECOMENDACIONES .....		67
BIBLIOGRAFIA		
ANEXOS		

## ÍNDICE DE TABLAS

<b>Tabla 1-1:</b> Clasificación del estado nutricional basado en z-scores .....	12
<b>Tabla 1-2:</b> Descripción de variables cuantitativas .....	35
<b>Tabla 2-2:</b> Descripción de variables cualitativas .....	36
<b>Tabla 1-3:</b> Análisis univariado variable dependiente.....	49
<b>Tabla 2-3:</b> Análisis univariado variables independientes cualitativas (Factores Básicos) .....	49
<b>Tabla 3-3:</b> Análisis univariado variables independientes cuantitativas (Factores Básicos) .....	51
<b>Tabla 4-3:</b> Análisis univariado variables independientes cualitativas (Factores Subyacentes) ..	52
<b>Tabla 5-3:</b> Análisis univariado variables independientes cuantitativas (Factores Subyacentes) ..	53
<b>Tabla 6-3:</b> Análisis univariado variables independientes cualitativas (Factores Inmediatos) ....	54
<b>Tabla 7-3:</b> Análisis univariado variables independientes cuantitativas (Factores Inmediatos) ..	55
<b>Tabla 8-3:</b> Modelo de regresión logística binaria .....	56
<b>Tabla 9-3:</b> Prueba ómnibus de la significancia del modelo .....	58
<b>Tabla 10-4:</b> Matriz de confusión regresión logística .....	59
<b>Tabla 11-3:</b> Matriz de confusión árbol de decisión .....	62
<b>Tabla 12-3:</b> Matriz de confusión Gradient Boosting .....	63
<b>Tabla 13-3:</b> Áreas bajo la curva (AUC) .....	65

## ÍNDICE DE FIGURAS

<b>Figura 1-1:</b> Ciclo de la malnutrición .....	11
<b>Figura 2-1:</b> Clasificación de los posibles factores asociados a la desnutrición crónica infantil, de acuerdo al marco teórico de la UNICEF .....	14
<b>Figura 3-1:</b> Estructura básica de un árbol de decisión .....	26
<b>Figura 4-1:</b> Matriz de confusión.....	31
<b>Figura 5-1:</b> Esquema Curva ROC .....	32
<b>Figura 1-3:</b> Comparación de factores significativos asociados a la desnutrición crónica infantil a través de los modelos de árboles de decisión y regresión logística.....	65

## ÍNDICE DE GRÁFICOS

<b>Gráfico 1-3:</b> Árbol de clasificación .....	59
<b>Gráfico 2-3:</b> Curva ROC: Árbol de clasificación .....	63
<b>Gráfico 3-3:</b> Curva ROC: Gradient Boosting .....	64
<b>Gráfico 4-3:</b> Curva ROC: Regresión logística.....	64

## ÍNDICE DE ANEXOS

- ANEXO A:** Código en R utilizado para selección y recodificación de variables de acuerdo a la UNICEF.
- ANEXO B:** Código en R para la unión de base de datos
- ANEXO C:** Código en R para el análisis exploratorio de datos (AED)
- ANEXO D:** Código en R para preprocesado de datos
- ANEXO E:** Código en R utilizado para construir y evaluar el modelo de árboles de decisión
- ANEXO F:** Código en R utilizado para construir y evaluar el modelo Gradient Boosting
- ANEXO G:** Código en R utilizado para construir y evaluar el modelo de Regresión logística
- ANEXO H:** Código en R utilizado para construir la aplicación web, pasos detallados para su publicación e Interfaz de usuario de la aplicación web interactiva

## ÍNDICE DE ABREVIATURAS

<b>INEC:</b>	Instituto Nacional de Estadística y Censos
<b>ENSANUT:</b>	Encuesta Nacional de Salud y Nutrición
<b>UNICEF:</b>	Fondo de las Naciones Unidas para la Infancia
<b>BID:</b>	Banco Interamericano de Desarrollo
<b>DANS:</b>	Encuesta Nacional sobre la Situación Alimentaria, Nutricional y de Salud de la Población de Niños Ecuatorianos menores de Cinco Años.
<b>CEPAR:</b>	Centro de Estudios de Población y Desarrollo Social
<b>ENDEMAIN:</b>	Encuesta Demográfica y de Salud Materna e Infantil
<b>MEF:</b>	Mujeres en edad fértil
<b>OMS:</b>	Organización Mundial de la Salud
<b>UPM:</b>	Unidad primaria de Muestreo
<b>PANN:</b>	Programa Nacional de Alimentación y Nutrición
<b>PND:</b>	Plan Nacional de Desarrollo
<b>DIF:</b>	Desarrollo Integral de la Familia
<b>GLM:</b>	Modelos Lineales Generalizados
<b>RL:</b>	Regresión Logística
<b>AIC:</b>	Criterio de Información Akaike
<b>ROC:</b>	Receiver Operating Characteristic
<b>AUC:</b>	Área bajo la curva ROC
<b>kNN:</b>	k Nearest Neighbours

## RESUMEN

El presente trabajo de titulación compara dos modelos de clasificación: Regresión logística y árboles de decisión con el fin de establecer el mejor modelo que determine los factores significativamente asociados a la desnutrición crónica en niños menores de cinco años y realice predicciones sobre esta variable. La información necesaria fue obtenida del repositorio del Instituto Nacional de Estadísticas y Censos (INEC) a través de la Encuesta Nacional de Salud y Nutrición (ENSANUT 2018). El estudio consideró una muestra de 11.231 niños menores de cinco años, del cual se utilizó para el entrenamiento de los modelos el 70% de la muestra y para validación el 30%. El rendimiento de los modelos planteados fue medido por dos métodos de bondad de ajuste: Tasa de error y Curva ROC en base al poder de predicción de los modelos obtenidos. Del análisis estadístico se extrajeron los siguientes resultados: el modelo de Regresión logística fue el mejor por tener mayor poder predictivo con un AUC = 62.19%, y una menor tasa de error 35%, además los factores asociados a la desnutrición crónica considerando los dos modelos planteados fueron: grupo étnico (indígena), escolaridad de la madre (Educación Media Bachillerato/Superior), algún miembro del hogar tiene teléfono celular(no), meses de embarazo cuando se hizo el primer control, con respecto a otros bebés el tamaño de su hijo era (más grande), grupo de edad (19-23). Se recomienda socializar con entes gubernamentales los resultados obtenidos en esta investigación puesto que ayudarán a la toma de decisiones en planes de contingencia que den solución a la desnutrición crónica infantil en el Ecuador.

**Palabras clave:** <ESTADÍSTICA>, <REGRESIÓN LOGÍSTICA>, <ÁRBOLES DE DECISIÓN>, <TASA DE ERROR>, <CURVA ROC>, <DESNUTRICIÓN CRÓNICA INFANTIL>.



Firmado electrónicamente por:  
**LUIS ALBERTO  
CAMINOS  
VARGAS**



0539-DBRAI-UPT-2021

## **ABSTRACT**

The present degree work compares two classification models: Logistic regression and decision trees in order to establish the best model that determines the factors significantly associated with chronic malnutrition in children under the age of five and makes predictions about this variable. The necessary information was obtained from the repository of the National Institute of Statistics and Censuses (INEC) through the National Health and Nutrition Survey (ENSANUT 2018). The study considered a sample of 11,231 children under five years of age, of which 70% of the sample was used for model training and 30% for validation. The performance of the proposed models was measured by two goodness-of-fit methods: Error Rate and ROC Curve based on the predictive power of the obtained models. The following results were extracted from the statistical analysis: the Logistic regression model was best for having the highest predictive power with an AUC of 62.19%, and a lower error rate of 35%, in addition the factors associated with chronic malnutrition considering the two models proposed were: Ethnic group (indigenous), mother's schooling (High School/University), at least one household member has a cell phone (no), months of pregnancy when the first check was done, the size of the child compared to that of other babies (bigger), age group (19-23). It is recommended that the results obtained in this research be socialized with government agencies as they will assist in decision-making in contingency plans to solve chronic child malnutrition in Ecuador.

**Keywords:** <STATISTICS>, <LOGISTICS REGRESSION>, <DECISION TREES>, <ERROR RATE>, <ROC CURVE>, <CHRONIC CHILD MALNUTRITION>.



## INTRODUCCIÓN

La presente investigación busca analizar la información sobre desnutrición crónica en niños de 0 a 5 años de edad, con el fin de encontrar los factores influyentes a esta problemática social y un modelo aceptable para su predicción. La importancia de enfocarse en los indicadores para niños menores de 5 años, radica en la vulnerabilidad de este grupo etario. A nivel individual la desnutrición está asociada fuertemente con la mortalidad infantil, así como con problemas de desarrollo físico y cognitivo, los que tienen impacto en el rendimiento escolar y posteriormente en la capacidad de trabajo (Paredes, 2016: p.9).

La principal fuente de información para la realización de la presente investigación es la Encuesta Nacional de Salud y Nutrición - ENSANUT 2018, fue levantada por el Ministerio de Salud Pública (MSP) en coordinación con el Instituto Nacional de Estadísticas y Censos (INEC), con el objetivo de describir la situación de la población ecuatoriana en temas de salud y poder formular políticas públicas eficaces.

La ENSANUT divide a la población en estudio en función de 5 formularios para analizar las siguientes temáticas: Hogar; Mujeres en edad fértil de 12 a 49 años de edad; Salud sexual y reproductiva, hombres de 12 años y más; Factores de riesgo, niños y niñas de 5 a menores de 18 años de edad; Desarrollo infantil para niños y niñas menores de 5 años a nivel Nacional. La información recogida a través de los formularios se divide en 9 bases de datos (personas, hogar, etiquetado, MEF, lactancia, salud\_niñez, ssr\_hombres, fact\_riesgo, desarrollo\_infantil), según la población objetivo a la que se dirige, cada base de datos tiene su propio factor de expansión.

La ENSANUT 2018 empleó un periodo de duración de 965 días (05/07/2016 – 21/02/2020) para la planificación, diseño, construcción, recolección, capacitación, procesamiento, análisis, difusión y evaluación de la misma. La periodicidad de esta operación estadística es quinquenal, basa la metodología de sus principales indicadores en las normas, conceptos y procedimientos establecidos por la Organización Mundial de la Salud-OMS y UNICEF, estándares utilizados internacionalmente que permiten tener una alta calidad de contenido para analizar las problemáticas sociales planteadas, no solo a nivel Nacional (Ministerio de Salud Pública, Secretaría Técnica del Plan Toda una Vida, Ministerio de Inclusión Económica y Social), sino a nivel internacional (Organismos internacionales, ONG) (INEC, 2019: pp.23-65).

La naturaleza de la encuesta permite extrapolar los datos a nivel nacional, subregional, por zonas de planificación, por condición social, por rangos de edad, por etnia y por sexo, lo que facilita el análisis focalizado en los niños menores de 5 años.

La disponibilidad de los datos captados por la ENSANUT 2018, más los avances tecnológicos para el procesamiento de grandes volúmenes de datos, se han convertido en oportunidades para usar estos datos con diversos propósitos en la producción estadística; sin embargo, el INEC efectúa únicamente un análisis descriptivo, a través de porcentajes que muestran los principales hallazgos encontrados en la aplicación de la encuesta.

Si bien las estadísticas descriptivas pueden mostrar señales de los principales problemas que afrontan las familias con niños que presentan desnutrición, desde el enfoque de utilización, no permiten identificar en qué medida las características analizadas influyen y si son relevantes dentro del problema. Para resolver esto, en este trabajo de investigación se procede a realizar modelos de clasificación (regresión logística y árboles de decisión), mediante los cuales se han identificado las principales características que estarían influyendo, los mismos que ayudaran a tomar decisiones en las políticas gubernamentales en el campo de la salud, en particular a la desnutrición crónica infantil.

Este trabajo está dividido en tres capítulos los mismos que se describen a continuación:

En el **Capítulo I** se desarrolla la base teórica necesaria para la presente investigación, en la cual se presenta información sobre la desnutrición crónica, la ENSANUT, se describe y conceptualizan los métodos y técnicas estadísticas utilizadas.

En el **Capítulo II** se plantea la parte metodológica de la investigación, el tipo y diseño de investigación llevada a cabo, localización del estudio, población, muestra, método de recolección de información, operacionalización de variables, instrumentos de procesamiento y análisis de la información, descripción del análisis exploratorio de datos y preprocesado de datos. Se detalla la construcción de los modelos de clasificación: Regresión logística y Árboles de decisión y la elaboración de una aplicación web interactiva.

En el **Capítulo III** se exponen los resultados que sustentan la investigación, mediante tablas se presentan las principales estadísticas descriptivas de las variables significativas en los modelos desarrollados, se exhiben los modelos obtenidos a través de las técnicas de regresión logística y árboles de decisión, finalmente a través de un cuadro comparativo se detallan las variables significativas obtenidas de los modelos planteados. Para la selección del mejor modelo predictivo, los modelos fueron evaluados por dos técnicas de bondad de ajuste: tasa de error y Curva ROC.

Finalmente, en la última sección se presentan las conclusiones y recomendaciones extraídas de la investigación realizada, se presenta la bibliografía utilizada para la investigación y en el apartado de Anexos se presentan los códigos realizados para el desarrollo de los modelos implementados.

## **Antecedentes**

### ***Antecedentes metodológicos***

Existen diversos estudios comparativos sobre modelos de clasificación (predicen una variable categórica en función de otras variables). En el siguiente apartado se mencionan algunas de estas investigaciones por su similaridad de la información con la que se trabajó (variables del mismo tipo, variables de respuesta dicotómica) y las técnicas comparativas utilizadas (regresión logística, árboles de decisión):

En el artículo titulado “Árboles de clasificación vs regresión logística en el desarrollo de competencias genéricas en ingeniería” desde un enfoque experimental compara el desempeño de los dos modelos en el contexto de competencias genéricas en ingeniería (razonamiento cuantitativo y comprensión lectora), para ello incorporan dos escenarios de predictores por separado (indicadores y constructos) con la finalidad de mejorar el desempeño de estos, concluyendo que para el uso exclusivo de constructos bajo regresión logística, ni ANOVA ni curvas de ROC con AUC reflejan diferencias en el desempeño de los métodos, sin embargo en árboles de clasificación se notó un detrimento en la curva ROC con AUC cuando se utilizó constructos en vez de indicadores. Cabe resaltar la capacidad de interpretación al usar constructos, pues facilitó la búsqueda de patrones de condiciones del estudiante que pueden verse como limitadores para el desarrollo de competencias. En general ambos métodos presentan similar desempeño en el escenario de solo indicadores, pero no en el escenario de constructos en el cual se mostró mejor desempeño en Regresión logística (Pérez y Echavarría, 2018: pp.1519-1541).

En la tesis titulada “Comparación de modelos de clasificación: Regresión logística y árboles de clasificación para evaluar el rendimiento académico” compara dos modelos de clasificación: regresión logística Binaria y árboles de clasificación CHAID para evaluar el rendimiento académico en estudiantes universitarios de primer semestre matriculados en el curso de Matemática. Para analizar el comportamiento de estos dos modelos utiliza cuatro indicadores: Sensibilidad, Curva ROC, Índice de GINI e Índice Kappa en base al poder de clasificación y predicción de los modelos obtenidos sobre la problemática planteada, se concluye que el mejor modelo por tener mayor poder de clasificación y predicción fue el de árboles de clasificación (Lizares, 2017: p.53).

En la tesis “Comparación de árboles de regresión y clasificación y regresión logística” se presenta la comparación mediante simulación Monte Carlo de dos técnicas estadísticas de clasificación: Árboles de Regresión y Clasificación (CART) y Regresión Logística, el comportamiento de las técnicas fue medido a través de la Tasa de Mala Clasificación (TMC). En general, se observó que cuando se tiene igual separación entre los grupos el TMC de los árboles de clasificación y la

Regresión logística tienen una mínima variación al incrementar la correlación entre las variables explicativas, sin embargo al aumentar la separación, la regresión logística presenta problemas de separación completa al no converger el algoritmo de estimación, mientras que los árboles de clasificación presentan una clasificación perfecta, estas comparaciones llevan a la conclusión de que el mejor modelo por tener mayor precisión son los árboles de clasificación (Serna, 2009: p.38).

En la investigación titulada “Técnicas de ML en medicina cardiovascular”, compara técnicas avanzadas de Machine Learning (ML) como árboles de decisión, máquinas de soporte vectorial y regresión logística en un conjunto de datos relacionados con enfermedades cardiovasculares, después de validar las técnicas mencionadas concluyó que la regresión logística ofrece mayores niveles de precisión que las técnicas de máquinas de soporte vectorial y árboles de decisión (De la Hoz Manotas, et.al; 2013: p.45).

Otro artículo muy similar al anterior titulado: “Árboles de decisión en el diagnóstico de enfermedades cardiovasculares” presenta una descripción de los árboles de decisión y del algoritmo ID3 (Inducción decision tree) para determinar si se debe o no aplicar fármacos a pacientes con enfermedades cardiovasculares usando las variables presión arterial, índice de colesterol, azúcar en la sangre, alergias a antibióticos y otras alergias, mediante la utilización de árboles de decisión con el algoritmo ID3, se concluye que los valores generados son muy similares utilizando las dos técnicas mencionadas, en ambos métodos la variable que tiene mayor ganancia es presión arterial, siendo este el nodo raíz de las dos técnicas, cabe recalcar que la técnica de árbol de decisión conjuntamente con el algoritmo ID3 entrega un conjunto de reglas entendibles que le permiten al médico o al tomador de decisión hacerlo de manera rápida (Solarte, 2011: pp.104-109).

En la investigación titulada “*A Decision Support Tool for Predicting Patients at Risk of Readmission: A Comparison of Classification Trees, Logistic Regression, Generalized Additive Models, and Multivariate Adaptive Regression Splines*” compara la precisión predictiva de la regresión logística, árboles de regresión, modelos aditivos generalizados (GAM) y Splines de regresión adaptativa multivariante (MARS) para predecir los reingresos de emergencia dentro de los 45 días posteriores al alta del hospital, la precisión predictiva se evaluó a través de las Curvas de ROC en la muestra de validación, concluyendo que los modelos de regresión logística y árboles de decisión tuvieron un rendimiento comparable al de los modelos más flexibles basados en GAM y MARS (Demir, 2014: p.875).

El artículo denominado “Análisis comparativo de los algoritmos de árboles de decisión en el procesamiento de datos biológicos” evalúan el desempeño de varios algoritmos de árboles de decisión para así encontrar por medio de comparaciones, cuales son más efectivos en el análisis de datos biológicos. Con esta comparación se determinó la pertinencia de los árboles de decisión,

concluyendo que al momento de analizar datos con variables cualitativas en el estudio de audiología los árboles de clasificación fueron los más competentes para estos datos, mientras que al trabajar sobre la base de antígeno prostático específico el cual manejaba variables cuantitativas los árboles de regresión fueron los más óptimos (Charris et al., 2018: pp.26-34).

### ***Antecedentes aplicativos***

La desnutrición infantil es una problemática social que cobra la vida de más de 6800 niños al día, por esta razón varios investigadores se han visto en la necesidad de profundizar en el estudio de la identificación de factores determinantes para la desnutrición especialmente en niños menores de 5 años ya que a esta edad el ser humano empieza a desarrollar sus facultades cognitivas e inmunológicas (UNICEF, 2013: p.9). A continuación, se detallan algunas de estas investigaciones:

En la tesis titulada “Determinantes de la desnutrición crónica de menores de 5 años y análisis del consumo alimenticio de los hogares del cantón San Miguel de Urququi” se identificaron factores básicos y subyacentes asociados a la desnutrición a través de un análisis de Regresión logística utilizando el marco referencial de la UNICEF y la Encuesta de Condiciones de Vida de Yachay-2013, se estudiaron patrones de consumo de alimentos y composición nutricional de los hogares que tienen y no tienen desnutrición. Los factores determinantes identificados luego de aplicar el modelo de regresión logística fueron: escolaridad de la madre, peso corporal de la madre, lactancia durante los 6 primeros meses, el padre y la madre deciden en conjunto la cantidad horaria que debe trabajar en jefe de hogar, el hogar posee red de alcantarillado y pozo séptico, lugar de residencia, medidas antropométricas, el niño cumple con todas las vacunas, así como todas las variables subyacentes establecidas en el marco referencial de la UNICEF (Paz, 2017: p.68).

El artículo titulado “Determinantes de la desnutrición crónica de los menores de tres años en las regiones del Perú: subanálisis de la encuesta ENDES 2000” identifican los determinantes más importantes de la desnutrición crónica en niños menores de tres años de edad de las diferentes regiones del Perú, según la información recopilada en la Encuesta de Demográfica y de Salud Familiar (ENDES 2000) del Instituto Nacional de Estadística e Informática (INEI). Se seleccionaron las variables significativas que entraron en el modelo multivariado (análisis bivariados, con una significancia de 10%). Se concluye que la desnutrición crónica en niños menores de tres años tiene particularidades en cada región, por lo que es fundamental plantear la resolución de este problema nutricional de diferentes maneras, sin embargo las más importantes en todos los ámbitos fueron: educación de la madre, controles de crecimiento de los niños, número de controles pre- natales, lugar del parto, peso al nacer del niño y el número de hijos vivos de la madre (Bullón y Astete, 2016: p.10).

Otro estudio muy similar al anterior titulado “Prevalencia y determinantes sociales de malnutrición en menores de 5 años afiliados al Sistema de Selección de Beneficiarios para Programas Sociales (SISBEN) del área urbana del municipio de Palermo en Colombia, 2017” identifica la prevalencia de malnutrición y su asociación con DSS, en menores de 5 años, para ello realizan un estudio de corte transversal con enfoque analítico. Se concluye que coexisten los dos extremos de malnutrición: por exceso y por defecto, ésta última, relacionada con determinantes: hacinamiento, bajo ingreso económico familiar y disposición inadecuada de basuras (Barrera et al., 2018: p.244).

El artículo titulado “Mortalidad por desnutrición en menores de cinco años. Pobreza y desarrollos regionales. Colombia. 2003-2012”, analiza uno de los problemas más relevantes en el contexto mundial: la mortalidad por desnutrición. En Colombia, esta situación se concentra en los menores de cinco años y los mayores de 65. La investigación evidenció la relación entre la mortalidad por desnutrición y las desigualdades socioeconómicas y territoriales, expresadas en los indicadores de pobreza, uso de la tierra y diferencias productivas. Dicha relación se manifiesta en el comportamiento diferencial de las tasas y en la configuración de clúster de municipios con altas y bajas tasas de mortalidad por desnutrición, especialmente en los menores de cinco años (Ruiz, 2018: p.66).

El artículo “Influencia de la dinámica familiar y otros factores asociados al déficit en el estado nutrición de preescolares en guarderías del sistema Desarrollo Integral de la Familia (DIF) Jalisco” identifica la influencia de factores asociados al estado nutricional de los preescolares y concluye en que la disfunción de la dinámica familiar, menor ingreso económico, y mayor número de miembros fueron factores de riesgo del estado nutricional. Menor ingreso familiar y escolaridad del padre influyeron en la percepción de ambos padres sobre su dinámica familiar (Ceballos et al., 2005: p.113).

El artículo titulado: “La desnutrición infantil y su relación con los pisos ecológicos en Vinto, Cochabamba, Bolivia” analiza la relación del estado nutricional de los niños menores de 5 años con los diferentes pisos ecológicos en el municipio de Vinto, para llevar a cabo esta investigación se realiza una evaluación antropométrica de niños menores de 5 años, de acuerdo con los indicadores estandarizados por la OMS. Se concluye que la prevalencia de desnutrición varía en relación al piso ecológico en el que habitan los niños, pero por si sola, no es un factor que defina la misma, depende de otros factores asociados. La desnutrición crónica es directamente proporcional a la altura de la población de origen del niño o niña. La residencia en la zona alta representa una mayor prevalencia de desnutrición crónica en este grupo de riesgo ( Mamani et al., 2012: p.18).

En la investigación titulada: “Factores asociados a la desnutrición crónica infantil en el Perú: una aplicación de modelos multinivel”, se identifica y establece la relación entre los factores básicos, subyacentes e inmediatos asociados a la desnutrición infantil de los menores entre seis y treinta y cinco meses de edad, basándose en el marco conceptual propuesto por el Fondo de las Naciones Unidas para la infancia (UNICEF), el cual postula que la desnutrición infantil es consecuencia de tres conjuntos de causas: inmediatas, subyacentes y básicas. Metodológicamente, se estimó porcentajes de desnutrición crónica infantil, luego realizó un análisis bivariado para identificar la asociación entre la variable dependiente: desnutrición crónica infantil y cada una de las variables independientes, concluyendo que las variables consideradas en el marco conceptual que representan a factores básicos, subyacentes e inmediatos se encuentran estadísticamente asociadas a la desnutrición crónica infantil, excepto la variable tos, la cual no resulta significativa (Canazas, 2010: p.53).

## **Planteamiento del problema**

### ***Enunciado del problema***

El uso de modelos es un proceso inherente en el ser humano, que lo ha implementado para entender cómo funcionan los fenómenos e incluso realizar predicciones sobre ellos a partir de observaciones detalladas de los eventos. Sin embargo, los modelos existentes son modelos teóricos ideales (parámetros que se supone que existen, pero nadie los conoce), todos ellos acarrear errores (que incluso en ocasiones podrían ser imperceptibles) y la falta de comparación entre dos puntos de vista diferentes (matemático vs algorítmico) no permiten que se aprovechen las ventajas que cada uno de ellos podría ofrecer.

Ecuador, al igual que el resto de países de Latinoamérica durante los últimos años ha implementado varios programas de alimentación, nutrición y asistencia alimentaria orientados a grupos específicos y vulnerables (escolares, menores de cinco años, mujeres embarazadas, madres en periodo de lactancia) con el propósito de prevenir, atender y mejorar la situación nutricional de la población infantil (Carranza, 2011: p.9). A pesar de estos esfuerzos, el informe del Instituto Nacional de Estadísticas y Censos (INEC), en el periodo 2006-2014 muestra que la tasa de desnutrición crónica infantil a disminuido de un 27% en 2006 a un 23.9% en 2014, sin embargo, la disminución en estos índices no corresponde a los esperados según el Plan Nacional de Desarrollo del Buen Vivir, que se propuso la meta de reducir a un 14 % en el 2013.

Al ser la desnutrición crónica infantil un fenómeno multifactorial originado por un conjunto de elementos que actúan en diferentes niveles de relación, para prevenirla y atenderla es necesario identificar cuáles son los factores que forman estos perfiles. Por ello, se consideran las técnicas de clasificación: regresión logística y árboles de decisión que nos permitan identificar

acertadamente los factores asociados a la desnutrición crónica infantil, en este sentido el primer reto es encontrar la técnica más adecuada con la finalidad de obtener un modelo con un alto poder predictivo.

### ***Formulación del problema***

Pregunta general:

¿Qué modelo (regresión logística o árboles de decisión), permite predecir adecuadamente la desnutrición crónica infantil y determina los factores significativos que la provocan?

### **Justificación**

#### ***Justificación teórica***

La gran cantidad de datos que en la actualidad se producen y la capacidad para procesarlos de forma rápida y precisa por medio de tecnologías de Big Data, han hecho que el aprendizaje automático tome gran importancia especialmente para resolver tres tipos de problemas: variedad, escalabilidad y velocidad. Estas herramientas son útiles en sectores donde se genera gran cantidad de información, sobre todo donde existan actividades o comportamientos repetitivos que puedan optimizarse a través de técnicas algorítmicas (Lazcano, 2019).

Los árboles de decisión son uno de los algoritmos más utilizados en minería de datos, principalmente para problemas de clasificación supervisada, estos realizan diagramas de construcción lógicas, que sirven para representar y categorizar una serie de condiciones que ocurren de forma sucesiva. Sus principales ventajas son: fácil interpretación, rapidez, visualización, robustez a valores perdidos y excelente para aprender relaciones complejas, altamente no lineales. Se suelen utilizar en múltiples campos, principalmente en medicina y economía (Ravina et al., 2018: p.32).

Por otra parte, tomando en cuenta un enfoque diferente, una técnica de estadística clásica muy aplicada en el análisis de datos clínicos y epidemiológicos, son los modelos de regresión logística, en los últimos años se ha verificado una presencia muy marcada de esta técnica, tanto en literatura orientada a tratar temas metodológicos como en los artículos científicos biomédicos, según la *New England Journal of Medicine* la regresión logística es el método multivariante más utilizado en la investigación sanitaria ocupando el quinto puesto, solo superada por cuatro técnicas convencionales: t de student, prueba Chi cuadrado, análisis de la varianza y prueba de Fisher (Fiuza y Rodríguez, 2000: pp.524-530). Varios estudios realizados, han determinado que el modelo de regresión logística es un análisis predictivo muy eficaz cuando la variable dependiente es binaria, comúnmente utilizada para explicar la relación entre una variable binaria dependiente y una o



más variables independientes sin restricción a su escala de medida estas pueden ser cualitativas (nominales u ordinales) o cuantitativas (intervalo o razón) (aprendeIA, 2020).

Teniendo presente la utilidad de las técnicas antes mencionadas, en esta investigación se propone compararlas con el fin de establecer el modelo que mejor prediga la desnutrición crónica en niños menores de cinco años y determine los factores asociados que la provocan.

### ***Justificación práctica***

El Ecuador, a través de la Encuesta Nacional de Salud y Nutrición-ENSANUT 2012 obtuvo la prevalencia de desnutrición crónica infantil, la cual para ese año se observó que afectaba alrededor de 3 de cada 10 niños (25.3%), posicionándose como el tipo de desnutrición que más padece la población infantil, habiendo alrededor de 400.000 niños en el país con esta afectación. Si bien el país se encuentra por debajo de las medias mundiales (FAO, 2018: p. 61), el problema de desnutrición crónica representa un reto de política pública que requiere respuesta urgente, no solo por el bienestar social y el cumplimiento de los niños, sino por la evidencia que demuestra que la desnutrición es un obstáculo para el desarrollo sostenible de un país sobre todo en su economía, principalmente debido a que el capital humano es deficiente por falta de un desarrollo cognitivo y físico adecuado en la niñez (Ortiz et al., 2006: p.534).

Si bien en la literatura se enuncian los determinantes de la desnutrición, es necesario realizar un análisis actual de la realidad en nuestro país. Tomando en cuenta este contexto, la presente investigación a través de la Encuesta Nacional de Salud y Nutrición-ENSANUT 2018, pretende determinar los factores significativamente asociados a la desnutrición crónica y realizar predicciones sobre la misma, enfocándose en los niños de 0 a 5 años, por considerarse una edad crítica para el desarrollo integral de las personas, debido a que ha esta edad los niños desarrollan habilidades en cuatro áreas principales: desarrollo cognitivo (Aprendizaje y pensamiento), desarrollo social y emocional, desarrollo del habla, lenguaje y desarrollo físico (ALAMEDA, 2020).

Los resultados de este estudio, serán de gran utilidad para los hacedores de política pública, profesionales estadísticos inmiscuidos en esta área de investigación y médicos especializados en medicina familiar y pediatría.

## **Objetivos**

### ***Objetivo general***

Comparar las técnicas (regresión logística y árboles de decisión) con el fin de establecer el modelo que mejor prediga la desnutrición crónica en niños menores de cinco años y determine los factores asociados que la provocan.

### ***Objetivos específicos***

- Operacionalizar los módulos de la ENSANUT para conseguir una base de datos consolidada.
- Identificar y operacionalizar los posibles factores asociados a la desnutrición crónica infantil, tomando como referencia el marco teórico establecido a nivel mundial por la UNICEF.
- Operacionalizar la variable respuesta para identificar la desnutrición en niños de 0 a 5 años de edad.
- Determinar los factores significativamente asociados a la desnutrición crónica a través de los dos modelos.
- Comparar la efectividad de predicción de los dos modelos a través de medidas de bondad de ajuste.
- Implementar los modelos a través de una aplicación interactiva.

# CAPÍTULO I

## 1 MARCO TEÓRICO REFERENCIAL

### 1.1 Conceptos generales de desnutrición

#### 1.1.1 Desnutrición

La desnutrición es una condición patológica inespecífica, sistemática y reversible en potencia, que resulta de la deficiente utilización de los nutrientes por las células del organismo que se acompaña de variadas manifestaciones clínicas relacionadas con diversos factores ecológicos y que reviste diferentes grados de intensidad (Márquez et al., 2012: pp.64-67).

Los efectos de la desnutrición se pueden manifestar a lo largo de todo ciclo de vida porque las necesidades nutricionales cambian. En este proceso, cabe enfatizar las fases relacionadas a la vida intrauterina y neonatal, lactante, preescolar, vida escolar y vida adulta.

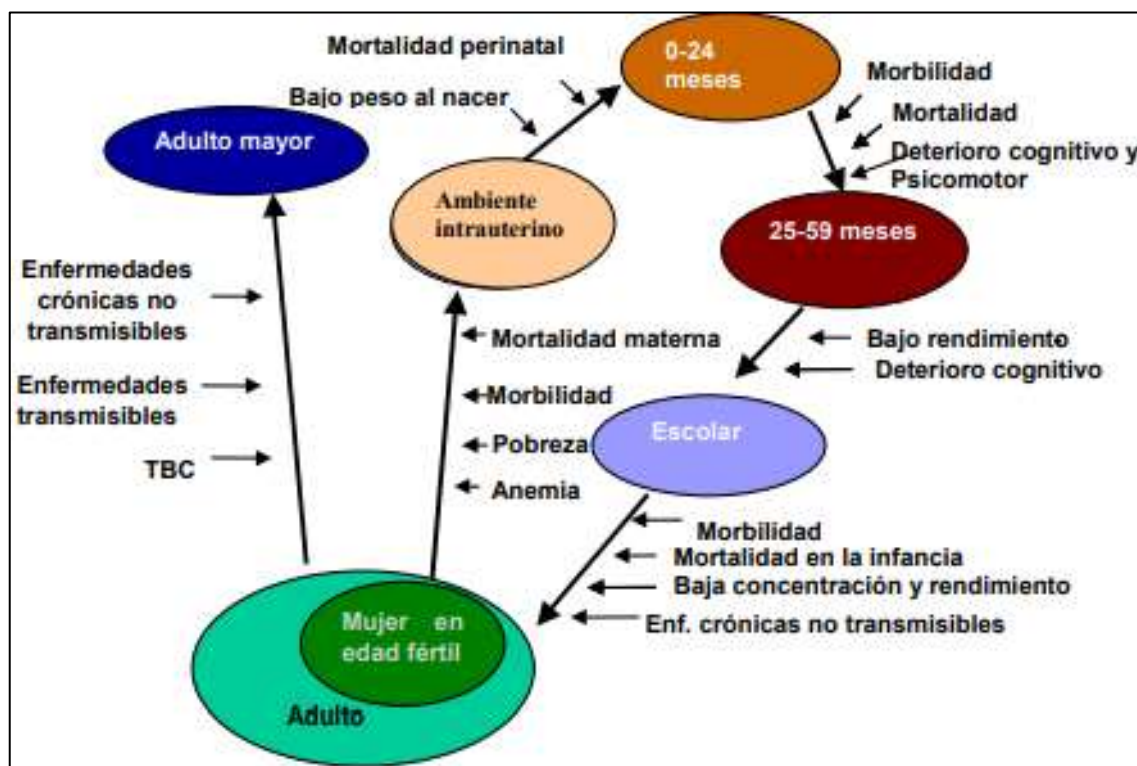


Figura 1-1: Ciclo de la malnutrición

Fuente: Branca y Ferrari, 2002

### 1.1.2 Clasificación de desnutrición infantil

Para evaluar el estado nutricional de un niño se utilizan indicadores obtenidos a través de medidas antropométricas (peso, talla/altura), estos indicadores son las herramientas más utilizadas en el análisis de la desnutrición infantil por su facilidad de uso. Para optimizar el uso de los indicadores de desnutrición, generalmente se convierten en z scores o desviaciones estándar (Palau, 2017: p.116).

$$z \text{ score} = \frac{(\text{peso del niño} - \text{media del peso de referencia})}{\text{desviación estándar del peso de la población de referencia}} \quad (1.1)$$

El “z score” es una medida de desviación estándar que describe en qué medida y dirección se desvía la medición antropométrica de un individuo del promedio de su sexo establecido por la OMS a través de los Patrones de Crecimiento Infantil.

**Tabla 1-1:** Clasificación del estado nutricional basado en z-scores

Clasificación	Valores de los z-scores
Adecuado	$-2 < z \text{ score} < 2$
Malnutrición Moderada	$-3 < z \text{ score} < -2$
Malnutrición Severa	$z \text{ score} < -3$

Fuente: OMS, 2005

Realizado por: Congacha, Giorgia, 2020

Al tomar en cuenta la desviación estándar se logra estandarizar las diferencias de peso independientemente de la altura del niño, a diferencia de otros indicadores de malnutrición, logra clasificar de manera acertada a los niños desnutridos. Es una expresión reconocida mundialmente, validada por la OMS.

Según la OMS existen tres tipos de desnutrición: crónica, aguda y global

**Desnutrición Crónica:** Produce retraso en el crecimiento del niño, se mide comparando la talla del niño con el estándar recomendado para su edad. Indica una carencia de los nutrientes necesarios durante un tiempo prolongado, por lo que aumenta el riesgo de que contraiga enfermedades y afecta el desarrollo físico e intelectual del niño. La desnutrición crónica presenta diferentes categorías; es moderada cuando el puntaje z se encuentra dentro de un rango  $-3 < z < -2$  y se considera severa cuando  $z < -3$  (UNICEF, 2013: p.6).

**Desnutrición aguda:** Un niño con desnutrición aguda pesa menos de lo que le corresponde con relación a su altura (disminución de masa corporal). Resulta de una pérdida de peso asociada con periodos recientes de hambruna o una enfermedad infecciosa que se desarrolla muy rápidamente. Se la puede clasificar como aguda moderada cuando el puntaje z se encuentra dentro de un rango

$-3 < z \leq -2$ ; o aguda grave cuando  $z < -3$ . En ambas situaciones el niño pesa menos de lo que mide, sin embargo en la segunda el peso es significativamente bajo para su altura, por lo que el riesgo de muerte es nueve veces mayor que para un niño en condiciones normales (UNICEF, 2011: pp.7-9).

**Desnutrición global:** También conocida como insuficiencia ponderal, es un indicador que engloba las anteriores. Se lo puede definir como la relación del peso con la edad. Captura la situación nutricional pasada y presente de los infantes; sin embargo, este indicador no es apropiado para estudios poblacionales, puesto que es más impreciso que los dos mencionados anteriormente. Se la puede clasificar como moderada cuando el peso es dos desviaciones estándar inferior a la mediana poblacional y como severa cuando se encuentra a tres desviaciones estándar por debajo de la mediana (UNICEF, 2011: p.12).

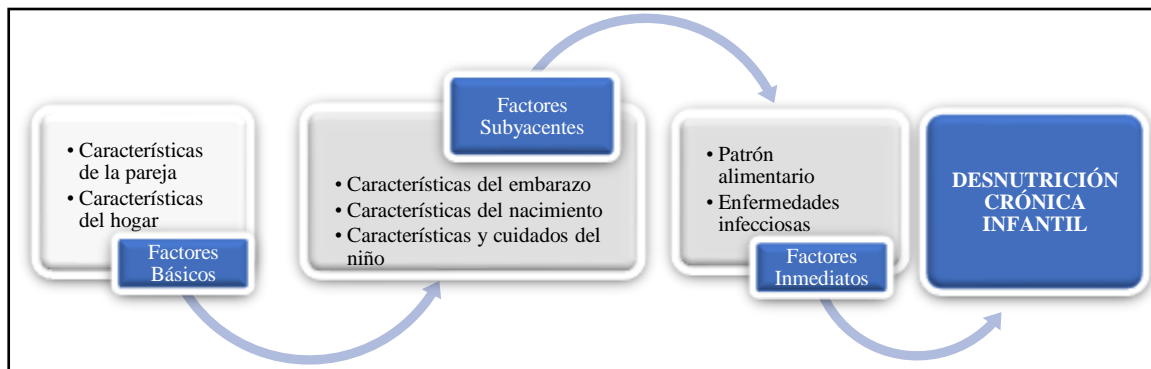
### ***1.1.3 Factores asociados a la desnutrición infantil***

En la década de los noventa Urban Jonsson propuso un marco conceptual para analizar los factores asociados a la desnutrición infantil que más tarde este fue adoptado por la UNICEF, el cual considera que dichos factores están asociados a una serie de condiciones sociales y pueden clasificarse en tres grupos: Básicas, asociadas a los sistemas políticos, contextos culturales, económicos y sociales; Subyacentes, asociadas a la educación, acceso a información y actitudes discriminatorias que limitan el acceso de las familias a recursos y servicios; e Inmediatas, que se asocian al estado de salud y el consumo de alimentos durante la niñez (UNICEF, 1998: pp.97-103).

Existe una adaptación al marco conceptual de la UNICEF realizado en 2005 por Mukuria J. con el fin de estudiar el estado nutricional de los niños de entre 0 y 35 meses de edad de 41 países de África, Europa, Asia y América Latina, donde la fuente de información fueron las encuestas Demográficas y de Salud Familiar (ENDES) del período 1994-2001.

Esta adaptación considera cuatro conjuntos de factores o causas asociadas a la desnutrición infantil: Factores inmediatos como enfermedades infecciosas e inadecuada alimentación, detrás de las cuales se encuentran factores biológicos y de comportamiento como el patrón alimentario, inmunización, cuidado de la salud del niño y características de la madre y del niño; factores socioeconómicos como educación, actividad económica de la madre y características del hogar, finalmente factores básicos como la estructura política, económica e ideológica, área de residencia (urbana/rural) (Mukuria et al., 2005: pp.49-52).

En este contexto y con la finalidad de identificar y establecer posibles relaciones entre factores asociados a la desnutrición crónica infantil en el Ecuador, se utilizará el marco conceptual planteado por la UNICEF y su adaptación propuesta por Mukuria J.



**Figura 2-1:** Clasificación de los posibles factores asociados a la desnutrición crónica infantil, de acuerdo al marco teórico de la UNICEF

**Fuente:** UNICEF, 1998

Asumimos, por tanto, que existen tres conjuntos de factores: Básicos, Subyacentes e Inmediatos, los cuales están asociados a distintos niveles de la desnutrición crónica infantil. Los factores inmediatos están relacionados con características del patrón alimentario y de enfermedades transmisibles, los factores básicos están relacionados con las características de la pareja y del hogar y los factores subyacentes se hallan relacionados con características del embarazo, del nacimiento y cuidado del niño.

Los factores inmediatos estarán asociados directamente a la desnutrición crónica infantil, mismos que a su vez son intermediadores de una asociación indirecta que existe entre los factores básicos y subyacentes con la desnutrición crónica infantil.

### **1.1.4 Encuesta Nacional de Salud y Nutrición - ENSANUT**

#### **1.1.4.1 Estructura de la ENSANUT**

El Instituto Nacional de Estadísticas y Censos (INEC) en acompañamiento técnico del Banco Interamericano de Desarrollo (BID), Fondo de las Naciones Unidas para la Infancia (UNICEF) e instituciones públicas como el Ministerio de Salud Pública (MSP) y el Ministerio de Inclusión Económica y Social (MIES) realizan la Encuesta Nacional de Salud y Nutrición (ENSANUT), la cual es una operación estadística con periodicidad quinquenal, cuya población objetivo son todos los miembros del hogar; su cobertura geográfica comprende las 24 provincias del país. Está compuesta por 5 formularios los mismos que se detallan a continuación:

#### **Formulario 1: Del Hogar**

Recaba información socio-económica de todos los miembros del hogar, el mismo que se conforma por 8 secciones que tratan las siguientes temáticas: datos de la vivienda y el hogar; información de los miembros del hogar; actividades económicas para personas de 10 años o más; uso de

servicios y gastos en salud; seguridad alimentaria; antropometría y Etiquetado de alimentos y bebidas procesadas.

*Formulario 2: Mujeres en edad fértil*

Recoge información sobre temas relacionados a la salud sexual, materna, conocimiento de métodos anticonceptivos, lactancia materna, salud en la niñez a las mujeres en edad fértil de 10 a 49 años de edad. Este formulario está compuesto por 10 secciones: Características generales de la entrevistada; historia de embarazos y nacimientos; lactancia materna (niños menores de 3 años); salud en la niñez (niños menores de 5 años); servicios asociados a la salud materna; planificación familiar; preferencias reproductivas; actividad sexual y salud reproductiva en mujeres; nupcialidad; infecciones de transmisión sexual (ITS/VIH/SIDA)

*Formulario 3: Salud sexual y reproductiva hombres 12 años y más*

Recoge información referente a los principales problemas de salud sexual y reproductiva de los hombres de 12 años. Se compone de 4 secciones: Selección del hombre de 12 años o más; actividad sexual y salud reproductiva; planificación familiar; infecciones de transmisión sexual (ITS/VIH/SIDA).

*Formulario 4: Factores de riesgo 5 a menores de 18 años*

Este formulario está compuesto por 6 secciones que recogen información de las siguientes temáticas: selección de la persona de 5 a menores de 18 años; salud oral; actividad física; alimentación y nutrición; consumo de bebidas alcohólicas y consumo de tabaco.

*Formulario 5: Desarrollo infantil para niños menores de 5 años.*

Recopila información relacionada al nivel de desarrollo integral de niños y niñas de 0 a 5 años, considerando aspectos como el lenguaje, aprendizaje, juego, desarrollo emocional y motor, cuidado y buen trato, calidad del ambiente en el hogar e interacciones. Está conformado por 13 secciones: selección del niño menor de 5 años; programas de primera infancia (niños/as de 0 a menores de 5 años); oportunidades de juego en el hogar para niños/as menores de 5 años; disciplina infantil; desarrollo, aprendizaje y educación para niños/as de 3 a menores de 5 años; lenguaje de niños/as de 12 a 18 meses; lenguaje de niños/as de 19 a 30 meses; lenguaje de niños/as de 31 a 42 meses; lenguaje-Peabody para niños/as de 43 a 59 meses; inventario home para niños/as menores de 3 años; inventario home para niños/as de 3 a menores de 5 años; motricidad gruesa y desarrollo para niños de 0 a 23 meses; madurez emocional para niños/as de 4 a menores de 5 años(INEC, 2018: pp.14-26).

#### *1.1.4.2 Historia de la ENSANUT*

En 1986, se realizó por primera vez en el país un análisis de la situación alimentaria, nutricional y de salud de niños menores de cinco años a través de la Encuesta Nacional sobre la situación Alimentaria, Nutricional y de Salud de la Población de niños ecuatorianos menores de cinco años-DANS. Esta encuesta reveló la existencia de elevadas tasas de desnutrición aguda, global y crónica, también evidenció la existencia de deficiencias específicas de micronutrientes, en particular deficiencias de hierro y zinc.

En 1987, se realiza la Encuesta Demográfica y de Salud Materna e Infantil (ENDEMAIN) de manera periódica por el Centro de Estudios de Población y Desarrollo Social (CEPAR), en ella se fueron incorporando algunos temas como: roles de género, violencia intra-familiar, prácticas, conocimientos y actitudes sobre enfermedades de transmisión sexual y SIDA, cuidado de la salud y aspectos laborales de la mujer. La última ENDEMAIN se realizó en 2004 en ella se incluyó información de mortalidad materna, antropometría, asistencia escolar, uso de servicios, gastos en salud y gastos de consumo de los hogares; tuvo una cobertura Nacional y representatividad a nivel urbano y rural, por regiones y provincias. Entre los principales resultados se evidenció una disminución en las tasas globales de desnutrición infantil y también se constató una epidemia de sobrepeso y obesidad en las mujeres en edad fértil.

En el año 2012 se realizó el levantamiento de la primera Encuesta Nacional de Salud y Nutrición (ENSANUT), con una cobertura nacional (área urbana y rural) la cual permitió representatividad a nivel de subregiones, grupos étnicos, provincias y áreas rurales y urbanas. A través de esta encuesta el Ministerio de Salud Pública se propuso recabar información relevante para el diseño de políticas públicas, planes y programas sobre la situación de salud reproductiva, enfermedades crónicas no transmisibles, actividad física, situación alimentaria y nutricional en la población menor de 60 años, considerando la diversidad geográfica, demográfica, cultural, étnica, social y económica del país. Los principales resultados obtenidos con la ENSANUT 2012 muestran que el 25.3% de los menores de 5 años en Ecuador tienen desnutrición crónica.

Con la finalidad de contar con información actualizada que permita evaluar políticas asociadas a la erradicación de la desnutrición crónica en menores de cinco años, generar nuevas políticas, planes y proyectos que permitan contribuir al desarrollo de la población, el INEC desarrolla la ENSANUT 2018-2019 en dos etapas, la primera a través de un proceso de actualización del marco de muestreo ya que no se contaba con toda la información requerida a nivel de Unidad primaria de muestreo – UPM y la segunda etapa implicó realizar la repartición por cuotas considerando la distribución de cada una de las poblaciones objetivo dentro de las UPM seleccionadas (INEC, 2018: pp. 5-8).



#### *1.1.4.3 Objetivo ENSANUT*

El principal objetivo de la Encuesta Nacional de Salud y Nutrición es generar indicadores sobre los principales problemas y la situación de salud de la población ecuatoriana con la finalidad de proporcionar a las autoridades, hacedores de política pública, sector privado, academia y población en general, información actualizada de la salud sexual y reproductiva, estado nutricional de la población, actividad física y acceso a programas de complementación alimentaria, que les permita comparar los resultados anteriores, así como evaluar la efectividad de los programas implementados(INEC, 2019: p.17).

#### *1.1.4.4 Justificación ENSANUT*

La ENSANUT responde a la necesidad del Estado Ecuatoriano de contar con información actualizada sobre salud y nutrición de la población a fin de que sirva de base para el diseño de políticas públicas y programas que permitan controlar problemas sociales y si es posible reducirlos a niveles que dejen de constituir problemas de salud pública. Además proporcionar información estadística que permita dar seguimiento a los objetivos del Plan Nacional de Desarrollo (PND) y demás agendas de desarrollo Nacional e Internacional(INEC, 2019: p.7).

### **1.2 Teoría Estadística**

#### *1.2.1 Análisis exploratorio de datos*

Previo a aplicar técnicas inferenciales, es importante llevar a cabo una exploración de los datos que se va a manipular, con la finalidad de detectar errores en la codificación de variables, eliminar inconsistencias, evaluar la magnitud y tipo de valores perdidos, conocer características básicas de la distribución de las variables. (López, 2008: p.2).

#### *1.2.2 Imputación de datos faltantes*

La presencia de datos faltantes en estudios censales, es un problema común especialmente en el ámbito de las ciencias sociales. Hace varias décadas se han estudiado técnicas de llenado de datos con el fin de obtener un conjunto de datos completos para analizarlos a través de métodos estadísticos tradicionales; sin embargo, este escenario se complica cuando se trabaja con grandes conjuntos de datos formadas por diversas variables sobre la cual se realizan estudios multivariantes, haciéndose necesaria la aplicación de métodos que imputen conjuntamente los datos.

La proporción de datos faltantes en un conjunto de datos puede variar dependiendo del estudio o la dificultad de la medición de una unidad, sin embargo, es decisión del investigador determinar hasta que porcentaje de pérdida de valores faltantes se considera tratable mediante imputación. Si

existe un alto porcentaje de valores ausentes en una variable, el investigador discernirá de acuerdo a su experiencia si debe eliminarla o no. El número de valores ausentes para eliminar una variable es muy relativo, pero se suele eliminar cuando el 50% de los valores están ausentes.

En la práctica se habla de pérdidas máximas entre 1 y 20% de los datos dependiendo de la exactitud del estudio y del área de investigación. Por ejemplo, en las ciencias médicas, la precisión es un factor determinante en la obtención de resultados, no pueden permitir la imputación de muchos valores que nunca serán reales sólo para poderlos analizar, mientras que en las ciencias sociales permite porcentajes de valores faltantes imputados más altos, siempre estará en manos del investigador esta decisión (Useche y Mesa, 2006: p.130).

El correcto tratamiento de valores faltantes es una etapa fundamental, ya que si no se aplica la técnica correcta se pueden cometer grandes errores y falsos resultados durante el procesamiento de los mismos.

#### *1.2.2.1 Técnicas para el tratamiento de valores faltantes*

Las técnicas utilizadas para el tratamiento de datos perdidos pueden clasificarse en tres tipos: técnicas de borrado, técnicas tolerantes y técnicas de imputación.

Las técnicas de borrado eliminan valores perdidos y se dividen en dos tipos: LD (*Listwise Deletion*) y PD (*Pairwise Deletion*), LD elimina por completo los casos con valores perdidos, mientras que PD considera cada característica de forma independiente, a la hora de operar se toma en cuenta únicamente los casos con valores completos.

Las técnicas tolerantes permiten trabajar con una base de datos incompleta, este método tolera los datos perdidos para todas las variables, menos la variable de clase.

Las técnicas de imputación estiman y rellenan valores perdidos usando toda la información disponible para generar un conjunto de datos completo. Las técnicas más utilizadas para la imputación son:

*Mean Imputation* (MI): Los valores ausentes se sustituyen por la media aritmética de los valores observados. Presenta el problema de subestimación de la varianza.

Imputación múltiple: consiste en imputar un número  $m > 1$  de veces la variable aleatoria a partir de la función distribución de la variable a imputar, obteniendo así  $m$  datasets completos con el objetivo de analizar el comportamiento de estos datos con técnicas que no toleren datos faltantes y así reducir el problema de la varianza. Presenta el inconveniente de requerir un elevado coste computacional.

*K Nearest Neighbours* (kNN): identifica los k casos más similares al valor a imputar. Para calcular la similitud entre los casos, considera a cada caso como un vector y utiliza una medida de distancia (Euclídea, Minkowski, Manhattan, Chebysev) para medir la diferencia entre los valores de una misma variable. Una vez obtenidas las distancias entre variables, calcula la distancia entre casos como la suma de las distancias entre sus variables, cuanto menor es la distancia entre dos casos más similares se consideran. Este método de imputación es muy acertado al trabajar en una base de datos grande (Cubas, 2017: pp.26-28).

### **1.2.3 Balanceo de clases**

Cuando nos enfrentamos a un problema de clasificación la mayoría de veces la variable a predecir no está balanceada, es decir, uno o varias de las clases son minoritarias frente al resto de las clases, lo que puede hacer que nuestro algoritmo no obtenga la información necesaria sobre esa clase o clases minoritarias.

Existen métodos que corrigen este desequilibrio entre las distintas clases, modificando el tamaño original del conjunto de datos y proporcionando un equilibrio entre clases.

Entre los más importantes se encuentran:

- **Submuestreo:** se basa en reducir la clase mayoritaria, reduciendo las observaciones del conjunto de datos para que sea equilibrado. Esta eliminación de clases puede hacerse de manera aleatoria o con un criterio de selección preestablecido.
- **Sobremuestreo:** Se basa en aumentar la clase minoritaria, replicando las observaciones de dicha clase y aumentando los datos hasta que el conjunto de datos este balanceado. Puede utilizar un criterio aleatorio o preestablecido.
- **Sintético:** Crea datos artificiales utilizando bootstrapping y k-vecinos más cercanos. Se puede considerar un método de sobremuestreo.
- **Matriz de costos:** Se basa en el costo asociado con las observaciones de clasificación equivocada o errónea. No crea una distribución de datos balanceada (Analytics Vidhya, 2016).

### **1.2.4 Regresión logística**

La regresión logística, es una técnica estadística multivariante que ayuda a modelar una variable de respuesta categórica en función de variables explicativas o predictoras cuantitativas o categóricas. Forma parte de los GLM (modelos lineales generalizados) y se aplica en diferentes campos como las ciencias de la salud, ciencias sociales, economía, ecología, entre otros campos. El objetivo de los modelos logit es estimar probabilidades o predecir un suceso definido por la variable respuesta categórica en función de las variables predictoras.

### 1.2.4.1 Modelo logístico binario

En el modelo de regresión logística binaria, la variable respuesta  $Y$  es una variable binaria (o dicotómica) que solo puede tomar dos valores 0 y 1 con probabilidad  $\pi_i$  para  $Y_i = 1$  y probabilidad  $1 - \pi_i$  para  $Y_i = 0$ .

El análisis de regresión logística comprende la estimación de la probabilidad de que ocurra un evento (variable respuesta dicotómica) como función de los valores de  $p$  variables independientes.

Consideremos  $Y$  una variable respuesta y  $p$  una colección de variables independientes expresadas por el vector  $X' = (x_1, x_2, \dots, x_p)$ .

La forma específica del modelo logístico con  $p$  variables predictoras está representada por:

$$\pi = \pi(x) = P(Y = 1|X = x) = \frac{e^{x'\beta}}{1 + e^{x'\beta}} \quad (2.1)$$

Representa la probabilidad condicional de que el evento  $Y = 1$  ocurra dada la ocurrencia de un conjunto de variables  $X$  (probabilidad de éxito).

Una transformación de  $\pi(x)$  que es fundamental para el estudio de regresión logística es la transformación logit. Esta transformación se define en términos de  $\pi(x)$ , como:

$$g(x) = \ln \left[ \frac{\pi(x)}{1 - \pi(x)} \right] = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p \quad (3.1)$$

Donde  $\beta_0$  es la constante y los  $\beta_i$  son los coeficientes de los predictores  $x_i$  del modelo. La importancia de esta transformación es que  $g(x)$  posee muchas de las propiedades deseables de un modelo de regresión lineal. La función logit es lineal en sus parámetros, puede ser continuo y variar de  $-\infty$  a  $+\infty$ , dependiendo del rango de  $x$ .

El modelo logístico puede expresarse en términos de **odds** (disparidad o ventaja) de ocurrencia de eventos. Esta razón se define como el cociente entre la probabilidad de éxito y la probabilidad de fracaso. Esto es:

$$odds = \frac{\pi(x)}{1 - \pi(x)} \quad (4.1)$$

La odds constituye una manera diferente de parametrizar una variable dicotómica, de modo alternativo a hacerlo mediante la probabilidad de éxito. Mientras la probabilidad de éxito toma valores de intervalo  $[0,1]$ , la odds puede tomar valores en el intervalo  $[0, +\infty]$ .

#### 1.2.4.2 Estimación de los parámetros del modelo

Debido a que la distribución de  $Y$  dado un conjunto de variables  $X = (x_1, x_2, \dots, x_p)$  no es normal y no existe homocedasticidad en los errores, la estimación del vector  $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p)$  por el método de mínimos cuadrados no tiene propiedades óptimas en su lugar emplearemos el método de máxima verosimilitud para obtener los valores de los parámetros desconocidos que maximizan la probabilidad de obtener el conjunto observado de datos.

Para estimar los parámetros  $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p)$  del modelo se utiliza el método de máxima verosimilitud, con la cual encontramos el valor de  $\beta$  que maximiza  $l(\beta)$ .

La función de verosimilitud adopta la forma:

$$l(\beta) = \prod_{i=1}^n (\pi(x_i)^{y_i} [1 - \pi(x_i)]^{1-y_i}) \quad (5.1)$$

Aplicando logaritmo neperiano, la expresión  $L(\beta)$  se define como:

$$L(\beta) = \ln[l(\beta)] = \sum_{i=1}^n \{y_i \ln[\pi(x_i)] + (1 - y_i) \ln [1 - \pi(x_i)]\} \quad (6.1)$$

Para encontrar el valor de  $\beta$  se deriva  $L(\beta)$  con respecto a  $\beta_0, \beta_1, \dots, \beta_p$  y se iguala al valor cero obteniéndose:

$$\sum_{i=1}^n [y_i - \pi(x_i)] = 0 \quad (7.1)$$

$$\sum_{i=1}^n x_{ij} [y_i - \pi(x_i)] = 0 \quad \forall j = 1, \dots, p \quad (8.1)$$

Para encontrar la solución de este conjunto de ecuaciones se utiliza el método iterativo de Newton-Raphson. Actualmente existen paquetes estadísticos para estimar los parámetros  $\beta$ .

### 1.2.4.3 Contrastes o pruebas de significancia del modelo

En el análisis de regresión logística se consideran los siguientes contrastes o pruebas estadísticas de significancia:

#### Contraste de Wald

En este contraste se evalúa estadísticamente los coeficientes de regresión logística. Se quiere contrastar si un parámetro  $\beta_i = 0$ , con  $i = 1, 2, \dots, k$  frente a si son significativamente distintos de 0, es decir:

$$H_0: \beta_i = 0,$$

$$H_1: \beta_i \neq 0$$

Y se contrasta mediante el estadístico de Wald:

$$W_i = \frac{\hat{\beta}_i}{\widehat{SE}(\beta_i)} \sim N(0,1) \quad (9.1)$$

$\hat{\beta}_i$  y  $\widehat{SE}(\beta_i)$  son las estimaciones del modelo para  $\beta_i$  y el error estándar de  $\beta_i$ . Los coeficientes son significativos si tienen un valor p inferior a 0.05.

Se pueden también determinar intervalos de confianza para  $\beta_i$  puesto que el estadístico de Wald se distribuye de forma normal estándar, entonces los extremos inferior y superior son respectivamente

$$\hat{\beta}_i - z_{\alpha/2} \widehat{SE}(\beta_i) \text{ y } \hat{\beta}_i + z_{\alpha/2} \widehat{SE}(\beta_i) \quad (10.1)$$

Existe una estrecha relación entre contrastes de hipótesis e intervalos de confianza. Si el intervalo de confianza incluye el 0, significa que al nivel  $\alpha$  elegido no se podría rechazar la hipótesis nula de que  $\beta_i = 0$

### 1.2.4.4 Pseudo estadísticas $R^2$

**$R^2$  de Cox y Snell:** Coeficiente de determinación generalizado que se utiliza para estimar la proporción de la varianza de la variable dependiente explicada por las variables independientes. Se basa en la comparación del logaritmo de la verosimilitud para el modelo respecto al logaritmo de la verosimilitud para un modelo de línea base. Los valores oscilan entre 0 y 1.

**R<sup>2</sup> de Nagelkerke:** Es una versión corregida de la R<sup>2</sup> de Cox y Snell, esta tiene un valor máximo inferior a 1, incluso para el modelo perfecto. La R<sup>2</sup> de Nagelkerke corrige la escala del estadístico para cubrir el rango completo de 0 a 1.

#### 1.2.4.5 Evaluación de la bondad de ajuste del modelo

### AIC

Una herramienta estadística útil para elegir el número de retrasos de p y q, es el criterio de información de Akaike (AIC), la cual se basa en la suma de los cuadrados de los errores, buscando minimizar a partir de diversas combinaciones de p y q.

$$AIC = \ln \hat{\sigma}^2 + \frac{2}{n}r \quad (11.1)$$

Donde:

$\ln$ : logaritmo neperiano

$\hat{\sigma}^2$ : Suma residual de cuadrados dividida entre el número de observaciones

$n$ : Número de observaciones

$r$ : Número total de parámetros (incluyendo el término constante)

### Test de Hosmer y Lemeshow

Para evaluar la bondad de ajuste del modelo, Hosmer Lemeshow utiliza una estrategia de agrupamiento para obtener la estadística de bondad de ajuste, obtenida por el cálculo de la estadística Chi-cuadrado de Pearson de una tabla de frecuencias observadas y frecuencias esperadas estimadas.

Hosmer Lemeshow prueba las siguientes hipótesis:

$H_0$ : *No existen diferencias entre los valores observados y predichos*

$H_1$ : *Existen diferencias entre los valores observados y predichos*

Si rechazamos  $H_0$ , implica que el modelo ajustado no es el adecuado.

Se dividen todos los casos en deciles basados en las probabilidades predichas, el primer decil cuenta los casos con las probabilidades más altas, siendo el estadístico de prueba:

$$\hat{C} = \sum_{k=1}^g \frac{(O_k - n_k \bar{\pi}_k)^2}{n_k \bar{\pi}_k (1 - \bar{\pi}_k)} \quad (12.1)$$

Donde:

$O_k$ : Número de respuestas entre las covariables

$n_k$ : Número de covariables en el k-ésimo decil

$\bar{\pi}_k$ : Probabilidad media estimada

La estadística  $\hat{C}$  tiene aproximadamente una distribución chi-cuadrado con  $g-2$  grados de libertad, bajo la  $H_0$  a un nivel de significancia  $\alpha$ , rechazamos  $H_0$  si:

$$\hat{C} > \chi_{1-\alpha}^2(g - 2) \quad (13.1)$$

Y concluimos que el modelo no es el adecuado (Hosmer y Lemeshow, 2000: pp.7-23).

#### 1.2.4.6 *Ventajas y desventajas del modelo de regresión logística*

Las principales ventajas al utilizar regresión logística son:

- Fácil de entender y explicar
- Rara vez existe sobreajuste
- Fácil de entrenar sobre grandes datos gracias a su versión estocástica
- Los resultados son altamente interpretables

Las desventajas al utilizar este modelo son:

- En algunas ocasiones es muy simple para captar relaciones complejas entre variables
- Puede sufrir con valores atípicos (aprendeIA, 2020).

#### 1.2.5 *Machine Learning*

Aprendizaje Automático o Machine Learning es una subrama de las ciencias de la computación que permite a las computadoras aprender de los datos patrones y relaciones que existen en ellos y que nos ayudan en la toma de decisiones.

El aprendizaje automático se apoya en las tecnologías de Big Data que surgen para resolver tres tipos de problemas:

1. Variedad de los datos: Han surgido nuevos tipos de datos como los datos no estructurados.



2. Escalabilidad: se busca la rapidez en el rendimiento y procesamiento de los datos, por lo que se escala en horizontal.
3. Velocidad: la velocidad de generación de datos, necesita velocidad de procesamiento.

Los algoritmos de aprendizaje automático se clasifican en dos tipos:

- **Aprendizaje supervisado:** Son aquellos problemas de aprendizaje automático en los que el algoritmo aprende de datos previos que es la variable a predecir. Su objetivo es aprender de estos datos para identificar patrones y reglas que permitan predecir la variable dependiente al enfrentarse a nuevos casos. Dentro del aprendizaje supervisado podemos encontrarnos con dos tipos de problemas en función de la variable a predecir: problemas de regresión y problemas de clasificación.
  - **Problemas de regresión:** Nos enfrentamos a un problema de regresión cuando la variable a predecir es una variable continua.
  - **Problemas de clasificación:** Nos enfrentamos a un problema de clasificación cuando la variable a predecir es una variable categórica. Si la variable a predecir solo puede tomar dos valores nos enfrentamos a un problema de clasificación binaria y si la variable toma más de dos valores nos enfrentamos a un problema de clasificación multiclase.
- **Aprendizaje no supervisado:** Este tipo de aprendizaje automático no requiere de una variable a predecir, busca obtener relaciones, diferencias o asociaciones entre las distintas observaciones (Hernández-Leal et al., 2017: pp.17-21).

## ***1.2.6 Árboles de decisión individuales***

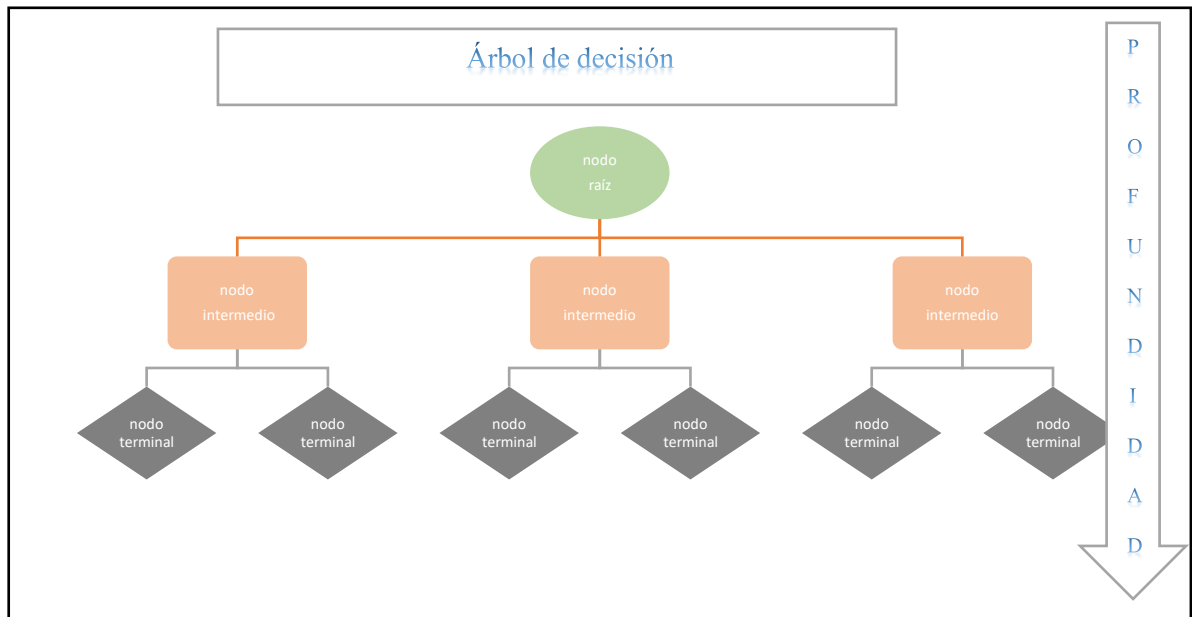
### *1.2.6.1 Definición*

Modelo de predicción que divide el espacio de los predictores agrupando observaciones con valores similares para la variable respuesta o dependiente. Divide el espacio muestral en subregiones, mediante la aplicación de una serie de reglas o decisiones, buscando que cada subregión contenga la mayor proporción posible de individuos de una de las poblaciones. Si una subregión contiene datos de diferentes clases, es subdividida en regiones más pequeñas hasta dividir el espacio en subregiones menores que contengan datos de la misma clase (Ferrero y López, 2020).

### *1.2.6.2 Estructura básica de un árbol de decisión*

Los árboles de decisión, tanto de clasificación como de regresión, están formados por nodos. Su lectura se realiza de arriba hacia abajo, donde el primer nodo es el nodo raíz, en el cual se realiza la primera división en función de la variable más importante; después de esta división se

encuentran los nodos intermedios que vuelven a dividir el conjunto de datos en función del resto de variables. En la parte inferior se encuentran los nodos terminales u hojas donde se indica la clasificación definitiva. La profundidad de un árbol es el número máximo de nodos de una rama. A continuación, se observa su estructura:



**Figura 3-1:** Estructura básica de un árbol de decisión

**Fuente:** Ferrero y López, 2020

### 1.2.6.3 Generación de un árbol de decisión

La creación de un árbol de decisión de un problema de clasificación se realiza por medio del algoritmo de Hunt que se basa en la división en subconjuntos, buscando una separación óptima. Dado un conjunto de registros de entrenamiento de un nodo, si pertenecen a la misma clase se considera un nodo terminal, pero si pertenecen a varias clases, se dividen los datos en subconjuntos más pequeños en función de una variable y se repite el proceso. Para seleccionar que variable elegir para obtener la mejor división se puede considerar el Error de Clasificación, el índice Gini o la Entropía.

El índice de Gini mide el grado de pureza de un nodo, mide la probabilidad de no sacar dos registros de la misma clase del nodo. A mayor índice de Gini menor pureza, por lo que se selecciona la variable con menor Gini ponderado.

Suele seleccionar divisiones desbalanceadas, donde normalmente aísla en un nodo a una clase mayoritaria y el resto de clases los clasifica en otros nodos (Ferrero, 2020).

El índice de Gini se define como:

$$GINI(t) = 1 - \sum_{i=1}^n (P_i)^2 \quad (14.1)$$

Donde  $P_i$  es la probabilidad de que un ejemplo sea de clase  $i$ .

La entropía es una medida que permite cuantificar el desorden de un sistema. Si un nodo es puro su entropía es 0 y solo tiene observaciones de una clase; pero si la entropía es igual a 1 existe la misma frecuencia para cada una de las clases de observaciones. La entropía tiende a crear nodos balanceados en el número de observaciones.

La entropía se define como:

$$H = - \sum_{i=1}^n P_i \times \log_2 P_i \quad (15.1)$$

Donde  $P_i$  es la probabilidad de que un ejemplo sea de clase  $i$ .

En el caso de árboles de decisión para problemas de regresión se utiliza el RSS (*Residual Sum of Squares*), esta mide la discrepancia entre los datos reales y los predichos por el modelo. Un RSS bajo indica un buen ajuste del modelo a los datos, por lo tanto, se busca minimizar el RSS.

El RSS se define como:

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (16.1)$$

Donde  $y_i$  es el valor real de la variable a predecir y  $\hat{y}_i$  es el valor predicho.

#### 1.2.6.4 Ventajas e inconvenientes de los árboles de decisión

Las principales ventajas al utilizar árboles de decisión son:

- Son fáciles de construir, interpretar y visualizar. Además, selecciona las variables más importantes y en su creación no siempre utiliza todos los predictores.
- No necesita que se cumplan los supuestos (linealidad, normalidad, homogeneidad).
- Permite relaciones no lineales entre las variables explicativas y la variable dependiente.
- Sirven para categorizar variables numéricas

Algunos de los inconvenientes al momento de utilizar árboles de decisión son:

- Suelen tender al sobreajuste u *overfitting*, por lo que al predecir nuevos casos el modelo no estima con el mismo índice de acierto.
- Poco eficientes en problemas de regresión.

- En conjuntos desbalanceados se pueden crear árboles sesgados.
- Se ven influenciados por datos atípicos, creando árboles con ramas muy profundas que no predicen bien para nuevos casos.
- Tienen una alta variabilidad.

#### 1.2.6.5 Técnicas de ensemble

Una forma de mejorar el poder predictivo de los árboles de decisión y reducir su variabilidad es por medio de técnicas de ensemble o métodos combinados, estas son: Bagging, Random Forest y Gradient Boosting que consisten en dividir el conjunto de entrenamiento en varios subconjuntos, creando un árbol para cada uno de los subconjuntos. Cada árbol es entrenado sobre el conjunto de datos aportando una predicción. La predicción final será la media de las predicciones de cada árbol en los problemas de regresión y en los problemas de clasificación será la categoría más frecuente. Cuanto más diferentes sean cada uno de los modelos promediados, mejorará la predicción final conjunta. A continuación, se menciona brevemente cada una de estas técnicas, sus ventajas y desventajas (Amat, 2017).

**Bagging (Bootstrap Agregation):** Crea subconjuntos de datos del conjunto de datos original, manteniendo la misma distribución de los datos. Después para cada subconjunto de datos creados se entrena un árbol de decisión individual sin podar creando un bosque de árboles en paralelo. Cada modelo se construye con el  $2/3$  (63.2%) de los datos, dejando  $1/3$  para predecir las observaciones de *out of bag* (OOB), es decir, sobre datos con los que no ha sido entrenado el modelo y que sirve como estimación del error del test. Cuantos más árboles creamos menor varianza tenemos y por tanto menor error. Por último, se promedia las predicciones individuales de cada árbol si nos enfrentamos a un problema de regresión o se elige la clase con más frecuencia para los problemas de clasificación.

#### Ventajas e inconvenientes

Las ventajas de los modelos Bagging son:

- Mayor poder predictivo que los árboles de decisión individuales.
- Se reduce la varianza con respecto a los árboles individuales.
- Nos indica las variables principales o las más influyentes en el modelo.

Los inconvenientes de estos modelos son:

- Crea modelos más complicados que son difíciles de interpretar.

- Utiliza todas las variables para crear los modelos por lo que, si hay una variable más influyente sobre el resto, se utilizará en todos los modelos y sus predicciones serán muy parecidas (alta correlación de las predicciones).
- No podemos obtener una visión grafica del modelo.

**Random Forest:** Es un caso particular dentro del método Bagging para reducir la varianza. Cuando se crea cada modelo solo se le deja elegir entre los  $m$  predictores elegidos aleatoriamente de todos los predictores considerados. No permite elegir entre todas las variables independientes, sino que solo puede elegir entre un numero establecido y elegidas aleatoriamente, no seleccionando siempre las variables más influyentes, descorrelacionando las predicciones y mejorando el poder predictivo del modelo (Orellana, 2018).

### Ventajas e inconvenientes

Las ventajas de Random Forest son:

- Alto poder predictivo. Mejoran las predicciones de los árboles individuales.
- Nos sirven tanto para problemas de regresión como de clasificación.
- Robusto a valores atípicos.
- Nos da una aproximación del error (*out of bag*) al entrenar el modelo (es una aproximación, no el error que vamos a cometer al predecir).
- Se pueden realizar modelos con alto poder predictivo sin tener que ajustar parámetros.
- Se puede utilizar como herramienta de selección de predictores.

Los inconvenientes de estos modelos son:

- No son fáciles de interpretar ni de representar
- Con grandes conjuntos de datos el modelo se vuelve lento

**Gradient Boosting:** Este método consiste en ir ajustando modelos de árboles de decisión, de modo que cada árbol aprende del error del modelo anterior, los árboles se van creando de forma secuencial y no paralela, donde cada árbol que se crea tiene como objetivo reducir los errores del árbol anterior. Cada árbol que se crea en la continuación de la secuencia aprenderá de una versión actualizada de los residuos. Los árboles iniciales se denominan aprendices débiles cuyo sesgo es grande y su poder predictivo es bajo, pero la unión secuencial de estos aprendices débiles permite crear un modelo final donde se reduce tanto el sesgo como la varianza (Amat, 2017).

### Ventajas e inconvenientes

Las ventajas que tienen los modelos Boosting son:

- Alto poder predictivo.

- No necesita un preprocesamiento de los datos, aunque este puede mejorar su poder predictivo.
- No necesita imputación de datos faltantes.
- Nos facilitan las variables más importantes para la construcción del modelo.

Los inconvenientes de estos modelos son:

- Pueden mejorar hasta crear overfitting.
- Son difíciles de interpretar.
- Al tener tantos parámetros que se pueden utilizar en su afinamiento, hace que sean algoritmos complejos.
- Altos costes computacionales.

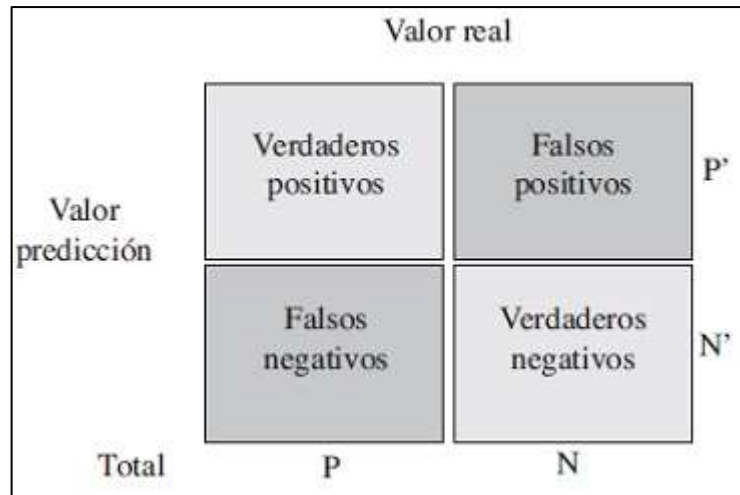
#### *1.2.6.6 Overfitting y Underfitting*

Cuando entrenamos un modelo nos podemos encontrar con dos problemas:

- **Overfitting:** Sobreentrena el modelo ajustándolo perfectamente a los datos de entrenamiento. El sobreajuste en los modelos de árboles de decisión produce árboles muy profundos, creando estructuras complejas que pueden llegar a tener tantos nodos terminales como observaciones tiene el conjunto de entrenamiento, clasificando cada una de las observaciones en un nodo terminal.
- **Underfitting:** Surge cuando el modelo no puede capturar correctamente la estructura de los datos por falta de entrenamiento del modelo. Esto puede suceder por tener escasas observaciones que no representan al conjunto de datos, falta de variables explicativas, excesivas limitaciones de parámetros del modelo, es decir tendrá poco poder predictivo.

#### *1.2.6.7 Métricas de error en problemas de clasificación*

Para entender las métricas que se puede utilizar en los problemas de clasificación se debe entender la matriz de confusión. En ella se comparan los valores predichos con los valores reales. En la diagonal de izquierda a derecha de la matriz de confusión se encuentran las predicciones correctas, tanto las clasificadas positivas como las negativas.



**Figura 4-1:** Matriz de confusión

Fuente: Sánchez, 2015

A partir de esta matriz podemos definir las siguientes métricas:

- Exactitud: Se puede definir como el porcentaje de predicciones que el modelo realizó correctamente. El objetivo es maximizar la exactitud.

$$Exactitud = \frac{VP + VN}{TP} \quad (17.1)$$

Donde **VP** son los verdaderos positivos y **VN** son los verdaderos negativos.

- Precisión: La precisión indica cuantas identificaciones positivas fueron correctas, nos indica los positivos reales. Se utiliza cuando determinar un falso positivo tiene un gran coste.

$$Precisión = \frac{VP}{VP + FP} \quad (18.1)$$

Donde **VP** son los verdaderos positivos y **FP** son los falsos positivos.

- Sensibilidad: Es importante en los casos que tienen un alto coste asociado a los falsos negativos.

$$Sensibilidad = \frac{VP}{VP + FN} \quad (19.1)$$

Donde **VP** son los verdaderos positivos y **FN** son los falsos negativos.

- Especificidad: Nos indica los verdaderos negativos entre los verdaderos negativos y los falsos positivos.

$$\text{Especificidad} = \frac{VN}{VN + FP} \quad (20.1)$$

Donde  $VN$  son los verdaderos negativos y  $FP$  son los falsos positivos.

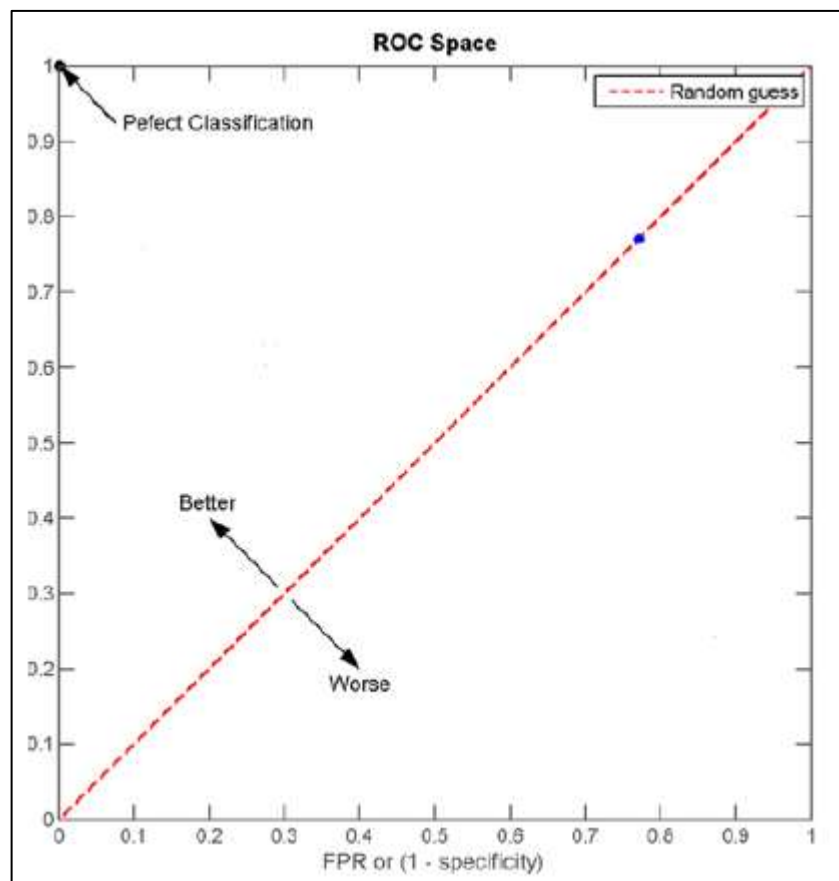
- Tasa de error: Nos indica el porcentaje de predicciones que el modelo realizó incorrectamente.

$$\text{Tasa de error} = \frac{FP + FN}{TP} \quad (21.1)$$

Donde  $FP$  son los falsos positivos y  $FN$  son los falsos negativos.

- Curva ROC: la curva ROC (Receiver Operating Characteristic) es la representación gráfica del ratio de verdaderos positivos frente al ratio de falsos positivos según el umbral de discriminación. Aquellos modelos de predicción que están por encima de la línea discriminante son mejores cuanto más separados están de la línea.

Los modelos que coinciden con la línea discriminante pueden clasificarse como aleatorios y los que están por debajo de la línea discriminante son peores o tienen algún error en la variable explicativa.



**Figura 5-1:** Esquema Curva ROC

Fuente: Gaspar, 2015



- AUC: La elección de la mejor curva ROC se hace mediante la comparación del espacio bajo la curva denominado AUC, está comprendida entre 0.5 y 1, donde 1 representa la predicción perfecta y 0.5 indica una predicción aleatoria. Si se obtiene valores menores a 0.5 puede haber problemas de concepto. Por lo tanto, siempre se elige aquella curva que tenga mayor AUC con respecto a otras (Srivastava, 2019).

## CAPITULO II

### 2 MARCO METODOLÓGICO

#### 2.1 Tipo y Diseño de investigación

La presente investigación es:

- Según el método de investigación es de tipo Mixta, puesto que los métodos a utilizar son cualitativos y cuantitativos;
- Según el objetivo es Aplicada, pues la investigación busca determinar los factores asociados que contribuyen a la problemática social (desnutrición crónica infantil);
- Según el nivel de profundización en el objeto de estudio es Explicativa, ya que existe una variable respuesta (el niño tiene o no desnutrición crónica), la cual será explicada por un conjunto de variables predictoras;
- Según la manipulación de variables es No experimental puesto que la matriz de datos para esta investigación proviene de una fuente de información secundaria (ENSANUT2018);
- Según el tipo de inferencia es inductiva ya que busca determinar cuáles son los factores influyentes en la desnutrición crónica infantil, además se pretende realizar predicciones con la técnica que proporcione menor error en el ajuste;
- Según el periodo temporal es de tipo Transversal esto se debe a que el objetivo principal de la investigación se enfoca en un único periodo de tiempo (año 2018) y el tiempo no representa un factor a considerarse en la creación de los modelos.

##### 2.1.1 *Localización del estudio*

La presente investigación se llevó a cabo a nivel nacional con información proporcionada por el Instituto Nacional de Estadísticas y Censos (INEC) a través de la encuesta Nacional de Salud y Nutrición ENSANUT 2018.

##### 2.1.2 *Población de estudio*

La ENSANUT 2018 está enfocada a todos los miembros del hogar a nivel nacional, en edades comprendidas de 0 a 49 años, sin embargo, la población objetivo del presente estudio son todos los niños ecuatorianos en edades de 0 a 5 años.

##### 2.1.3 *Tamaño de la muestra*

Tomando en cuenta los objetivos de la presente investigación, se filtró la información presentada por la ENSANUT de los niños menores a 5 años que fueron seleccionados aleatoriamente para

ser parte de la encuesta ENSANUT, con un total de individuos muestreados de 11.231, del cual se considera 70% para entrenamiento (7.862) y el 30% para validación (3.369).

#### 2.1.4 Método de muestreo

Se trabajó con información proveniente de la encuesta Nacional de Salud y Nutrición ENSANUT 2018 la cual utiliza un tipo de muestreo probabilístico estratificado trietápico de conglomerados.

#### 2.1.5 Recolección de información

La base de datos ENSANUT 2018 se obtuvo del repositorio que se encuentra en el sitio oficial del Instituto Nacional de Estadísticas y Censos INEC, los cuales fueron descargados en formato CSV de la página: [https://www.ecuadorencifras.gob.ec/documentos/web-inec/Estadisticas\\_Sociales/ENSANUT/ENSANUT\\_2018/BDD\\_DATOS\\_ABIERTOS\\_ENSANUT\\_2018\\_CSV.zip](https://www.ecuadorencifras.gob.ec/documentos/web-inec/Estadisticas_Sociales/ENSANUT/ENSANUT_2018/BDD_DATOS_ABIERTOS_ENSANUT_2018_CSV.zip)

## 2.2 Variables en estudio

Se estableció como variable explicada a “dcrónica”, la cual caracteriza a un infante como desnutrido crónico o no; la variable se codifica como: 0: no desnutrido crónico y 1: desnutrido crónico.

Para establecer las variables explicativas se tomó en cuenta las variables que propone la UNICEF como posibles factores asociados a la desnutrición crónica las cuales están repartidas en tres conjuntos de factores: Básicos (comprenden variables relacionadas a características de la pareja y características del hogar), Subyacentes (características del embarazo, nacimiento, características y cuidados del niño) e Inmediatos (patrón alimentario y enfermedades infecciosas).

Las variables en estudio se describen en la **Tabla 1-2** y **Tabla 2-2**.

#### 2.2.1 Operacionalización de variables

**Tabla 1-2:** Descripción de variables cuantitativas

Factores	Código	Variable	Tipo	Escala de medición
<b>Factores Básicos</b>	FBCP1_1	Años cumplidos de la madre	Cuantitativa	De razón
	FBCP1_5	Número de hijos nacidos vivos	Cuantitativa	De razón
	FBCH_11	Número de cuartos de la vivienda	Cuantitativa	De razón
	FBCH_12	Número de dormitorios de la vivienda	Cuantitativa	De razón
<b>Factores Subyacentes</b>	FSCE_3	Meses de embarazo cuando se hizo el primer control	Cuantitativa	De intervalo
	FSCE_4	Cuántos controles tuvo antes del parto	Cuantitativa	De razón
	FSCE_11	Primer control postparto-días	Cuantitativa	De razón
	FSCE_12	Primer control postparto-semanas	Cuantitativa	De razón

	FSCCE_13	Primer control postparto-meses	Cuantitativa	De razón
	FSCN_10	Talla al nacer	Cuantitativa	De razón
	FSCCN_2	control por primera vez-días	Cuantitativa	De razón
	FSCCN_3	control por primera vez-semanas	Cuantitativa	De razón
	FSCCN_4	control por primera vez-meses	Cuantitativa	De razón
	FSCCN_8	bcg-dosisf41002a	Cuantitativa	De razón
	FSCCN_9	hepatitis b-dosis	Cuantitativa	De razón
	FSCCN_10	pentavalente 1-dosis	Cuantitativa	De razón
	FSCCN_11	pentavalente 2-dosis	Cuantitativa	De razón
	FSCCN_12	pentavalente 3-dosis	Cuantitativa	De razón
	FSCCN_13	rotavirus 1-dosis	Cuantitativa	De razón
	FSCCN_14	rotavirus 2-dosis	Cuantitativa	De razón
	FSCCN_15	antipolio(opv) 1-dosis	Cuantitativa	De razón
	FSCCN_16	antipolio(opv) 2-dosis	Cuantitativa	De razón
	FSCCN_17	antipolio(opv) 3-dosis	Cuantitativa	De razón
	FSCCN_18	neumococo 1-dosis	Cuantitativa	De razón
	FSCCN_19	neumococo 2-dosis	Cuantitativa	De razón
	FSCCN_22	Edad (en meses)	Cuantitativa	De razón
<b>Factores Inmediatos</b>	FIPA_1	Hasta qué edad le dio el seno-días	Cuantitativa	De razón
	FIPA_2	Hasta qué edad le dio el seno-meses	Cuantitativa	De razón
	FIPA_3	Hasta qué edad le dio el seno-años	Cuantitativa	De razón
	FIPA_14	Consumió ayer -agua pura - cuántas veces	Cuantitativa	De razón
	FIPA_21	Consumió ayer -sopa- cuántas veces	Cuantitativa	De razón

Fuente: Diccionario de variables ENSANUT 2018

Realizado por: Congacha, Giorgia, 2020

**Tabla 2-2:** Descripción de variables cualitativas

Factores	Código	Variable	Tipo	Escala de medición	Categoría
<b>Variable dependiente</b>	dcronica	Desnutrición crónica menores de 5 años	Cualitativa	Nominal	0: Si 1: No
<b>FACTORES BÁSICOS</b>	FBCP1_4	El padre vive con usted en el hogar	Cualitativa	Nominal	Si No
	FBCP1_6	Grupo étnico	Cualitativa	Nominal	Mestizo/a Indígena Montuvio/a Blanco/a Afroecuatoriano/a Negro/a Otro
	FBCP1_7	Estado civil	Cualitativa	Nominal	Casado Divorciado Separado Soltero Unión de hecho

---

				Unión libre
				Viudo
FBCPI_8	Escolaridad de la madre	Cualitativa	Ordinal	Ninguno o Centro de Alfabetización Educación Básica Educación Media/Bachillerato Superior
FBCH_1	Vía de acceso principal a la vivienda	Cualitativa	Nominal	Calle pavimentada o adoquinada Empedrado Lastrado/calle de tierra otro Río/mar Sendero
FBCH_2	Tipo de vivienda	Cualitativa	Nominal	Casa o villa Departamento Mediagua Rancho Cuartos inquilinato Choza Otro
FBCH_3	Material predominante del techo	Cualitativa	Nominal	Asbesto (Eternit) hormigón/losa/cemento otro palma/paja/hoja teja zinc
FBCH_4	Material predominante de las paredes	Cualitativa	Nominal	adobe/tapia asbesto/cemento bahareque (caña, carrizo revestido) caña o estera hormigón/bloque/ladrillo madera otra
FBCH_5	Material predominante del piso	Cualitativa	Nominal	cemento/ladrillo cerámica/baldosa/vinyl tabla/tablón no tratado duela/parquet/tabloncillo/piso flotante tierra mármol/marmetón otro
FBCH_6	De donde obtienen el agua	Cualitativa	Nominal	carro repartidor/triciclo otra fuente por tubería otro pila o llave pública pozo red pública

---

				río/vertiente/acequia
FBCH_7	El agua que recibe la vivienda es	Cualitativa	Nominal	No recibe agua por tubería recibe por tubería dentro de la vivienda recibe por tubería fuera de la vivienda, pero dentro del edificio recibe por tubería fuera del edificio
FBCH_8	El servicio higiénico de la vivienda es:	Cualitativa	Nominal	excusado y alcantarillado excusado y pozo ciego excusado y pozo séptico letrina no tiene
FBCH_9	El servicio de luz eléctrica es:	Cualitativa	Nominal	empresa eléctrica pública Ninguno planta eléctrica privada vela/candil/mechero/gas
FBCH_10	Principalment e cómo eliminan la basura:	Cualitativa	Nominal	botan a la calle/quebrada/río contratan el servicio la entierran la queman otra servicio municipal
FBCH_13	El agua que toman los miembros del hogar:	Cualitativa	Nominal	la beben tal como llega al hogar la hierven no sabe otro tratamiento
FBCH_14	Combustible que utilizan para cocinar	Cualitativa	Nominal	electricidad(inducción) gas leña/carbón no cocina
FBCH_15	¿El servicio higiénico del hogar es exclusivo?	Cualitativa	Nominal	Si No
FBCH_16	¿Disponen de servicio telefónico convencional?	Cualitativa	Nominal	Si No
FBCH_17	¿Algún miembro del hogar tiene telf. celular?	Cualitativa	Nominal	Si No
FBCH_18	La vivienda que ocupa el hogar es:	Cualitativa	Nominal	anticresis y arriendo cedida en arriendo otra propia y la está pagando propia y totalmente pagada recibida por servicios
FSCE_1		Cualitativa	Nominal	Si

<b>Factores Subyacentes</b>					
		Tuvo algún control prenatal			No
FSCE_2	Dónde se hizo el control con mayor frecuencia	Cualitativa	Nominal	establecimiento de salud del MSP clínica/consultorio privado en casa hospital/dispensario del IESS junta de beneficencia seguro social campesino otro	
FSCE_5	En el embarazo le vacunaron contra el tétanos	Cualitativa	Nominal	Si No No sabe/no responde	
FSCE_6	En qué lugar tuvo el parto	Cualitativa	Nominal	establecimiento de salud del MSP clínica/consultorio privado en casa hospital/dispensario del IESS junta de beneficencia seguro social campesino otro	
FSCE_7	Qué persona o profesional le atendió	Cualitativa	Nominal	médico obstetrix familiar comadrona o partera enfermera auxiliar de enfermería otra	
FSCE_8	El parto fue:	Cualitativa	Nominal	cesaria normal	
FSCE_9	El nacimiento fue a los 9 meses o antes de tiempo	Cualitativa	Nominal	a tiempo no sabe posmaduro prematuro	
FSCE_10	Tuvo algún control después del parto	Cualitativa	Nominal	Si No	
FSCE_14	Dónde tuvo el control postparto	Cualitativa	Nominal	establecimiento de salud del MSP clínica/consultorio privado hospital/dispensario del IESS otro, ¿cuál? seguro social campesino hospital ff.aa /policía otro	
FSCN_1	Le pesaron en el momento de nacer	Cualitativa	Nominal	Si No	
FSCN_2		Cualitativa	Nominal	gramos	

	En que unidad de medida fue pesado			kilogramos libras-onzas no sabe
FSCN_5	Con respecto a otros bebes el tamaño de su hijo era:	Cualitativa	Nominal	igual no sabe pequeño muy pequeño más grande
FSCN_6	Tiene el carnet de salud infantil	Cualitativa	Nominal	Si No Si le entregaron, pero se perdió
FSCN_7	Registró el peso al nacer	Cualitativa	Nominal	Si No
FSCN_9	Registró la talla al nacer	Cualitativa	Nominal	Si No
FSCN_11	Registró el perímetro cefálico al nacer	Cualitativa	Nominal	Si No
FSCN_13	El carnet registra puntos en la curva de crecimiento	Cualitativa	Nominal	Si No
FSCCN_1	Después que nació le llevó a control médico	Cualitativa	Nominal	Si No
FSCCN_5	Porqué o para que lo llevó	Cualitativa	Nominal	estaba enfermo para control niño sano no recuerda
FSCCN_6	A qué establecimiento de salud lo llevó	Cualitativa	Nominal	establecimiento de salud del MSP clínica/consultorio privado hospital/dispensario del IESS seguro social campesino otro unidad municipal de salud otro, ¿cuál?
FSCCN_7	Vive con usted actualmente	Cualitativa	Nominal	Si No
FSCCN_21	Grupo de edad	Cualitativa	Ordinal	0-11 12-18 19-23 24-30 31-35 48-59 otro
FSCCN_23	Sexo del niño	Cualitativa	Nominal	Hombre Mujer
FIPA_4		Cualitativa	Nominal	Si



<b>Factores Inmediatos</b>	Fue alimentado con leche materna el día de ayer			No
FIPA_5	A qué tiempo después del nacimiento empezó a mamar o lactar	Cualitativa	Nominal	entre una hora y menos de 24 horas después del parto más de un día menos de una hora
FIPA_6	Los tres primeros días después del nacimiento le dió algo de beber aparte de leche materna	Cualitativa	Nominal	Si No
FIPA_11	Le dio pecho cada vez que le pidió	Cualitativa	Nominal	Si No
FIPA_12	Consumió algún líquido diferente a la leche materna ayer	Cualitativa	Nominal	Si No
FIPA_13	Consumió ayer-agua pura	Cualitativa	Nominal	Si No No sabe/no responde
FIPA_15	Consumió ayer-leche de fórmula	Cualitativa	Nominal	Si No No sabe/no responde
FIPA_17	Consumió ayer-leche en polvo	Cualitativa	Nominal	Si No No sabe/no responde
FIPA_19	Consumió ayer-jugos naturales	Cualitativa	Nominal	Si No No sabe/no responde
FIPA_20	Consumió ayer-sopa	Cualitativa	Nominal	Si No No sabe/no responde
FIPA_22	Consumió ayer-yogurt	Cualitativa	Nominal	Si No No sabe/no responde
FIPA_24	Consumió ayer-colada	Cualitativa	Nominal	Si No No sabe/no responde
FIPA_26	Consumió ayer-gaseosa	Cualitativa	Nominal	Si No No sabe/no responde
FIPA_28	Consumió ayer-agua aromática	Cualitativa	Nominal	Si No No sabe/no responde

FIPA_30	Consumió ayer-cualquier otro líquido	Cualitativa	Nominal	Si No No sabe/no responde
FIPA_32	Comió algún alimento ayer	Cualitativa	Nominal	Si No
FIPA_33	Comió colada espesa de harina de trigo	Cualitativa	Nominal	Si No No sabe/no responde
FIPA_34	Comió colada espesa de granos	Cualitativa	Nominal	Si No No sabe/no responde
FIPA_35	Comió zapallo, zanahoria, pepino	Cualitativa	Nominal	Si No No sabe/no responde
FIPA_36	Comió papa blanca, yuca, camote	Cualitativa	Nominal	Si No No sabe/no responde
FIEI_1	Ha tenido diarrea en las últimas dos semanas	Cualitativa	Nominal	Si No No sabe/no responde
FIEI_4	En las dos últimas semanas ha tenido tos o moquera	Cualitativa	Nominal	Si No

**Fuente:** Diccionario de variables ENSANUT 2018

**Realizado por:** Congacha, Giorgia, 2020

## 2.3 Análisis estadístico

### 2.3.1 Instrumentos de procesamiento y análisis de información

La presente investigación utiliza el software estadístico R (versión 4.0.0) a través de su interfaz RStudio para el análisis, procesamiento y obtención de resultados.

### 2.3.2 Análisis exploratorio de datos

Se importó la información de los archivos .csv: 1\_BDD\_ENS2018\_f1\_personas, 2\_BDD\_ENS2018\_f1\_hogar, 4\_BDD\_ENS2018\_f2\_mef, 5\_BDD\_ENS2018\_f2\_lactancia, 6\_BDD\_ENS2018\_f2\_salud\_ninez, se unieron las bases de datos a través de la función merge (), tomando en cuenta las variables iguales en cada base para su unión : “id\_viv”, “id\_hogar”, “id\_per”, dejando como resultado una sola base de datos.

Considerando la dimensión del conjunto de datos se utilizó el paquete “data.table” (tiene un mejor manejo de memoria RAM), con el cual se filtró y recodificó las variables de interés tomando en

cuenta al factor al que pertenecían (factores Básicos, Subyacentes e Inmediatos), quedando en total de 35937 individuos y 123 variables en estudio.

Con la función `str()` se analizó la estructura de los datos, se transformaron las variables cualitativas "dcronica", "id\_viv", "id\_hogar", "id\_per", "FBCP1\_4", "FBCP1\_6", "FBCP1\_7", "FBCH\_17", "FSCCN\_23", "FBCH\_1", "FBCH\_2", "FBCH\_3", "FBCH\_4", "FBCH\_5", "FBCH\_6", "FBCH\_7", "FBCH\_8", "FBCH\_9", "FBCH\_10", "FBCH\_13", "FBCH\_14", "FBCH\_15", "FBCH\_16", "FBCH\_18", "FBCP1\_2", "FBCP1\_8", "FIPA\_4", "FIPA\_5", "FIPA\_6", "FIPA\_7", "FIPA\_10", "FIPA\_11", "FIPA\_12", "FIPA\_13", "FIPA\_15", "FIPA\_17", "FIPA\_19", "FIPA\_20", "FIPA\_22", "FIPA\_24", "FIPA\_26", "FIPA\_28", "FIPA\_30", "FIPA\_32", "FIPA\_33", "FIPA\_34", "FIPA\_35", "FIPA\_36", "FBCP1\_3", "FSCE\_1", "FSCE\_2", "FSCE\_5", "FSCE\_6", "FSCE\_7", "FSCE\_8", "FSCE\_9", "FSCE\_10", "FSCE\_14", "FSCN\_1", "FSCN\_2", "FSCN\_4", "FSCN\_5", "FSCN\_6", "FSCN\_7", "FSCN\_9", "FSCN\_11", "FSCN\_13", "FSCCN\_1", "FSCCN\_5", "FSCCN\_6", "FSCCN\_7", "FSCCN\_21", "FIEI\_1", "FIEI\_3", "FIEI\_4", "FIEI\_6", "FIEI\_7" a factor, ya que al momento de importar la base de datos, estas variables fueron reconocidas como variables numéricas y carácter. Posteriormente se elaboró un resumen de la base de datos y se encontró datos duplicados por lo que se procedió a eliminarlos, quedando un total de 11231 individuos.

Antes de empezar a preprocesar los datos se dividió la base de datos en un conjunto de entrenamiento (70%) y un conjunto de validación (30%). Para realizar esta división de datos se utilizó la función `createDataPartition()` la cual nos asegura una división aleatoria, manteniendo la proporción de la variable dependiente.

Como la variable dependiente (dcronica) no está balanceada, se procedió a balancear las clases para el conjunto de entrenamiento utilizando el método de submuestreo. Para realizar el balanceo de clases se utilizó la función `ovun.sample()` de la librería ROSE.

### **2.3.3 Preprocesado de datos**

Antes de empezar a modelar, se realizó una serie de transformaciones en los datos para que los modelos mejoren su precisión.

#### **2.3.3.1 Análisis de datos faltantes**

Para realizar el análisis de datos faltantes se utilizó la función `df_status()`, la cual nos proporciona el número de datos faltantes por variable y el porcentaje que este representa con respecto al total de datos, las variables "FBCP1\_2", "FIPA\_7", "FIPA\_8", "FIPA\_9", "FIPA\_10", "FIPA\_16", "FIPA\_18", "FIPA\_23", "FIPA\_25", "FIPA\_27", "FIPA\_29",

“FIPA\_31”, “FSCN\_3”, “FSCN\_4”, “FSCN\_10”, “FSCN\_12”, “FSCN\_14”, “FSCCN\_20”, “FIEL\_2”, “FIEL\_3”, “FIEL\_5”, “FIEL\_6”, “FIEL\_7” presentaron porcentajes superiores al 50% por lo que se decidió eliminarlos de la base de datos. Las variables restantes fueron imputadas por el método kNN vecinos más cercanos mediante la función “knnImputation ()”, la cual utiliza un promedio ponderado de los valores de los vecinos más cercanos para rellenar los datos faltantes. Los pesos vienen dados por  $\exp(-\text{dist}(k, x))$  donde  $\text{dist}(k, x)$  es la distancia euclídea entre el caso con NA’s (x) y el vecino (k), este método se utilizó para las variables cuantitativas, para las variables cualitativas knnImputation imputa los datos a través de la moda.

#### 2.3.3.2 *Centrado y escalado de variables numéricas*

Las variables continuas en estudio presentan distintas magnitudes, unidades y rangos, para nuestros modelos representa esto un problema, puesto que las variables con magnitudes altas pesan más en los cálculos que aquellas variables con magnitudes bajas. Por ello se llevó al mismo nivel todas las variables. Se utilizó la función preProcess () de la librería caret para centrar y escalar el conjunto de datos de entrenamiento y validación.

#### 2.3.3.3 *Variables Dummies*

El tratamiento de variables categóricas es una fase muy importante antes de modelar el conjunto de datos. Una forma de tratarlas es por medio de la binarización de estas variables, la cual consiste en crear tantas variables como niveles tenga la variable categórica, asignando ceros a todas las observaciones excepto a las observaciones que tienen el nivel que representa la nueva variable creada que se le asigna 1. Para la creación de estas variables ficticias se utilizó la función “dummyVars ()” de la librería caret. Como las variables categóricas tenían varios niveles, se creó una variable por cada nivel, cuyo nombre es el nombre de la variable original concatenado con un “.” y el nombre del nivel. El conjunto de entrenamiento como de validación ahora tiene 301 variables.

#### 2.3.3.4 *Variables con varianza cero*

Es necesario eliminar del conjunto de entrenamiento aquellas variables que sus valores en la mayoría de los casos es el mismo o presentan baja variabilidad ya que no aportan información, para ello se utilizó la función “nearZeroVar ()” del paquete caret el cual indica que variables presentan las características antes mencionadas para proceder a eliminarlas. Las variables “FBCP1\_6.afroecuatoriano/afrodescendiente?”, “FBCP1\_6.blanco/a?”, “FBCP1\_6.montuvio/a?”, “FBCP1\_6.mulato/a?”, “FBCP1\_6.negro/a?”, “FBCP1\_6.otra, cuál?(especifique)”, “FBCP1\_7. unión de hecho”, “FBCP1\_7.viudo”, “FBCH\_1.otro, cuál?”, “FBCH\_1.rio/mar”, “FBCH\_2.choza”, “FBCH\_2.covacha”, “FBCH\_2.cuarto/s en casa de inquilinato”, “FBCH\_2.otra, cuál?”,

"FBCH\_3.otro, cuál?", "FBCH\_3.palma/paja/hoja?", "FBCH\_4.adobe/tapia?", "FBCH\_4.bahareque (caña, carrizo revestido)?", "FBCH\_4.caña o estera?", "FBCH\_4.otra, cuál?", "FBCH\_5.caña?", "FBCH\_5.mármol/marmeton?", "FBCH\_5.otro, cuál?", "FBCH\_5.tierra?", "FBCH\_6.carrorepartidor/triciclo?", "FBCH\_6.otro, cuál?", "FBCH\_6.pila o llave pública?", "FBCH\_7.por tubería fuera del edificio, lote o terreno?", "FBCH\_8.letrina?", "FBCH\_9.empresa eléctrica pública?", "FBCH\_9.ninguno?", "FBCH\_9.planta eléctrica privada?", "FBCH\_9.vela, candil, mechero, gas?", "FBCH\_10.botan a la calle/ quebrada/ río?", "FBCH\_10.contratan el servicio?", "FBCH\_10.la entierran?", "FBCH\_10.otra, cuál?", "FBCH\_13.no sabe", "FBCH\_13.otro tratamiento?", "FBCH\_14.electricidad? (inducción)", "FBCH\_14.no cocina", "FBCH\_18.anticresis y arriendo?", "FBCH\_18.otra, cuál?", "FBCH\_18.recibida por servicios?", "FBCH\_8.Ninguno o Centro de Alfabetización", "FIPA\_13.no sabe / no responde", "FIPA\_15.no sabe / no responde", "FIPA\_17.no sabe / no responde", "FIPA\_20.no sabe / no responde", "FIPA\_22.no sabe / no responde", "FIPA\_26.no sabe / no responde", "FIPA\_30.no", "FIPA\_30.no sabe / no responde", "FIPA\_30.si", "FIPA\_34.no sabe / no responde", "FIPA\_35.no sabe / no responde", "FBCH\_3.no quería tener hijos?", "FSCE\_1.no", "FSCE\_1.si", "FSCE\_2.consejo provincial/unidad municipal de salud", "FSCE\_2.fundación/ ong", "FSCE\_2.hospital ff.aa/policia", "FSCE\_2.junta de beneficencia", "FSCE\_2.no recuerda", "FSCE\_2.otro, cuál?", "FSCE\_2.partera", "FSCE\_2.seguro social campesino", "FSCE\_5.no sabe / no responde", "FSCE\_6.consejo provincial/unidad municipal de salud", "FSCE\_6.fundación/ ong", "FSCE\_6.hospital ff.aa/ policia", "FSCE\_6.junta de beneficencia", "FSCE\_6.otro, cuál?", "FSCE\_6.seguro social campesino", "FSCE\_7.aux. enfermería", "FSCE\_7.comadrona o partera", "FSCE\_7.enfermera", "FSCE\_7.familiar", "FSCE\_7.otro, cuál?", "FSCE\_7.usted misma", "FSCE\_9.no sabe", "FSCE\_9.posmaduro", "FSCE\_14.fundación/ ong", "FSCE\_14.hospital ff.aa/policia", "FSCE\_14.hospital/clínica/dispensario del iess", "FSCE\_14.junta de beneficencia", "FSCE\_14.no recuerda", "FSCE\_14.otro, cuál?", "FSCE\_14.partera", "FSCE\_14.seguro social campesino", "FSCN\_1.no", "FSCN\_1.si", "FSCN\_5.muy pequeño", "FSCN\_5.no sabe", "FSCCN\_1.no", "FSCCN\_1.si", "FSCCN\_5.no recuerda", "FSCCN\_6.consejo provincial/unidad municipal de salud", "FSCCN\_6.fundación/ ong", "FSCCN\_6.hospital ff.aa/policia", "FSCCN\_6.junta de beneficencia", "FSCCN\_6.no recuerda", "FSCCN\_6.otro, cuál?", "FSCCN\_6.partera", "FSCCN\_6.seguro social campesino", "FSCCN\_7.no", "FSCCN\_7.si", "FSCCN\_21.36-42", "FSCCN\_21.43-47", "FSCCN\_21.48-59", "FIEI\_1.no sabe / no responde", "FSCCN\_21.36-42", "FIPA\_28.no sabe / no responde;

presentaron varianza cero por lo cual se procedió a eliminarlas quedando en total 190 variables para el estudio.

## 2.3.4 Construcción de modelos de clasificación

### 2.3.4.1 Árboles de decisión

La técnica de árboles de clasificación construye el modelo seleccionando aleatoriamente un subconjunto de variables, de entre ellas escoge la que divide de manera significativa a la muestra en cada uno de los nodos de clasificación que se van formando, su objetivo principal es minimizar la tasa de error de clasificación asignándole a cada observación la clase más común en su región del espacio de predictores.

Para la construcción del modelo se utilizaron los paquetes “rpart” y “rpart. plot”, el modelo fue realizado sobre el conjunto de entrenamiento. Se utilizaron todas las variables disponibles para predecir la variable “dcronica”, además se añadió el parámetro “control” de la función rpart para que la división mejore el ajuste por lo menos en un 0.001, creándose un árbol muy profundo, para poder solventar este problema de sobreajuste se usó el árbol utilizando la CP óptima, es decir se buscó la CP que minimice el error promedio obtenido de la validación cruzada. El árbol resultante se presentó en la parte de resultados.

Con la finalidad de reducir la variabilidad y aumentar el poder predictivo del modelo de árboles de decisión se aplicó el método de ensemble Gradient Boosting, a través del algoritmo XGBoost, se escogió este algoritmo por aumentar la precisión de predicción frente a otros modelos de árboles secuenciales, recordando que los modelos de Boosting se basan en crear modelos de árboles de forma secuencial, donde los árboles aprenden del error del modelo anterior. Para la realización de este modelo se utilizó el paquete “xgboost”, el modelo fue realizado sobre el conjunto de entrenamiento. Lo primero que se realizó fue buscar los parámetros óptimos para el modelo de Boosting, empezando por la profundidad; para ello se realizó una búsqueda en cuadrícula asignando los valores 3, 4, y 5 en el parámetro “interaction. depth”, no se asignó más valores, tomando en cuenta que cuantas más opciones le demos, más tiempo tardará en entrenar el modelo. La profundidad óptima se obtuvo en 4. La búsqueda del resto de hiperparámetros se realizó de forma conjunta, tomando en cuenta la profundidad óptima. Una vez obtenidos los parámetros óptimos se creó el modelo Boosting tomando en cuenta estos valores, además se añadió en el parámetro “distribution” = “bernoulli”, para indicar al modelo que estamos ante una clasificación binaria, además se entrenó al modelo con 3000 iteraciones. Las variables más importantes en el modelo fueron determinadas a través del método de influencia relativa, el cual calcula la mejora en promedio realizada por cada variable en todos los árboles en los que se utiliza la variable.

#### 2.3.4.2 Regresión logística binaria

A diferencia de la técnica de árboles de clasificación, los modelos de regresión logística binaria no presentan aleatoriedad intrínseca.

Para realizar el modelo de regresión logística se utilizó la función “glm ()”, en el cual se especificó la familia a la que pertenece “binomial”, especificando la función link  $\log\left(\frac{\pi}{1-\pi}\right)$  de  $\pi$ , donde se especifica que se va a utilizar el modelo *logit*, de modo que así se asegura que no exista ningún problema estructural respecto al posible rango de valores de  $\pi$ .

El método de selección de variables para el modelo fue por pasos hacia atrás, el cual comienza incluyendo todos los predictores del modelo y va eliminando iterativamente los predictores que no aportan significativamente al modelo. El método por pasos hacia atrás se detiene cuando todos los predictores del modelo son estadísticamente significativos.

#### 2.3.4.3 Evaluación de modelos

La evaluación de los modelos a través de medidas de bondad de ajuste nos permite comparar los modelos y poder seleccionar el modelo que tenga mayor precisión. Para la evaluación de los modelos implementados se utilizó la tasa de error de los modelos a través de la matriz de confusión utilizando la función “confusionMatrix ()” de la librería “caret” y el área bajo la curva ROC a través del AUC, para este último se utilizaron las librerías “ROCR” y “pROC”.

#### 2.3.5 Construcción de la aplicación web interactiva

La aplicación web interactiva tiene como finalidad implementar los modelos realizados en la investigación (regresión logística y arboles de decision) para la base de datos analizada (ENSANUT 2018). Para la realización de la aplicación se utilizó el paquete “shiny” el cual permite la creación de aplicaciones web interactivas y de fácil desarrollo, las cuales pueden ser utilizadas desde un ordenador, Tablet o incluso desde un teléfono móvil; para mejorar su apariencia se utilizó la librería “shinydashboard”.

Los requerimientos técnicos para el uso de la aplicación son: contar con acceso a internet, tener instalado un navegador ya sea Microsoft Internet Explorer [versión 7.0 o superior], Chrome o Mozilla Firefox [versión 3.0 o superior].

El ingreso a la aplicación se realiza desde la siguiente dirección electrónica: [https://giorgiacongacha.shinyapps.io/MODELOS\\_CLASIFICACION/?\\_ga=2.25840125.1129436373.1608525416-872480578.1575579265](https://giorgiacongacha.shinyapps.io/MODELOS_CLASIFICACION/?_ga=2.25840125.1129436373.1608525416-872480578.1575579265)

Al ingresar en la dirección electrónica se visualiza en la parte izquierda un menú con 2 opciones “Modelo logit” o “Arboles de decision”, al dar click en la primera opción se despliega una serie de sub opciones “Datos”, “Resultados”, “Predicciones”. En el sub ítem “Datos” se puede observar la base de datos de la ENSANUT con la cual se entrenará el modelo, en “Resultados se puede observar el modelo obtenido a través de Regresión logística” y por último en la sección predicciones se puede apreciar las predicciones de los nuevos datos ingresados.

Por otra parte, al seleccionar la opción “Arboles de decision” se despliegan 3 sub opciones “Datos”, “Resultados” y “Gráfica”. En el sub ítem “Datos” se observa la base de datos de la ENSANUT que se utilizó para el entrenamiento del modelo, en el sub ítem “Resultados” se observa el modelo obtenido a través de árboles de decisión, en el sub ítem “Gráfica” se puede visualizar gráficamente el modelo del árbol de decisión. La interfaz de usuario se la puede observar en el Anexo H.



## CAPÍTULO III

### 3 MARCO DE RESULTADOS Y DISCUSIÓN DE LOS RESULTADOS

#### 3.1 Análisis exploratorio univariado

**Tabla 1-3:** Análisis univariado variable dependiente

Variable	Categoría	Porcentaje (%)
dcronica	1 (Desnutrido Crónico)	27,97
	0 (No Desnutrido Crónico)	72,02

Fuente: ENSANUT 2018

Realizado por: Congacha, Giorgia, 2020

En la **Tabla 1-3** se observa que del total de niños analizados en la ENSANUT 2018, el 27,97% presenta desnutrición crónica, mientras que el 72,02% no la presenta, es decir 1 de cada 4 niños tiene desnutrición crónica.

Como el número de variables independientes es demasiado extenso (190), se presenta el análisis univariado de las variables significativas en los modelos planteados distribuidos por factores.

**Tabla 2-3:** Análisis univariado variables independientes cualitativas (Factores Básicos)

Variable	Categoría	Porcentaje (%)
FBCP1_6	Mestizo/a	75,55
	Indígena	14,5
	Montuvio/a	4,25
	Blanco/a	1,34
	Afroecuatoriano/a	1,39
	Negro/a	1,42
	Otro	0,04
	Casado	28,67
FBCP1_7	Divorciado	0,98
	Separado	6,76
	Soltero	16,98
	Unión de hecho	1,29
	Unión libre	44,7
FBCP1_8	Viudo	0,6
	Ninguno o Centro de Alfabetización	0,99
	Educación Básica	37,17
	Educación Media/Bachillerato	42,76
FBCH_1	Superior	19,08
	Calle pavimentada o adoquinada	49,02
	Empedrado	8,41
	Lastrado/calle de tierra	34,83

	otro	0,08
	Río/mar	0,54
	Sendero	7,11
	Casa o villa	66,54
	Departamento	12,65
	Mediagua	10,41
<b>FBCH_2</b>	Rancho	6,21
	Cuartos inquilinato	3,02
	Choza	0,83
	Otro	0,02
	Asbesto (Eternit)	11,08
	hormigón/losa/cemento	28,95
<b>FBCH_3</b>	otro	0,37
	palma/paja/hoja	0,84
	teja	5,39
	zinc	53,39
	cemento/ladrillo	38,5
	cerámica/baldosa/vinyl	32,08
	tabla/tablon no tratado	17,09
<b>FBCH_5</b>	duela/parquet/tabloncillo/piso flotante	7,19
	tierra	3,85
	mármol/marmetón	0,71
	otro	0,009
	electricidad(inducción)	1,76
<b>FBCH_14</b>	gas	92,24
	leña/carbón	5,91
	no cocina	0,09
<b>FBCH_15</b>	Si	12,35
	No	87,65
	anticresis y arriendo	1,57
	cedida	19,44
	en arriendo	22,91
<b>FBCH_18</b>	Otra	0,14
	propia y la está pagando	6,45
	propia y totalmente pagada	47,97
	recibida por servicios	1,52

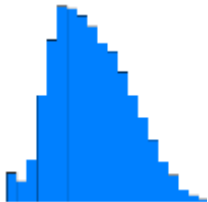

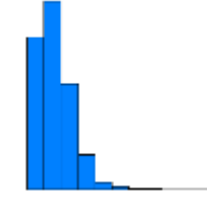
Fuente: ENSANUT 2018

Realizado por: Congacha, Giorgia, 2020

En la **Tabla 2-3** se indican las características predominantes de los individuos en estudio (características de la pareja y características del hogar), de los cuales podemos decir que: el 75.55% pertenece al grupo étnico mestizo , el 44.7% se encuentra en unión libre, el 42.76% de madres tienen un nivel de educación media Bachillerato, el 49.02% de viviendas tiene una vía de acceso principal de característica pavimentada o adoquinada, el 66.54% de individuos viven en casas o villas, el material predominante del techo es zinc en un 53.39%, el material predominante

del piso es cemento/ladrillo en un 38.5%, el 92.24% de viviendas utilizan gas como combustible para cocinar, el servicio higiénico del hogar no es exclusivo en un 87.65% de los hogares, la vivienda que ocupan el 47.97% de los hogares es propia y totalmente pagada.

**Tabla 3-3:** Análisis univariado variables independientes cuantitativas (Factores Básicos)

Variable		Estadísticas	Grafico	
<b>FBCPI_1</b>	Años cumplidos de la madre	Media	27,26	
		Mediana	27	
		Des. estandar	7,39	
		Kurtosis	-0,4	
		Asimetría	0,28	
		CV	27.10%	
<b>FBCPI_5</b>	Número de hijos nacidos vivos	Media	2,28	
		Mediana	2	
		Des. estandar	1,44	
		Kurtosis	5,27	
		Asimetría	1,84	
		CV	63.27%	
<b>F BCH_1 2</b>	Número de dormitorios de la vivienda	Media	2,066	
		Mediana	2	
		Des. estandar	1,129	
		Kurtosis	2,188	
		Asimetría	0,76	
		CV	54.68%	

Fuente: ENSANUT 2018

Realizado por: Congacha, Giorgia, 2020

En la **Tabla 3-3** se analizan las variables cuantitativas más representativas pertenecientes a Factores Básicos de los individuos en estudio de los cuales podemos decir que: el promedio de edad de las madres en estudio es aproximadamente 27 años, observando el histograma se ve que corresponde a una distribución normal. Dado que el coeficiente de variación es del 27.10% se puede concluir que los datos son homogéneos.

El número promedio de hijos nacidos vivos es aproximadamente 2, la misma que tiene una asimetría positiva como lo indica su histograma, su coeficiente de variación es 63.27% lo que indica que los datos no son homogéneos.

El promedio de dormitorios de la vivienda es 2, se observa que presenta una asimetría positiva en el histograma; el coeficiente de variación es 54.68% lo que indica que sus datos no son homogéneos.

**Tabla 4-3:** Análisis univariado variables independientes cualitativas (Factores Subyacentes)

<b>Variable</b>	<b>Categoría</b>	<b>Porcentaje (%)</b>
<b>FSCE_5</b>	Si	90,54
	No	5,92
	No sabe/no responde	3,54
<b>FSCE_9</b>	a tiempo	85,14
	no sabe	0,33
	posmaduro	2,69
	prematureo	11,83
	establecimiento de salud del MSP	76,91
<b>FSCE_14</b>	clínica/consultorio privado	11,51
	hospital/dispensario del IESS	7,77
	otro, ¿cuál?	1,021
	seguro social campesino	0,94
	hospital ff.aa/policía	0,59
	otro	0,18
	igual	57,49
<b>FSCN_5</b>	no sabe	4,89
	pequeño	14,96
	muy pequeño	4,18
	más grande	18,48
	establecimiento de salud del MSP	82,92
<b>FSCCN_6</b>	clínica/consultorio privado	9,83
	hospital/dispensario del IESS	5,81
	seguro social campesino	0,28
	otro	0,2
	unidad municipal de salud	0,28
<b>FSCCN_21</b>	otro, ¿cuál?	0,046
	0-11	27,74
	12-18	19,86
	19-23	12,87
	24-30	17,87
	31-35	12,76
	36-42	2,16
	43-47	1,93
	48-59	4,79

Fuente: ENSANUT 2018

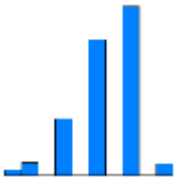
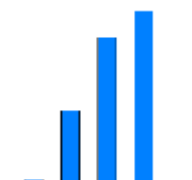
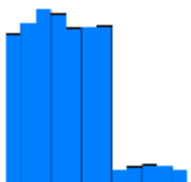
Realizado por: Congacha, Giorgia, 2020

En la **Tabla 4-3** se indican las características del embarazo, nacimiento y cuidados del niño pertenecientes a Factores Subyacentes, las más importantes son: en el embarazo el 90.54% de mujeres si fueron vacunadas contra el tétanos, el 85% de los nacimientos fue a los 9 meses. El 76.91% de madres tuvieron su control postparto en un establecimiento de salud del MSP. Con respecto a otros bebés el tamaño de los niños era igual en un 57.49%. El 82.92% de los niños

tuvieron su primer control médico en un establecimiento de salud del MSP. El 65.47% de niños analizados en la presente investigación están en edades comprendidas de 0 a 30 meses.

**Tabla 5-3:** Análisis univariado variables independientes cuantitativas (Factores Subyacentes)

Variable		Estadísticas	Grafico	
<b>FSCE_3</b>	Meses de embarazo cuando se hizo el primer control	Media	7,64	
		Mediana	6	
		Des. estandar	5,35	
		Kurtosis	4,94	
		Asimetría	1,84	
		CV	70.12%	
<b>FSCE_4</b>	Cuántos controles tuvo antes del parto	Media	7,09	
		Mediana	7	
		Des. estandar	2,86	
		Kurtosis	64,88	
		Asimetría	2,99	
CV	40.35%			
<b>FSCE_12</b>	Primer control postparto-semanas	Media	0,75	
		Mediana	1	
		Des. estandar	0,88	
		Kurtosis	20,72	
		Asimetría	2,18	
CV	118.44%			
<b>FSCE_13</b>	Primer control postparto-meses	Media	0,59	
		Mediana	0	
		Des. estandar	1,21	
		Kurtosis	25,19	
		Asimetría	4,46	
CV	203.44%			
<b>FSCCN_9</b>	hepatitis b-dosis	Media	16,99	
		Mediana	17	
		Des. estandar	1,04	
		Kurtosis	0,5	
		Asimetría	-0,78	
CV	6.09%			
<b>FSCCN_10</b>	pentavalente 1-dosis	Media	17,06	
		Mediana	17	
		Des. estandar	0,98	
		Kurtosis	0,46	
		Asimetría	-0,74	
CV	5.74%			

<b>FSCCN_12</b>	pentavalente 3-dosis	Media	17,25	
		Mediana	17	
		Des. estandar	0,93	
		Kurtosis	0,85	
		Asimetría	-0,88	
		CV	5,42%	
<b>FSCCN_19</b>	neumococo 2-dosis	Media	17,15	
		Mediana	17	
		Des. estandar	0,96	
		Kurtosis	0,59	
		Asimetría	-0,82	
		CV	5,60%	
<b>FSCCN_22</b>	Edad (en meses)	Media	20,63	
		Mediana	19	
		Des. estandar	12,99	
		Mínimo	0	
		Máximo	59	
		CV	62,95%	

Fuente: ENSANUT 2018

Realizado por: Congacha, Georgia, 2020

En la **Tabla 5-3** se presentan las variables cuantitativas pertenecientes a Factores Subyacentes. En promedio, las mujeres se realizan el primer control prenatal aproximadamente a los 8 meses, los controles que se realizan antes del parto en promedio son 7 veces. El promedio de controles postparto en semanas es 1 vez, correspondiente en meses a 0. De acuerdo al coeficiente de asimetría los controles pre y post parto son positivos como se lo corrobora observando los histogramas, los coeficientes de variación presentan alta variabilidad indicando la no homogeneidad de los mismos.

El número de dosis puestas al año para el control de hepatitis b, pentavalente 1, pentavalente 3 y neumococo 2 en promedio es de 17 con una baja variabilidad como lo indica su coeficiente de variación, de acuerdo a su asimetría esta es negativa y se observa en sus diagramas de barras.

La edad en meses de los niños analizados en la presente investigación se encuentra entre 0 y 59 meses, es decir de 0 a 5 años, notándose que el 75% de los niños tienen una edad inferior a 29 meses (3 años).

**Tabla 6-3:** Análisis univariado variables independientes cualitativas (Factores Inmediatos)

Variable	Categoría	Porcentaje (%)
<b>FIPA_5</b>	entre una hora y menos de 24 horas	16,23
	después del parto	58,86
	más de un día	7,38

	menos de una hora	17,52
<b>FIPA_6</b>	Si	27,96
	No	72,04
<b>FIPA_19</b>	Si	68,64
	No	31,12
<b>FIPA_20</b>	No sabe/no responde	0,24
	Si	83,96
	No	15,84
<b>FIPA_33</b>	No sabe/no responde	0,21
	Si	68,3
	No	31,49
<b>FIPA_36</b>	No sabe/no responde	0,21
	Si	66,45
	No	33,22
<b>FIEI_1</b>	No sabe/no responde	0,33
	Si	13,52
	No	85,74
	No sabe/no responde	0,75

Fuente: ENSANUT 2018

Realizado por: Congacha, Giorgia, 2020

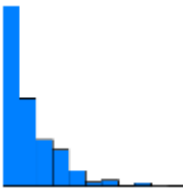
En la **Tabla 6-3** se describen características sobre patrón alimentario y enfermedades infecciosas pertenecientes a Factores Inmediatos, las cuales se detallan a continuación:

El 58.86% de los niños empezó a lactar inmediatamente después del parto, los tres primeros días después del nacimiento el 72.04% de niños no bebió ningún líquido a parte de leche materna, el 68.64% de niños consumió jugos naturales un día antes de ser encuestado, el 83.96% de los niños también consumió sopa, el 68.3% de niños consume colada espesa de harina de trigo, el 66.45% consume papa blanca, yuca o camote.

En cuanto a enfermedades infecciosas el 85.74% de niños no tuvo diarrea dos semanas antes de ser encuestado

**Tabla 7-3:** Análisis univariado variables independientes cuantitativas (Factores Inmediatos)

Variable		Estadísticas		Grafico	
<b>FIPA_2</b>	Hasta qué edad le dio el seno- meses	Media	31,58		
		Mediana	8		
		Des. estandar	31,83		
		Mínimo	0		
		Máximo	77		
		CV	100.78%		

<b>FIPA_14</b>	Consumió ayer -agua pura - cuántas veces	Media	2,93	
		Mediana	3	
		Des. estandar	1,74	
		Kurtosis	2,08	
		Asimetría	1,24	
		CV	59,47%	

Fuente: ENSANUT 2018

Realizado por: Congacha, Giorgia, 2020

En la **Tabla 7-3** se presentan las variables cuantitativas más significativas en los modelos planteados pertenecientes a Factores Inmediatos, de las cuales se puede decir que: la edad de lactancia materna máxima medida en meses, presenta datos atípicos, estos pueden ser observados en el histograma, la edad mínima que presenta esta variable es 0 lo que indica que hay niños que nunca recibieron leche materna su máximo es 77 meses, es decir existen niños que consumen leche materna pasando los 5 años. El coeficiente de variación indica una alta variabilidad en los datos.

El promedio de consumo de agua pura en niños menores de cinco años es de tres veces al día, con tendencia positiva, es decir toman agua pura más de tres veces al día y presenta menos variabilidad que la edad de lactancia materna.

### 3.2 Modelos de Clasificación: Regresión logística

El modelo de regresión logística se aplicó con la finalidad de predecir e identificar las variables que influyen en la desnutrición crónica infantil. Para la estimación del modelo se procedió a utilizar solo la muestra de entrenamiento.

Después de seleccionar las variables más significativas para el modelo de regresión logística a través del método selección por pasos hacia atrás, el modelo resultante se presenta a continuación:

**Tabla 8-3:** Modelo de regresión logística binaria

	<b>B</b>	<b>Error estándar</b>	<b>valor z</b>	<b>Pr(&gt; z )</b>	<b>exp(B)</b>
<b>(Intercepto)</b>	-0,925	0,434	-2,132	0,033	0,396
<b>FBCP1_6. indígena.</b>	0,476	0,089	5,351	0,000	1,608
<b>FBCP1_7. separado</b>	0,193	0,082	2,357	0,018	1,213
<b>FBCH_17.no</b>	0,207	0,069	3,008	0,003	1,229
<b>FBCH_12</b>	-0,073	0,028	-2,612	0,009	0,929
<b>FBCP1_1</b>	-0,013	0,005	-2,388	0,017	0,987
<b>FBCP1_5</b>	0,095	0,029	3,264	0,001	1,099
<b>FBCP1_8. Educación. Media. Bachillerato</b>	-0,164	0,071	-2,302	0,021	0,848
<b>FBCP1_8. Superior</b>	-0,277	0,100	-2,775	0,006	0,758
<b>FIPA_33.no.sabe...no. responde</b>	-0,330	0,083	-3,981	0,000	0,719



<b>FSCE_3</b>	0,015	0,006	2,632	0,008	1,015
<b>FSCE_12</b>	0,109	0,038	2,856	0,004	1,115
<b>FSCN_5. más. grande</b>	-0,402	0,084	-4,761	0,000	0,668
<b>FSCN_5. pequeño</b>	0,271	0,086	3,148	0,002	1,311
<b>FSCCN_21.19.23</b>	0,430	0,090	4,757	0,000	1,537
<b>FSCCN_21.31.35</b>	0,339	0,093	3,646	0,000	1,404
<b>FIEI_1.si</b>	0,842	0,416	2,024	0,043	2,321

Fuente: ENSANUT 2018

Realizado por: Congacha, Giorgia, 2020

De acuerdo a la **Tabla 8-3** se puede observar que las variables en mención resultaron significativas ( $p$ -valor  $<0.05$ ) de las cuales se puede decir lo siguiente:

- (FBCP1\_6. indígena) La variable grupo étnico indígena tiene una relación positiva con la variable dependiente, es decir si el niño es indígena tiene 60.8% mayor probabilidad de tener desnutrición crónica que niños de otros grupos étnicos.
- (FBCP1\_7. separado) La probabilidad de que el niño tenga desnutrición crónica es 21.3% mayor si los padres del niño están separados.
- (FBCH\_17.no) Si algún miembro del hogar no tiene teléfono celular, el niño tiene 22.9% mayor probabilidad de tener desnutrición crónica.
- (FBCH\_12) La variable número de dormitorios de la vivienda tiene una relación negativa con la variable dependiente es decir a medida que el número de dormitorios crezca la probabilidad de tener desnutrición crónica disminuye en un 7.1%.
- (FBCP1\_1) Por cada año adicional cumplido de la madre, la probabilidad de que el niño tenga desnutrición crónica disminuye en un 1.3%.
- (FBCP1\_5) La probabilidad de que el niño tenga desnutrición crónica es 9.9% mayor si el número de hijos nacidos vivos en el hogar aumenta.
- (FBCP1\_8. Educación. Media. Bachillerato) Si la madre tiene un nivel de escolaridad Media Bachillerato la probabilidad de que el niño tenga desnutrición crónica disminuirá en un 15.2%.
- (FBCP1\_8. Superior) Si la madre tiene un nivel de escolaridad Superior la probabilidad de que el niño tenga desnutrición crónica disminuirá en un 24.2%.
- (FIPA\_33.no.sabe...no. responde) Si la madre no sabe que su hijo consume colada espesa de harina de trigo la probabilidad de que el niño tenga desnutrición crónica disminuirá en un 28.1%.
- (FSCE\_3) Si la madre se tarda en realizar su primer control de embarazo la probabilidad de que el niño tenga desnutrición crónica aumentará en un 1.5%.

- (FSCE\_12) Si la madre se demora en realizar su primer control postparto, la probabilidad de que el niño tenga desnutrición crónica aumenta en un 11.5%
- (FSCN\_5. más. grande) Si el niño es más grande con respecto a otros bebés de su misma edad la probabilidad de que el niño tenga desnutrición crónica disminuye en un 33.2%
- (FSCN\_5. pequeño) Por el contrario si el niño resulta ser más pequeño en relación a otros niños de su misma edad, la probabilidad de que el niño tenga desnutrición crónica aumenta en un 31.1%
- (FSCCN\_21.19.23) Si el niño pertenece al grupo de edad de entre 19 a 23 meses la probabilidad de que tenga desnutrición crónica aumenta en un 53.7%
- (FSCCN\_21.31.35) Si el niño pertenece al grupo de edad de entre 31 a 35 meses la probabilidad de que tenga desnutrición crónica aumenta en un 40.4%
- (FIEI\_1.si) Si el niño tiene diarrea en las dos últimas semanas la probabilidad de que tenga desnutrición crónica se duplica en relación a los niños que no la presentan.

### 3.2.1 *Significatividad del Modelo*

$H_0$  : El modelo no es significativo

$H_1$  : El modelo es significativo

**Tabla 9-3:** Prueba ómnibus de la significancia del modelo

	<b>Chi-cuadrado</b>	<b>gl</b>	<b>Sig.</b>
<b>Modelo</b>	263,5353	18	1,43E-45

Fuente: Elaboración propia

**Realizado por:** Congacha, Giorgia, 2020

Para contrastar la significatividad global del modelo, se utilizó el estadístico de Razón de Verosimilitud (Prueba Ómnibus).

Se muestra un valor p ( $1,43E-45 < 0.05$ ), lo que indica que hay una relación significativa entre las variables independientes y la variable dependiente (dcronica), es decir, el modelo es significativo.

### 3.2.2 *Matriz de confusión*

El modelo construido mediante la técnica de regresión logística binaria presentó un error total de predicción de 35%, la matriz de confusión asociada a este modelo se presenta a continuación.

**Tabla 10-4:** Matriz de confusión regresión logística

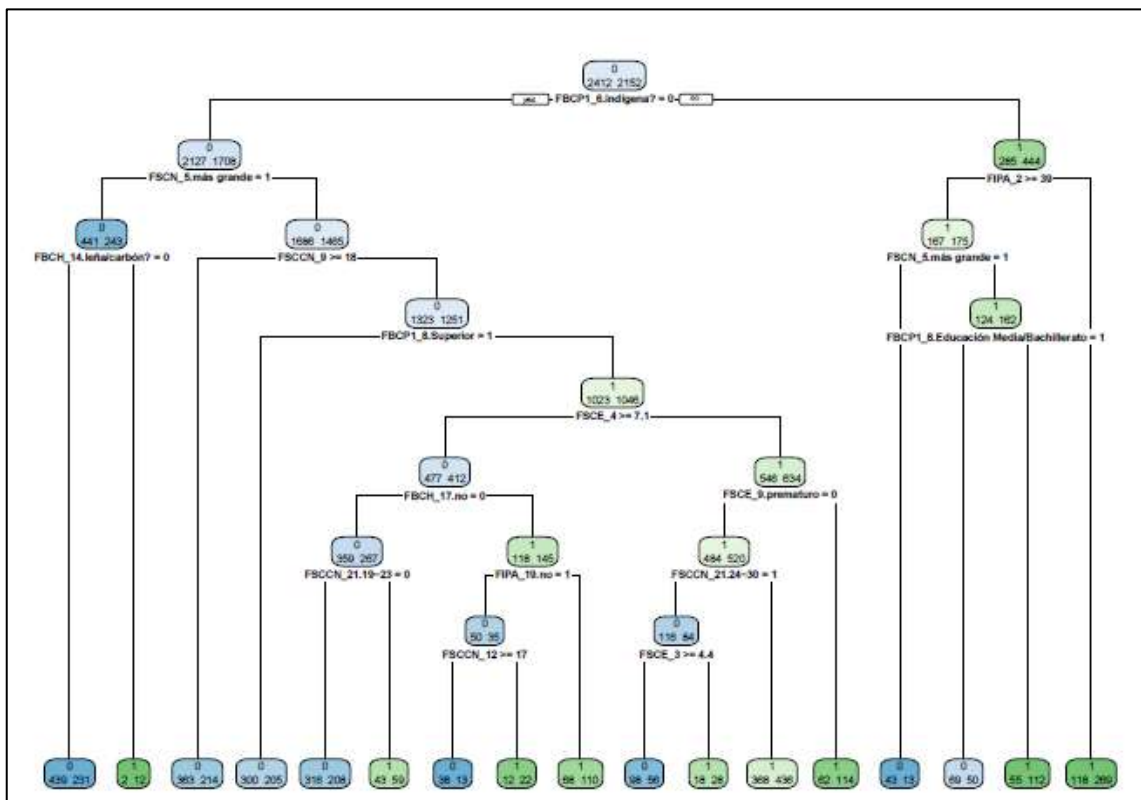
		Desnutrición crónica menores de 5 años	
		No	Si
Desnutrición crónica menores de 5 años	No	1774	719
	Si	473	403

Fuente: Elaboración propia

Realizado por: Congacha, Georgia, 2020

### 3.3 Modelo de clasificación: Árboles de decisión

Después de realizar el modelo con un CP = 0.001 para una mayor profundidad del árbol se procede a realizar la poda del árbol en función de la CP que minimice el error promedio obtenido de la validación cruzada. El árbol resultante se muestra a continuación:



**Gráfico 1-3:** Árbol de clasificación

Realizado por: Congacha, Georgia, 2020

Como se observa en el **gráfico1-3** el árbol presenta 15 nodos intermedios y 17 nodos terminales. De las 4564 observaciones utilizadas en el conjunto de entrenamiento 2152 (47.15%) pertenecen a la clase 1 (el niño presenta desnutrición crónica) y 2412 (52.85%) pertenecen a la clase 0 (el niño no presenta desnutrición crónica).

La variable más importante se encuentra en el nodo raíz la cual es grupo étnico de la categoría indígena (FBCP1\_6. indígena) en esta se realiza una primera división. Si FBCP1\_6. indígena es mayor o igual a 0.5 se clasifican 729 observaciones de las cuales 285 (39.1%) niños no presentan desnutrición crónica y 444 (60.9%) niños si la presentan; al no ser un nodo terminal se analiza la siguiente variable más importante en el modelo: hasta que edad le dió el seno – meses (FIPA\_2).

Si (FIPA\_2) es menor a 39 entonces se clasifican 387 observaciones de las cuales 118 (30.49%) pertenecen a la clase 0 y 269 (69.51%) pertenecen a la clase 1 llegando a un nodo terminal, siendo la clase mayoritaria 1 (el niño presenta desnutrición crónica).

Si (FIPA\_2) es mayor a 39 se clasifican 342 observaciones de las cuales 167 (48.83%) pertenecen a la clase 0 y 175 (51.17%) pertenecen a la clase 1, siendo esta última la clase mayoritaria; al no ser un nodo terminal analizamos la siguiente variable de importancia en el modelo: con respecto a otros bebés el tamaño del niño es más grande (FSCN\_5. más. grande).

Si (FSCN\_5. más. grande) es mayor o igual a 0.5 se clasifican 56 observaciones de las cuales 13 (23.21%) pertenecen a la clase 1 y 43 (76.79%) pertenecen a la clase 0, llegando a un nodo terminal. Por el contrario, si (FSCN\_5. más. grande) es menor a 0.5 se clasifican 286 observaciones de las cuales 124 (43.36%) pertenecen a la clase 0 y 162 (56.64%) pertenecen a la clase 1, siendo esta la clase mayoritaria, al no ser un nodo terminal se examina la siguiente variable de interés escolaridad de la madre (FBCP1\_8. Educación Media/Bachillerato).

Si (FBCP1\_8. Educación Media/Bachillerato) es mayor o igual a 0.5 se clasifican 119 observaciones de las cuales 69 (57.98%) pertenecen a la clase 0 y 50 (42.01%) pertenecen a la clase 1 llegando a un nodo terminal, si (FBCP1\_8. Educación Media/Bachillerato) es menor a 0.5 se clasifican 167 observaciones de las cuales 55 (32.93%) pertenecen a la clase 0 y 112 (67%) pertenecen a la clase 1, llegando a un nodo terminal.

Cuando FBCP1\_6.indígena es menor que 0.5 se clasifican las 3835 observaciones de las cuales 1708 (44.54%) corresponden a niños que tienen desnutrición crónica y 2127 (55.46%) no la presentan, al no ser un nodo terminal se analiza la siguiente variable en importancia: con respecto a otros bebés el tamaño del niño es más grande (FSCN\_5.más.grande) si esta variable es mayor o igual a 0.5 entonces se clasifican 684 observaciones de los cuales 441 (64.47%) observaciones pertenecen a la clase 0 (el niño no presenta desnutrición crónica) y 243 (35.52%) pertenecen a la clase 1 (el niño presenta desnutrición crónica), al no ser un nodo terminal pasamos a la siguiente variable de importancia combustible que utilizan para cocinar (FBCH\_14.leña/carbón) si esta es menor a 0.5 se clasifican 670 observaciones de las cuales 231 (34.47%) pertenecen a la clase 1 y 439 (65.52%) pertenecen a la clase 0 llegando a un nodo terminal. Si (FBCH\_14. leña/carbón) es

mayor o igual a 0.5 se clasifican 14 observaciones de las cuales 2 (14.29%) pertenecen a la clase 0 y 12 (85.71%) pertenecen a la clase 1 llegando a un nodo terminal.

Si (FSCN\_5. más.grande) es menor a 0.5 se clasifican 3151 observaciones de las cuales 1686 (53.51%) observaciones pertenecen a la clase 0 y 1465 (46.49%) pertenecen a la clase 1, al no ser un nodo terminal se analiza la variable hepatitis b-dosis (FSCCN\_9), si esta es mayor o igual a 17.52 se clasifican 577 observaciones de las cuales 363 (62.91%) pertenecen a la clase 0 y 214 (37.09%) pertenecen a la clase 1, llegando a un nodo terminal.

Si (FSCCN\_9) es menor a 17.52 se clasifican 2574 observaciones de las cuales 1323 (51.39%) pertenecen a la clase 0 y 1251 (48.60%) pertenecen a la clase 1, al no ser un nodo terminal pasamos a la siguiente variable escolaridad de la madre (FBCP1\_8. Superior) si esta es mayor o igual a 0.5 se clasifican 505 observaciones de las cuales 300 (59.41%) pertenecen a la clase 0 y 214 (37.09%) pertenecen a la clase 1, llegando a un nodo terminal.

Si (FBCP1\_8.Superior) es menor a 0.5 clasifican 2069 observaciones de las cuales 1023 (49.44%) pertenecen a la clase 0 y 1046 (50.55%) pertenecen a la clase 1, al no ser un nodo terminal analizamos la variable cuantos controles tuvo antes del parto (FSCE\_4) si esta es mayor o igual a 7.06 clasifican 889 observaciones de las cuales 477 (53.66%) pertenecen a la clase 0 y 412 (46.34%) pertenecen a la clase 1, seguimos con el análisis de la variable algún miembro del hogar tiene teléfono celular (FBCH\_17.no) si esta es menor a 0.5 clasifican 626 observaciones de las cuales 359 (57.34%) pertenecen a la clase 0 y 267 (42.65%) pertenecen a la clase 1, al no ser un nodo terminal seguimos con el análisis de la variable grupo de edad (FSCCN\_21.19-23) si esta es menor a 0.5 clasifican 524 observaciones de las cuales 316 (60.31%) pertenecen a la clase 0 y 208 (39.69%) pertenecen a la clase 1, llegando al nodo terminal; siendo 0 la clase mayoritaria.

Si (FSCCN\_21.19-23) es mayor o igual a 0.5 clasifican 102 observaciones de las cuales 43 (42.16%) pertenecen a la clase 0 y 59 (57.84%) pertenecen a la clase 1 llegando al nodo terminal, siendo 1 la clase mayoritaria.

Si (FBCH\_17.no) es mayor o igual a 0.5 clasifican 263 observaciones de las cuales 118 (44.87%) pertenecen a la clase 0 y 145 (55.13%) pertenecen a la clase 1, al no ser un nodo terminal analizamos la siguiente variable consumió ayer – jugos naturales (FIPA\_19.no) si esta es menor a 0.5 clasifican 178 observaciones de las cuales 68 (38.320%) pertenecen a la clase 0 y 110 (61.79%) pertenecen a la clase 1, llegando a un nodo terminal, siendo 1 la clase mayoritaria. Pero si (FIPA\_19.no) es mayor o igual a 0.5 clasifica 85 observaciones de las cuales 50 (58.82%) pertenecen a la clase 0 y 35 (41.17%) pertenecen a la clase 1, analizamos la variable pentavalente 3- dosis (FSCCN\_12) si esta es mayor o igual a 16.92 clasifican 51 observaciones de las cuales 38 (74.51%) pertenecen a la clase 0 y 13 (25.49%) pertenecen a la clase 1, llegando al nodo

terminal siendo 0 la clase mayoritaria. Si (FSCCN\_12) es menor a 16.92 clasifican 34 observaciones de las cuales 12 (35.29%) pertenecen a la clase 0 y 22 (64.71%) pertenecen a la clase 1 llegando a un nodo terminal, siendo 1 la clase mayoritaria.

Si (FSCE\_4) es menor a 7.06 clasifican 1180 observaciones de las cuales 546 (46.27%) pertenecen a la clase 0 y 634 (53.72%) pertenecen a la clase 1, al no ser un nodo terminal analizamos la variable el nacimiento fue a los 9 meses o antes de tiempo (FSCE\_9. prematuro) si esta es mayor o igual a 0.5 clasifican 176 observaciones de las cuales 62 (35.22%) pertenecen a la clase 0 y 114 (64.77%) pertenecen a la clase 1, llegando a un nodo terminal, siendo 1 la clase mayoritaria. Pero si (FSCE\_9. prematuro) es menor a 0.5 clasifican 1004 observaciones de las cuales 484 (48.20%) pertenecen a la clase 0 y 520 (51.79%) pertenecen a la clase 1, al no ser un nodo terminal analizamos la variable grupo de edad (FSCCN\_21.24-30) si esta es menor a 0.5 clasifican 804 observaciones de las cuales 368 (45.77%) pertenecen a la clase 0 y 436 (54.22%) pertenecen a la clase 1, llegando a un nodo terminal, siendo 1 la clase mayoritaria. Pero si (FSCCN\_21.24-30) es mayor o igual a 0.5 clasifican 200 observaciones de las cuales 116 (58%) pertenecen a la clase 0 y 84 (42%) pertenecen a la clase 1, al no ser un nodo terminal analizamos la variable meses de embarazo cuando se hizo el primer control (FSCE\_3), si esta es menor a 4.44 clasifican 46 observaciones de las cuales 18 (39.13%) pertenecen a la clase 0 y 28 (60.86%) pertenecen a la clase 1, llegando al nodo terminal, siendo 1 la clase mayoritaria. Si (FSCE\_3) es mayor a 4.44 clasifican 154 observaciones de las cuales 98 (63.63%) pertenecen a la clase 0 y 56 (36.36%) pertenecen a la clase 1, llegando al nodo terminal; siendo 0 la clase mayoritaria.

### 3.3.1 Matriz de confusión

El modelo construido mediante la técnica de árbol de decisión presentó un error total de predicción del 40%, la matriz de confusión asociada a este modelo se presenta a continuación.

**Tabla 11-3:** Matriz de confusión árbol de decisión

		Desnutrición crónica menores de 5 años	
		No	Si
Desnutrición crónica menores de 5 años	No	1546	465
	Si	888	470

Fuente: Elaboración propia

Realizado por: Congacha, Giorgia, 2020

### 3.3.2 Modelo Gradient Boosting

El nuevo modelo construido a través de la técnica de Gradient Boosting para reducir la variabilidad y mejorar la calidad predictiva del árbol de decisión, presentó un error total de predicción de 37.31%. La matriz de confusión asociada a este modelo se presenta a continuación.

**Tabla 12-3:** Matriz de confusión Gradient Boosting

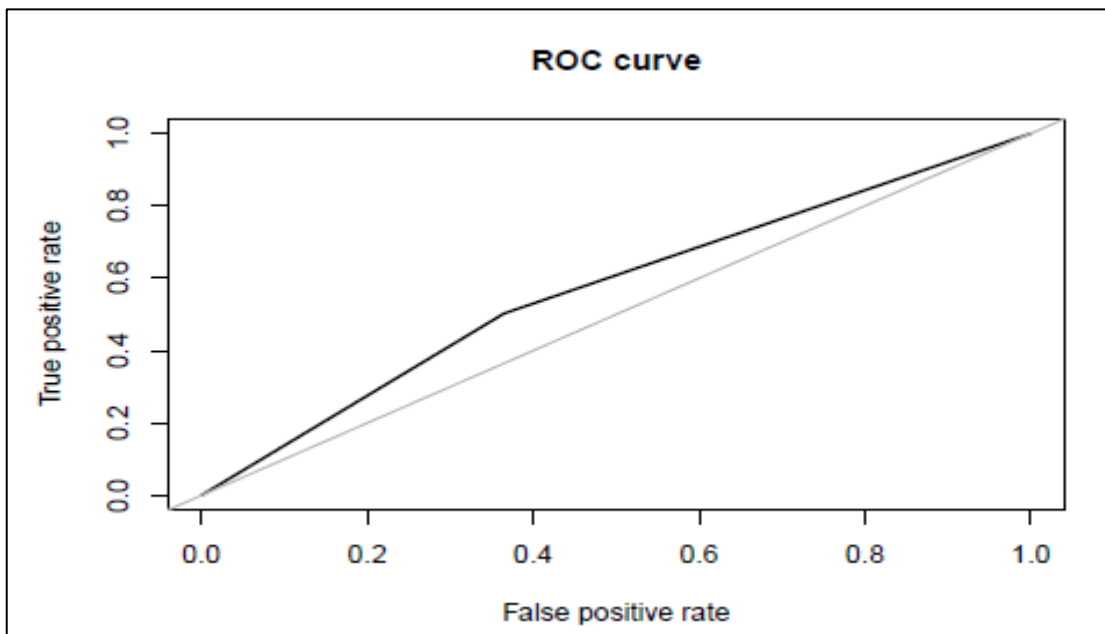
		Desnutrición crónica menores de 5 años	
		No	Si
Desnutrición crónica menores de 5 años	No	1687	510
	Si	747	425

Fuente: Elaboración propia

Realizado por: Congacha, Giorgia, 2020

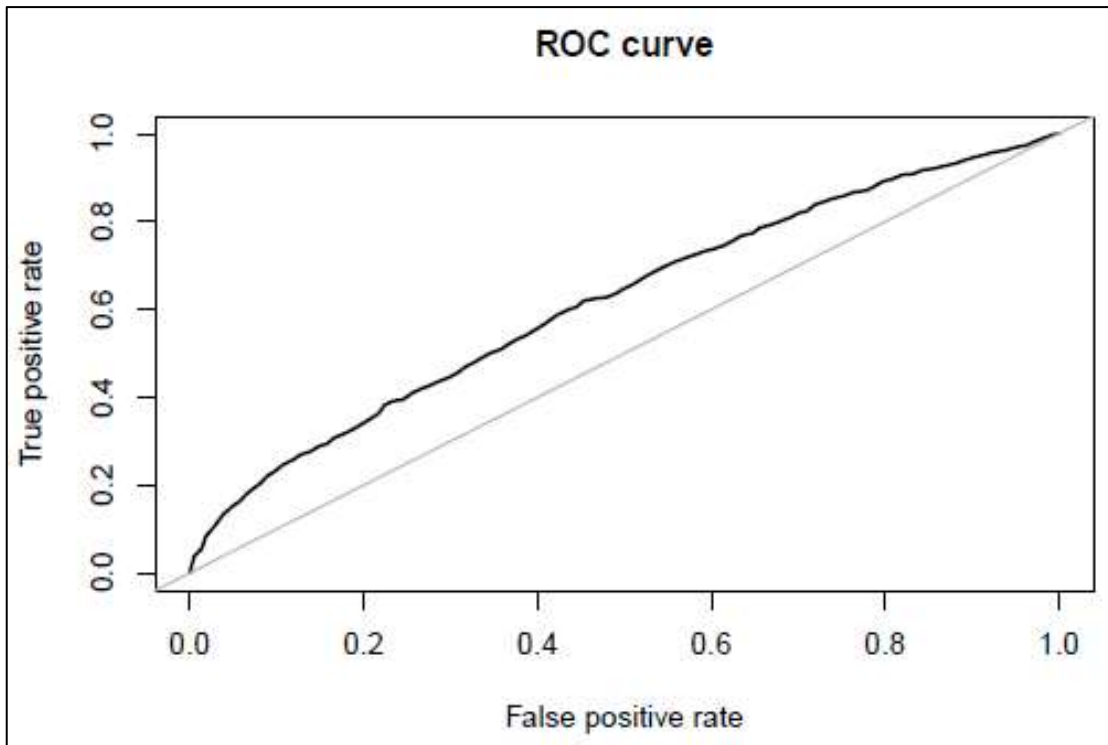
### 3.4 Comparativa de los modelos: Regresión logística, Árboles de decisión y Gradient Boosting

Tomando en cuenta la tasa de error a través de la matriz de confusión de cada modelo construido, (Tabla 10-4, Tabla 11-4 y Tabla 12-4) se estudió la calidad predictiva de los modelos generados. También se lo realiza analizando el AUC de las curvas de ROC, las gráficas de las curvas de ROC y el AUC de cada modelo se presentan a continuación:



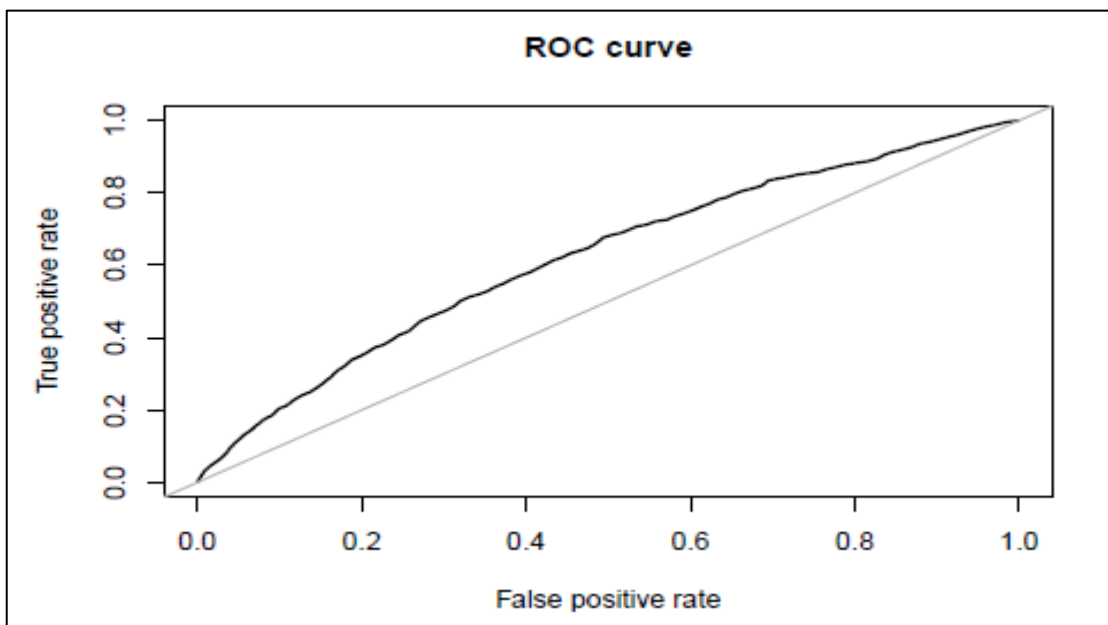
**Gráfico 2-3:** Curva ROC: Árbol de clasificación

Realizado por: Congacha, Giorgia, 2020



**Gráfico 3-3:** Curva ROC: Gradient Boosting

Realizado por: Congacha, Giorgia, 2020



**Gráfico 4-3:** Curva ROC: Regresión logística

Realizado por: Congacha, Giorgia, 2020



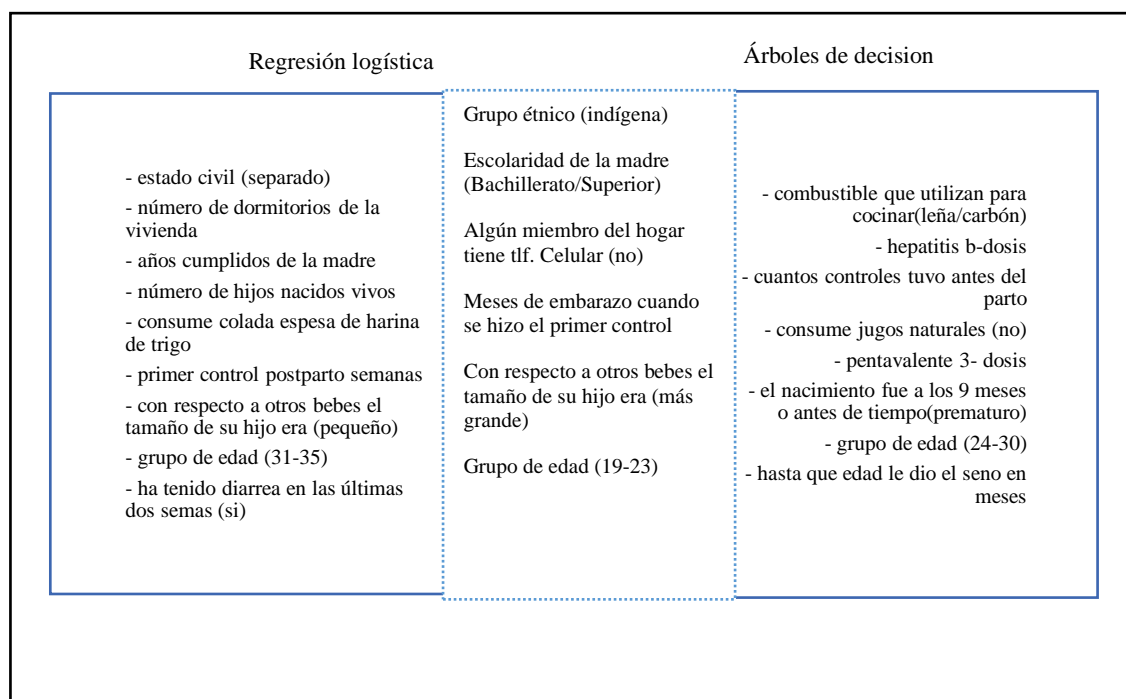
**Tabla 13-3:** Áreas bajo la curva (AUC)

	Áreas bajo la curva (AUC)
<b>Árbol de decisión</b>	58,19%
<b>Gradient Boosting</b>	61,7%
<b>Regresión Logística</b>	62,10%

Fuente: Elaboración propia

Realizado por: Congacha, Giorgia, 2020

Tomando en cuenta las representaciones gráficas de la curva de ROC y las áreas bajo la curva (AUC) para los modelos construidos se puede observar que el modelo de regresión logística presenta un mayor AUC = 62.10% por tanto este tiene un poder predictivo más alto que el árbol de decisión que presenta un AUC = 58.19% y que Gradient Boosting AUC = 61.7%. Esto también se puede corroborar a través de la matriz de confusión, puesto que la tasa de error para el modelo de regresión logística fue del 35% mientras que el de árbol de decisión presento una tasa de error del 40% y Gradient Boosting 37.31%.



**Figura 1-3:** Comparación de factores significativos asociados a la desnutrición crónica infantil a través de los modelos de árboles de decisión y regresión logística

Realizado por: Congacha, Giorgia, 2020

## CONCLUSIONES

- Al comparar los modelos: árboles de decisión y regresión logística mediante la tasa de error obtenida de la matriz de confusión y Curvas de ROC a través del AUC se reflejaron diferencias en el desempeño de estas técnicas, en árboles de decisión se notó un AUC del 58.19% y una tasa de error del 40% a diferencia de la técnica de regresión logística que presentó un AUC del 62.10% y una tasa de error del 35%. Con la finalidad de reducir la variabilidad en el modelo de árboles de decisión y aumentar el poder predictivo se utilizó el método de ensemble Gradient Boosting el cual obtuvo un AUC = 61.7% y una tasa de error de 37.31%; por tanto, el mejor modelo para predecir la desnutrición crónica infantil es regresión logística.
- Analizando los factores asociados a la desnutrición crónica infantil, como se puede ver en la **Figura 1-3** los que prevalecieron como significativos en ambos modelos son: grupo étnico indígena (FBCP1\_6.indígena), escolaridad de la madre (FBCP1\_8.Educación.Media.Bachillerato, FBCP1\_8.Superior), algún miembro del hogar tiene teléfono celular (FBCH\_17.no), meses de embarazo cuando se hizo el primer control (FSCE\_3), con respecto a otros bebés el tamaño de su hijo era igual, pequeño o más grande (FSCN\_5.más.grande) y grupo de edad (FSCCN\_21.19-23).
- Para el modelo de Regresión logística los factores que ayudaron a identificar la desnutrición crónica infantil a parte de los ya mencionados son: estado civil (FBCP1\_7.separado), número de dormitorios de la vivienda (FBCH\_12), años cumplidos de la madre (FBCP1\_1), número de hijos nacidos vivos (FBCP1\_5), consume colada espesa de harina de trigo (FIPA\_33.no.sabe...no.responde), primer control postparto-semanas (FSCE\_12), con respecto a otros bebés el tamaño de su hijo era pequeño (FSCN\_5.pequeño), grupo de edad (FSCCN\_21.31.35) y ha tenido diarrea en las últimas dos semanas (FIEI\_1.si).
- Para árboles de clasificación, los factores asociados a parte de los mencionados anteriormente son: combustible que utilizan para cocinar (FBCH\_14. leña/carbón), hepatitis b – dosis (FSCCN\_9), cuántos controles tuvo antes del parto (FSCE\_4), consume jugos naturales (FIPA\_19.no), pentavalente 3-dosis (FSCCN\_12), el nacimiento fue a los 9 meses o antes de tiempo (FSCE\_9. prematuro), grupo de edad (FSCCN\_21.24-30), hasta que edad le dio el seno-meses (FIPA\_2).
- Este trabajo también aporta con una aplicación web interactiva con la finalidad de implementar los modelos analizados en el estudio (árboles de decisión y regresión logística), en el análisis de la Encuesta Nacional de Salud y Nutrición - ENSANUT 2018-2019.
- Este estudio, además enfatiza la importancia del procedimiento de preparación y limpieza de datos como un requisito indispensable para llevar a cabo el análisis de los modelos.

## RECOMENDACIONES

- Para el desarrollo de árboles de decisión se recomienda ir probando los hiperparámetros: profundidad (maxdepth), mínimo de observaciones de cada nodo (minsplit), mínimo de observaciones en el árbol terminal (minbucket), con la finalidad de mejorar la precisión del modelo y reducir su variabilidad. Si bien es cierto que las técnicas de ensemble reducen la variabilidad de los árboles individuales y mejoran su calidad predictiva, son difíciles de interpretar y tienen altos costes computacionales, por lo que se recomienda utilizar modelos prácticos que permitan la fácil interpretación de sus resultados.
- Tomando en cuenta que el modelo de regresión logística resultó ser el más confiable al predecir la desnutrición crónica infantil; se recomienda su aplicación en bases de datos sobre salud similares a la utilizada en el presente estudio.
- Debido a la gran cantidad de información que se produce en la actualidad, que hace inviable el procesamiento manual o bajo software tradicional se recomienda para futuros estudios aprovechar el potencial de softwares especializados en ciencia de datos como R y Python.
- Actualmente, el proyecto de política pública que ejecuta el gobierno ecuatoriano para controlar la desnutrición crónica infantil, es el proyecto “Misión Ternura” el mismo que se plantea el bienestar humano desde la gestación y durante los primeros 5 años de vida; esta política incluye en sus componentes algunas variables que se asocian con la desnutrición crónica, pero no especifica acciones explícitas para los temas de etnia y grupos de edad a pesar de ser variables que juegan un papel importante en esta problemática, por ello se recomienda socializar con entes gubernamentales los resultados obtenidos en esta investigación, serán de provecho para la toma de decisiones en planes de contingencia que den solución a la desnutrición crónica infantil en el Ecuador.

## BIBLIOGRAFIA

AMAT, J., "Árboles de decisión, bagging, random forest, boosting y C5.0". [en línea], 2017.[Consulta: 2 noviembre 2020]. Disponible en: [https://rpubs.com/Joaquin\\_AR/255596](https://rpubs.com/Joaquin_AR/255596).

APRENDEIA., "Ventajas y Desventajas de los Algoritmos de Clasificación". [blog]. [Consulta: 18 noviembre 2020]. Disponible en: <https://aprendeia.com/ventajas-y-desventajas-de-los-algoritmos-de-clasificacion-machine-learning/>.

BARRERA-DUSSÁN, N.;et al. "Prevalencia y determinantes sociales de malnutrición en menores de 5 años afiliados al Sistema de Selección de Beneficiarios para Programas Sociales (SISBEN) del área urbana del municipio de Palermo en Colombia, 2017'. *Universidad y Salud* [en línea], 2018, vol. 20, no. 3, pp. 236-246.[Consulta: 23 enero 2020]. ISSN 2389-7066. Disponible en: <http://dx.doi.org/10.22267/rus.182003.126>

BULLÓN C. L.; & ASTETE R., L. "Determinantes de la desnutrición crónica de los menores de tres años en las regiones del Perú: sub-análisis de la encuesta ENDES 2000". *Anales Científicos* [en línea], 2016, vol. 77, no. 2, pp. 249-259. ISSN 2519-7398. Disponible en: <http://dx.doi.org/10.21704/ac.v77i2.636>

CANAZAS, V. "Factores asociados a la desnutrición crónica infantil en Perú". *Revista Latinoamericana de Población* [en línea], 2010, vol. 4, no. 6, pp. 41-56. ISSN 2175-8581. Disponible en: <http://dx.doi.org/10.31406/relap2010.v4.i1.n6.2>.

CARRANZA BARONA, C. "Políticas públicas en alimentación y nutrición: Los programas de alimentación social de Ecuador". Quito-Ecuador; Abya-Yala, 2011, ISBN 978-9978-67-270-9, pp.5-217

CEBALLOS-GONZÁLEZ, A.; et al. "Influencia de la dinámica familiar y otros factores asociados al déficit en el estado nutricional de preescolares en guarderías del sistema Desarrollo Integral de la Familia (DIF) Jalisco". *Boletín médico del hospital infantil de México* [en línea], 2005, vol.62, no.2, pp.104-116. [Consulta: 19 abril 2019]. ISSN 1665-1146. Disponible en: <http://www.scielo.org.mx/pdf/bmim/v62n2/v62n2a4.pdf>.

CEPAL, "Malnutrición en niños y niñas en América Latina y el Caribe | Enfoques | Comisión Económica para América Latina y el Caribe" [en línea], 2018. [Consulta: 30 octubre 2020]. Disponible en: <https://www.cepal.org/es/enfoques/malnutricion-ninos-ninas-america-latina-caribe>.

Charris, Luis; et.al; "Análisis comparativo de algoritmos de árboles de decisión en el procesamiento de datos biológicos". *Investigación y Desarrollo en TIC* [en línea], 2018, vol. 9,

no.1, pp. 26-34. [Consulta: 19 enero 2020]. ISSN: 2216-1570 Disponible en: <http://revistas.unisimon.edu.co/index.php/identific>.

CUBAS ROVIRA, G., "Análisis del algoritmo MINI para imputación de valores perdidos en conjuntos de datos pequeños y con variables continuas y categóricas" [En línea] (trabajo de titulación) (Tercer Nivel). Universitat Politecnica de Valencia, Valencia. 2017, pp.10-92. [Consulta: 30 agosto 2020]. Disponible en: <http://hdl.handle.net/10251/80506>

DE LA HOZ MANOTAS, A., et al. "Técnicas de ml en medicina cardiovascular". *Memorias* [en línea], 2013, vol. 11, pp. 41-46. [Consulta: 19 enero 2020]. Disponible en: [https://www.researchgate.net/publication/279850557\\_Tecnicas\\_de\\_ml\\_en\\_medicina\\_cardiovascular](https://www.researchgate.net/publication/279850557_Tecnicas_de_ml_en_medicina_cardiovascular).

DEMIR, E., "A Decision Support Tool for Predicting Patients at Risk of Readmission: A Comparison of Classification Trees, Logistic Regression, Generalized Additive Models, and Multivariate Adaptive Regression Splines". *Decision Sciences* [en línea], 2014, vol. 45, no. 5, pp. 849-880. [Consulta: 19 enero 2020]. ISSN 00117315, Disponible en: <http://doi.wiley.com/10.1111/deci.12094>.

DÍAZ, J; & CORREA, J.C., "Comparación entre árboles de regresión CART y regresión lineal". *Comunicaciones en Estadística* [en línea], 2013, vol. 6, no. 2, pp. 175-195. [Consulta: 19 enero 2020]. ISSN 71269839. Disponible en: <http://www.bdigital.unal.edu.co/9474/1/71269839.2013.pdf>

DUPOUY BERRIOS Carlos,. "Aplicación de árboles de decisión para la estimación del escenario económico y la estimación de movimiento la tasa de interés en Chile" [En línea] (trabajo de titulación) (Maestría) Universidad de Chile. 2014, pp. 8-91. [Consulta: 19 enero 2020]. Disponible en: <http://repositorio.uchile.cl/handle/2250/117556>

FAO,. "El estado de la seguridad alimentaria y la nutrición en el mundo. Fomentando la resiliencia climática en aras de la seguridad alimentaria y la nutrición". [en línea]. S.l.: s.n. [Consulta: 18 noviembre 2020]. ISBN 9789251308417. Disponible en: <http://www.fao.org/publications/es>.

FERRERO, R.; & LÓPEZ, J. "Qué son los árboles de decisión y para qué sirven". [blog] [Consulta: 2 septiembre 2020]. Disponible en: <https://www.maximaformacion.es/blog-dat/que-son-los-arboles-de-decision-y-para-que-sirven/>.

FIUZA, D. y RODRÍGUEZ, J., "La regresión logística: una herramienta versátil". *nefrología* [en línea], 2000, vol. 20, no. 6, pp. 477-565. [Consulta: 30 octubre 2020]. Disponible en: <https://www.revistanefrologia.com/es-la-regresion-logistica-una-herramienta-versatil-articulo->

X0211699500035664.

GUDE SAMPEDRO, F.; & PÉREZ GONZÁLEZ, A., "Imputación de datos faltantes en un modelo de tiempo de fallo acelerado". [En línea] (trabajo de titulación) (Maestría). Universidad de Vigo. 2014, pp. 23-50. [Consulta: 8 junio 2020]. Disponible en: [http://eio.usc.es/pub/mte/descargas/ProyectosFinMaster/Proyecto\\_940.pdf](http://eio.usc.es/pub/mte/descargas/ProyectosFinMaster/Proyecto_940.pdf)

HERNÁNDEZ Leal,; et al., "Big Data: una exploración de investigaciones, tecnologías y casos de aplicación". *Tecnológicas*, 2017, vol. 20, no. 39, pp. 9-24. [Consulta: 2 septiembre 2020]. ISSN 2256-5337. Disponible en: <http://www.scielo.org.co/pdf/teclo/v20n39/v20n39a02.pdf>.

HOSMER, D.; & LEMESHOW, S., "*Applied Logistic Regression*" [en línea]. Second Edition. Canada; A Wiley - Interscience Publication, 2000, [Consulta: 3 septiembre 2020]. ISBN 0-471-35632-8. Disponible en: [http://resource.heartonline.cn/20150528/1\\_3kOOSTg.pdf](http://resource.heartonline.cn/20150528/1_3kOOSTg.pdf).

INEC;. "Diseño muestral de la Encuesta Nacional de Salud y Nutrición". 2020 [en línea]. [Consulta: 14 agosto 2020]. Disponible en: [https://www.ecuadorencifras.gob.ec/documentos/web-inec/Estadisticas\\_Sociales/ENSANUT/ENSANUT\\_2018/Metodologia\\_del\\_disenio\\_muestral\\_ENSANUT\\_2018.pdf](https://www.ecuadorencifras.gob.ec/documentos/web-inec/Estadisticas_Sociales/ENSANUT/ENSANUT_2018/Metodologia_del_disenio_muestral_ENSANUT_2018.pdf).

INEC;. "Evolución histórica de la ENSANUT 2018-2019" . 2018 [en línea]. [Consulta: 14 agosto 2020]. Disponible en: [https://www.ecuadorencifras.gob.ec/documentos/web-inec/Estadisticas\\_Sociales/ENSANUT/ENSANUT\\_2018/Evolucion%20Historica%20de%20ENSANUT%202018.pdf](https://www.ecuadorencifras.gob.ec/documentos/web-inec/Estadisticas_Sociales/ENSANUT/ENSANUT_2018/Evolucion%20Historica%20de%20ENSANUT%202018.pdf)

INEC;. "Documento metodológico de la Encuesta Nacional de Salud y Nutrición" (ENSANUT). 2019 [en línea]. [Consulta: 14 agosto 2020]. Disponible en: [https://www.ecuadorencifras.gob.ec/documentos/web-inec/Estadisticas\\_Sociales/ENSANUT/ENSANUT\\_2018/Metodologia\\_ENSANUT\\_2018.pdf](https://www.ecuadorencifras.gob.ec/documentos/web-inec/Estadisticas_Sociales/ENSANUT/ENSANUT_2018/Metodologia_ENSANUT_2018.pdf).

KABACOFF, R. "*Tree - Based Models*". DataCamp [blog]. [Consulta: 7 septiembre 2020]. Disponible en: <https://www.statmethods.net/advstats/cart.html>.

LARREA, C. "Desnutrición, etnicidad y pobreza en el Ecuador y el área Andina. *UASB-DIGITAL*, 2006", pp.12-24 [Consulta: 12 enero 2020]. Disponible en: <http://repositorio.uasb.edu.ec/handle/10644/856>

LAZCANO, R., "Big data, machine learning y deep learning: conceptos y diferencias". [en línea], 2019. [Consulta: 30 octubre 2020]. Disponible en: <https://blog.enzymeadvisinggroup.com/big-data-machine-learning>.

LÓPEZ, E. y RUIZ, M., "Análisis de datos con el Modelo Lineal Generalizado. Una aplicación con R". *Revista Española De Pedagogía* [en línea], 2011. vol. 69, no. No.248, pp. 59-80. [Consulta: 2 noviembre 2020]. ISSN 23766383. Disponible en: <https://www.jstor.org/stable/23766383>.

LÓPEZ, Ana., "Análisis previo y exploratorio de datos" [blog]. [Consulta: 30 agosto 2020] Disponible en: [http://www.ub.edu/aplica\\_infor/spss/Cap2-2.htm](http://www.ub.edu/aplica_infor/spss/Cap2-2.htm).

MAMANI ORTIZ, Yercin; et al. "La desnutrición infantil y su relación con los pisos ecológicos en Vinto, Cochabamba, Bolivia". *Gaceta Médica Boliviana* [en línea], 2011, vol. 35, no. 1, pp. 16-21. [Consulta: 20 enero 2020]. Disponible en: [http://www.scielo.org.bo/scielo.php?script=sci\\_arttext&pid=S1012-29662012000100004&lang=es](http://www.scielo.org.bo/scielo.php?script=sci_arttext&pid=S1012-29662012000100004&lang=es).

MÁRQUEZ-GONZÁLEZ,; et al., "Clasificación y evaluación de la desnutrición en el paciente pediátrico". *El residente* [en línea], 2012. vol. VII, no. 2, pp. 56-69. Disponible en: [www.medigraphic.org.mx](http://www.medigraphic.org.mx).

MARTÍNEZ, R. y FERNÁNDEZ, A., "Modelo de análisis del impacto social y económico de la desnutrición infantil en América Latina". *PMA Naciones Unidas* [en línea], 2006. Santiago de Chile: [Consulta: 16 noviembre 2020]. Disponible en: [https://repositorio.cepal.org/bitstream/handle/11362/5491/S0600972\\_es.pdf?sequence=1](https://repositorio.cepal.org/bitstream/handle/11362/5491/S0600972_es.pdf?sequence=1).

MINISTERIO DE SALUD PÚBLICA DEL ECUADOR, "Plan Intersectorial de alimentación y nutrición Ecuador 2018-2025". Buena nutrición toda una vida. [en línea], 2018. Quito: [Consulta: 30 octubre 2020]. Disponible en: <https://www.salud.gob.ec/wp-content/uploads/2018/08/PIANE-2018-2025-final-compressed-.pdf>.

MUKURIA, A., et al., "*Nutritional status of children: results from the demographic and health surveys 1994-2001*". [en línea] 2005., Reporte no.10, [Consulta: 13 agosto 2020]. Disponible en: <http://www.measuredhs.comorbycontacting>.

OMS. "Preguntas y respuestas: malnutrición y emergencias" [blog]. 2017. [Consulta: 21 abril 2020]. Disponible en: <http://www.who.int/features/qa/malnutrition-emergencies/es/>.

ORELLANA, J., "Arboles de decision y Random Forest". [en línea], 2018. [Consulta: 2 noviembre 2020]. Disponible en: <https://bookdown.org/content/2031/>.

ORTIZ, A; et.al; "Desnutrición infantil, salud y pobreza: intervención desde un programa integral". *SCIELO* [en línea], 2006, vol. 21, no. 4, pp. 533-541. [Consulta: 30 octubre 2020]. ISSN 1699-5198. Disponible en: [http://scielo.isciii.es/scielo.php?script=sci\\_arttext&pid=S0212-16112006000700011](http://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S0212-16112006000700011).

PALAU, A.P., "Evaluación del crecimiento posnatal en los prematuros de muy bajo peso con edad gestacional menor o igual a 32 semanas desde el nacimiento hasta los 5 años de vida". (Tesis Doctoral)(Cuarto Nivel) [en línea]. Barcelona: Universitat Autònoma de Barcelona, 2017. pp.116-129 [Consulta: 16 noviembre 2020]. Disponible en: <https://www.tdx.cat/bitstream/handle/10803/457736/app1de1.pdf?sequence=1&isAllowed=y>.

PAREDES, G., "Factores que determinan el estado de inseguridad alimentaria en niños y niñas de 0 a 5 años, en el Ecuador 2012" [en línea], 2016. Quito: Pontificia Universidad Católica del Ecuador. [Consulta: 14 noviembre 2020]. Disponible en: [http://repositorio.puce.edu.ec/bitstream/handle/22000/12630/Disertación\\_Gabriela\\_Paredes.pdf?sequence=1&isAllowed=y](http://repositorio.puce.edu.ec/bitstream/handle/22000/12630/Disertación_Gabriela_Paredes.pdf?sequence=1&isAllowed=y).

PAZ VELOZ, A.K., "Determinantes de la desnutrición crónica de menores de 5 años y análisis del consumo alimenticio de los hogares del cantón San Miguel de Urququí". (trabajo de titulación) (Tercer Nivel). Universidad Central del Ecuador 2017, pp. 63-75. [Consulta: 1 enero 2020]. Disponible en: <http://www.dspace.uce.edu.ec/handle/25000/13975?mode=full>

PEÑA, M., et al., "La nueva situación epidemiológica de Ecuador". *Comunigraf* 2014. vol. 32, pp.5-101. [Consulta: 21 noviembre 2019]. Disponible en: [https://www.paho.org/ecu/index.php?option=com\\_docman&view=download&category\\_slug=comunicacion-social&alias=509-boletin-informativo-n0-32-junio-2014-1&Itemid=599](https://www.paho.org/ecu/index.php?option=com_docman&view=download&category_slug=comunicacion-social&alias=509-boletin-informativo-n0-32-junio-2014-1&Itemid=599).

PÉREZ-RAVE, J.; & ECHAVARRÍA, F.G., "*Classification trees vs. Logistics Regression in the Generic skill Development in Engineering*". *Computacion y Sistemas* [en línea], 2018. vol. 22, no. 4, pp. 1519-1541. [Consulta: 19 enero 2020] ISSN 20079737. Disponible en : <http://dx.doi.org/10.13053/CyS-22-4-2804>.

RAVINA, R ; et.al;"Los ÁrboLes de decisión. Una herramienta práctica para la toma de decisiones" [en línea], 2018. 1era. Edición. Cabimas: UNERMB. [Consulta: 30 octubre 2020]. ISBN 9789804270567. Disponible en: [http://150.185.9.18/fondo\\_editorial/](http://150.185.9.18/fondo_editorial/).

RUIZ-RUIZ Nubia.; "Mortalidad por desnutrición en menores de cinco años. Pobreza y desarrollos regionales. Colombia. 2003-2012". *Economía, Sociedad y Territorio* [en línea], 2018, vol. xviii, pp. 35-75. [Consulta: 20 enero 2020]. ISSN 20181077 Disponible en: <http://dx.doi.org/10.22136/est20181077>.

SERNA PINEDA, S., "Comparación de Árboles de Regresión y Clasificación y regresión logística". (trabajo de titulación) (Maestría). Universidad Nacional de Colombia 2009., pp. 14-60. [Consulta: 17 septiembre 2020]. Disponible en: [http://www.bdigital.unal.edu.co/671/1/42694070\\_2009.pdf](http://www.bdigital.unal.edu.co/671/1/42694070_2009.pdf).



SOLARTE MARTINEZ, G.R., "Arboles de decisiones en el diagnóstico de enfermedades cardiovasculares". *Scientia Et Technica* [en línea], 2011, vol. XVI, no. 49, pp. 104-109. [Consulta: 19 enero 2020] ISSN 0122-1701. Disponible en: <http://dx.doi.org/10.22517/23447214.1487>.

SRIVASTAVA, T., "*Evaluation Metrics Machine Learning*". 2019 [blog]. [Consulta: 3 septiembre 2020]. Disponible en: <https://www.analyticsvidhya.com/blog/2019/08/11-important-model-evaluation-error-metrics/>.

UNICEF. "El Estado Mundial de la Infancia". Nueva York . Palais des Nations 1998, ISBN 0-19-829401-8, pp. 93-1007 [Consulta: 19 enero 2020].

UNICEF. "La desnutrición crónica infantil.2013", [Consulta: 17 mayo 2019]. Disponible en: <https://www.unicef.org/peru/spanish/La-desnutricion-cronica-infantil.pdf>.

USECHE, L.; & MESA, D., "Una introducción a la imputación de valores perdidos". *Terra*. 2006, vol. XXII, no. 31, pp. 127-152. [Consulta: 30 agosto 2020]. ISSN 1012-7089. Disponible en: <https://www.redalyc.org/pdf/721/72103106.pdf>

VILLACÍS CRUZ, B.et al., "La desnutrición en la población indígena y afroecuatoriana menor de cinco años". [en línea]. S.l.: [Consulta: 17 mayo 2019]. Disponible en: [www.inec.gov.ec](http://www.inec.gov.ec).

WISBAUM, Wendy. "Causas, consecuencias y estrategias para su prevención y tratamiento: la desnutrición infantil". 2011, [Consulta: 21 junio 2019]. Disponible en: [www.unicef.es](http://www.unicef.es).

## ANEXOS

**ANEXO A:** Código en R utilizado para selección y recodificación de variables de acuerdo a la UNICEF.

```
library(data.table)
```

```
library(readr)
```

### Base de datos sobre personas

```
B1 <- read_csv("1_BDD_ENS2018_f1_personas.csv")
```

#### *Selección y recodificación de variables de interés*

```
BT1M <-
```

```
BT1[,.(id_viv,id_hogar,id_per,FBCHP1_4=f1_s2_14,FBCHP1_6=f1_s2_9,FBCHP1_7=f1_s2_16,FBCH_17=f1_s2_23,  
FSCCN_23=sexo)]
```

```
dim(BT1M)
```

### Base de datos sobre hogar

```
B2 <- read_csv("2_BDD_ENS2018_f1_hogar.csv")
```

#### *Selección y recodificación de variables de interés*

```
BT2M <-
```

```
BT2[,.(id_viv,id_hogar,FBCH_1=f1_s1_1,FBCH_2=f1_s1_2,FBCH_3=f1_s1_3,FBCH_4=f1_s1_5,FBCH_5=f1_s1_4,  
FBCH_6=f1_s1_21,FBCH_7=f1_s1_24,FBCH_8=f1_s1_13,FBCH_9=f1_s1_10,FBCH_10=f1_s1_11,FBCH_11=f1_s1_7,  
FBCH_12=f1_s1_8,FBCH_13=f1_s1_30,FBCH_14=f1_s1_9,FBCH_15=f1_s1_19,FBCH_16=f1_s1_44_10,  
FBCH_18=f1_s1_37)]
```

```
dim(BT2M)
```

### Base de datos sobre mujeres en edad fértil

```
B4 <- read_csv("4_BDD_ENS2018_f2_mef.csv")
```

#### *Selección y recodificación de variables de interés*

```
BT4M <-
```

```
BT4[,.(id_viv,id_hogar,id_per,FBCHP1_1=f2_s1_101,FBCHP1_2=f2_s2_201,FBCHP1_5=f2_s2_217_1+f2_s2_217_2,F  
BCP1_8=nivins)]
```

```
dim(BT4M)
```

### Base de datos sobre lactancia

```
B5 <- read_csv("5_BDD_ENS2018_f2_lactancia.csv")
```

#### *Selección y recodificación de variables de interés*

```
BT5M <-
```

```
BT5[,.(id_viv,id_hogar,id_per,FIPA_4=f2_s3b_305,FIPA_5=f2_s3a_304,FIPA_6=f2_s3b_306,FIPA_7=f2_s3c_307  
_1,FIPA_8=f2_s3c_307_2,FIPA_9=f2_s3c_307_3,FIPA_10=f2_s3c_308,FIPA_11=f2_s3c_309,FIPA_12=f2_s3d_3
```

```
10,FIPA_13=f2_s3d_311a_1,FIPA_14=f2_s3d_311a_2,FIPA_15=f2_s3d_311b_1,FIPA_16=f2_s3d_311b_2,FIPA_17=f2_s3d_311c_1,FIPA_18=f2_s3d_311c_2,FIPA_19=f2_s3d_311d_1,FIPA_20=f2_s3d_311e_1,FIPA_21=f2_s3d_311e_2,FIPA_22=f2_s3d_311f_1,FIPA_23=f2_s3d_311f_2,FIPA_24=f2_s3d_311g_1,FIPA_25=f2_s3d_311g_2,FIPA_26=f2_s3d_311h_1,FIPA_27=f2_s3d_311h_2,FIPA_28=f2_s3d_311i_1,FIPA_29=f2_s3d_311i_2,FIPA_30=f2_s3d_311j_1,FIPA_31=f2_s3d_311j_2,FIPA_32=f2_s3d_312,FIPA_33=f2_s3d_313_1,FIPA_34=f2_s3d_313_2,FIPA_35=f2_s3d_313_3,FIPA_36=f2_s3d_313_4])
```

```
dim(BT5M)
```

### Base de datos sobre salud niñez

```
B6 <- read_csv("6_BDD_ENS2018_f2_salud_ninez.csv")
```

#### *Selección y recodificación de variables de interés*

```
BT6M <-
```

```
BT6M[,.(id_viv,id_hogar,id_per,FBCP1_3=f2_s4a_405_,FSCE_1=f2_s4b_406_,FSCE_2=f2_s4b_407a_,FSCE_3=f2_s4b_420_,FSCE_4=f2_s4b_421_,FSCE_5=f2_s4b_417_,FSCE_6=f2_s4c_423a_,FSCE_7=f2_s4c_424a_,FSCE_8=f2_s4c_425_,FSCE_9=f2_s4d_426_,FSCE_10=f2_s4e_440_,FSCE_11=f2_s4e_441a_,FSCE_12=f2_s4e_441b_,FSCE_13=f2_s4e_441c_,FSCE_14=f2_s4e_442a_,FSCN_1=f2_s4d_428_,FSCN_2=f2_s4d_437a_,FSCN_3=f2_s4d_437b_,FSCN_4=f2_s4d_438_,FSCN_5=f2_s4d_439_,FSCN_6=f2_s4d_432_,FSCN_7=f2_s4d_436a_,FSCN_9=f2_s4d_434a_,FSCN_10=f2_s4d_434b_,FSCN_11=f2_s4d_435a_,FSCN_12=f2_s4d_435b_,FSCN_13=f2_s4d_433a_,FSCN_14=f2_s4d_433b_,FSCCN_1=f2_s4f_448_,FSCCN_2=f2_s4f_449dias_,FSCCN_3=f2_s4f_449semanas_,FSCCN_4=f2_s4f_449meses_,FSCCN_5=f2_s4f_450_,FSCCN_6=f2_s4f_452a_,FSCCN_7=f2_s4f_456_,FSCCN_8=f2_s4j_4871aanio_,FSCCN_9=f2_s4j_4872aanio_,FSCCN_10=f2_s4j_4875aanio_,FSCCN_11=f2_s4j_4876aanio_,FSCCN_12=f2_s4j_4877aanio_,FSCCN_13=f2_s4j_4873aanio_,FSCCN_14=f2_s4j_4874aanio_,FSCCN_15=f2_s4j_4878aanio_,FSCCN_16=f2_s4j_4879aanio_,FSCCN_17=f2_s4j_48710aanio_,FSCCN_18=f2_s4j_48711aanio_,FSCCN_19=f2_s4j_48712aanio_,FSCCN_20=f2_s4j_48713aanio_,FSCCN_21=edad_meses,FSCCN_22=edadmeses,FIPA_1=f2_s4f_454dias_,FIPA_2=f2_s4f_454meses_,FIPA_3=f2_s4f_454años_,FIEI_1=f2_s4g_457_,FIEI_2=f2_s4g_458_,FIEI_3=f2_s4g_459_,FIEI_4=f2_s4h_473_,FIEI_5=f2_s4h_474_,FIEI_6=f2_s4h_475_,FIEI_7=f2_s4h_476a_,dcronica)]
```

```
dim(BT6M)
```

## ANEXO B: Código en R para la unión de base de datos

```
library(data.table)
```

```
library(dplyr)
```

```
library(tidyr)
```

```
BASE1 <- read.csv("BASE1.csv")
```

```
BASE2 <- read.csv("BASE2.csv")
```

```
BASE4 <- read.csv("BASE4.csv")
```

```
BASE5 <- read.csv("BASE5.csv")
```

```
BASE6 <- read.csv("BASE6.csv")
```

*Variables necesarias para realizar combinación o cruce de variables*

*id\_viv*

*id\_hogar*

*id\_per*

*Unimos las bases con la función merge*

```
BPH <- merge(BASE1, BASE2, by=c("id_viv", "id_hogar"))
```

```
dim(BPH)
```

```
names(BPH)
```

```
BPHMEF <- merge(BPH, BASE4, by=c("id_viv", "id_hogar", "id_per"))
```

```
names(BPHMEF)
```

```
dim(BPHMEF)
```

```
BPHMEFLAC <- merge(BPHMEF, BASE5, by=c("id_viv", "id_hogar", "id_per"))
```

```
names(BPHMEFLAC)
```

```
dim(BPHMEFLAC)
```

```
basecompleta <- merge(BPHMEFLAC, BASE6, by=c("id_viv", "id_hogar", "id_per"))
```

```
names(basecompleta)
```

```
dim(basecompleta)
```

## ANEXO C: Código en R para el análisis exploratorio de datos (AED)

```
#----- librerias -----  
library(dplyr) #manipulacion de datos  
library(ggplot2) #visualizacion  
library(caret) #visualizacion  
library(ggpubr) #unir visualizaciones  
library(PASWR)  
library(outliers)  
library(data.table)  
  
Cargamos los datos  
BASECOMPLETA <- read.csv("BASECOMPLETA.csv")  
dim(BASECOMPLETA)  
names(BASECOMPLETA)  
  
Estructura de los datos  
glimpse(BASECOMPLETA)  
  
Transformamos en factor las variables que consideramos cualitativas  
factor <- c("dronica","id_viv","id_hogar","id_per","FBCP1_4","FBCP1_6","FBCP1_7","FBCH_17","FSCCN_23","  
FBCH_1","FBCH_2","FBCH_3","FBCH_4","FBCH_5","FBCH_6","FBCH_7","FBCH_8","FBCH_9","FBCH_10",  
"FBCH_13","FBCH_14","FBCH_15","FBCH_16","FBCH_18","FBCP1_2","FBCP1_8","FIPA_4","FIPA_5","FIP  
A_6","FIPA_7","FIPA_10","FIPA_11","FIPA_12","FIPA_13","FIPA_15","FIPA_17","FIPA_19","FIPA_20","FIP  
A_22","FIPA_24","FIPA_26","FIPA_28","FIPA_30","FIPA_32","FIPA_33","FIPA_34","FIPA_35","FIPA_36","F  
BCP1_3","FSCE_1","FSCE_2","FSCE_5","FSCE_6","FSCE_7","FSCE_8","FSCE_9","FSCE_10","FSCE_14","FS  
CN_1","FSCN_2","FSCN_4","FSCN_5","FSCN_6","FSCN_7","FSCN_9","FSCN_11","FSCN_13","FSCCN_1","F  
SCCN_5","FSCCN_6","FSCCN_7","FSCCN_21","FIEL_1","FIEL_3","FIEL_4","FIEL_6","FIEL_7")  
for(i in factor){  
  BASECOMPLETA[,i] <- as.factor(BASECOMPLETA[,i])  
}  
glimpse(BASECOMPLETA)  
  
Estadísticas básicas  
summary(BASECOMPLETA)  
dim(BASECOMPLETA)  
  
Eliminación de datos duplicados  
s_n_duplicados <- BASECOMPLETA[!duplicated(BASECOMPLETA[,c("id_viv","id_hogar","id_per")]),]  
dim(s_n_duplicados)  
summary(s_n_duplicados)  
glimpse(s_n_duplicados)  
  
La variable dependiente dronica es de tipo factor de 2 niveles, por tanto, nuestro problema es de clasificación  
Variable dependiente : dronica  
ggplot(data = s_n_duplicados, aes(x = dronica, fill = dronica)) +  
  geom_bar() +  
  labs(title = "Distribucion variable dronica") +  
  theme_classic2()  
  
Estamos ante una variable no balanceada donde la clase 0 es mas numerosa que la clase 1  
prop.table(table(s_n_duplicados$dronica))*100
```

```
library(data.table)
```

```
s_n_duplicados <- data.table(s_n_duplicados)
```

### VISUALIZACION DE VARIABLES CUANTITATIVAS

*Realizamos un análisis estadístico de las variables cuantitativas*

```
library(e1071)
```

```
s_n_duplicados %>%
```

```
summarise_at( vars(FBCP1_1),
```

```
  funks (
```

```
    MEDIA= mean(., na.rm=TRUE),
```

```
    MEDIANA= median(., na.rm = TRUE),
```

```
    STDEV = sd(., na.rm = TRUE),
```

```
    KURTOSIS = kurtosis(., na.rm = TRUE),
```

```
    ASIMETRIA = skewness(., na.rm = TRUE),
```

```
    CV = (STDEV/MEDIA)*100,
```

```
    MIN = min(., na.rm = TRUE),
```

```
    MAX = max(., na.rm = TRUE)
```

```
  )
```

```
) %>% View
```

```
s_n_duplicados %>%
```

```
summarise_at( vars(FBCP1_5),
```

```
  funks (
```

```
    MEDIA= mean(., na.rm=TRUE),
```

```
    MEDIANA= median(., na.rm = TRUE),
```

```
    STDEV = sd(., na.rm = TRUE),
```

```
    KURTOSIS = kurtosis(., na.rm = TRUE),
```

```
    ASIMETRIA = skewness(., na.rm = TRUE),
```

```
    CV = (STDEV/MEDIA)*100,
```

```
    MIN = min(., na.rm = TRUE),
```

```
    MAX = max(., na.rm = TRUE)
```

```
  )
```

```
) %>% View
```

```
s_n_duplicados %>%
```

```
summarise_at( vars(FBCH_12),
```

```
  funks (
```

```
    MEDIA= mean(., na.rm=TRUE),
```

```
    MEDIANA= median(., na.rm = TRUE),
```

```
    STDEV = sd(., na.rm = TRUE),
```

```
    KURTOSIS = kurtosis(., na.rm = TRUE),
```

```
    ASIMETRIA = skewness(., na.rm = TRUE),
```

```
    CV = (STDEV/MEDIA)*100,
```

```
    MIN = min(., na.rm = TRUE),
```

```
    MAX = max(., na.rm = TRUE)
```

```
  )
```

```
) %>% View
```

s\_n\_duplicados %>%

```
summarise_at( vars(FSCE_3),  
  funs (  
    MEDIA= mean(., na.rm=TRUE),  
    MEDIANA= median(., na.rm = TRUE),  
    STDEV = sd(., na.rm = TRUE),  
    KURTOSIS = kurtosis(., na.rm = TRUE),  
    ASIMETRIA = skewness(., na.rm = TRUE),  
    CV = (STDEV/MEDIA)*100,  
    MIN = min(., na.rm = TRUE),  
    MAX = max(., na.rm = TRUE)  
  )  
) %>% View
```

s\_n\_duplicados %>%

```
summarise_at( vars(FSCE_4),  
  funs (  
    MEDIA= mean(., na.rm=TRUE),  
    MEDIANA= median(., na.rm = TRUE),  
    STDEV = sd(., na.rm = TRUE),  
    KURTOSIS = kurtosis(., na.rm = TRUE),  
    ASIMETRIA = skewness(., na.rm = TRUE),  
    CV = (STDEV/MEDIA)*100,  
    MIN = min(., na.rm = TRUE),  
    MAX = max(., na.rm = TRUE)  
  )  
) %>% View
```

s\_n\_duplicados %>%

```
summarise_at( vars(FSCE_12),  
  funs (  
    MEDIA= mean(., na.rm=TRUE),  
    MEDIANA= median(., na.rm = TRUE),  
    STDEV = sd(., na.rm = TRUE),  
    KURTOSIS = kurtosis(., na.rm = TRUE),  
    ASIMETRIA = skewness(., na.rm = TRUE),  
    CV = (STDEV/MEDIA)*100,  
    MIN = min(., na.rm = TRUE),  
    MAX = max(., na.rm = TRUE)  
  )  
) %>% View
```

s\_n\_duplicados %>%

```
summarise_at( vars(FSCE_13),  
  funs (  
    MEDIA= mean(., na.rm=TRUE),  
    MEDIANA= median(., na.rm = TRUE),
```

```

    STDEV = sd(., na.rm = TRUE),
    KURTOSIS = kurtosis(., na.rm = TRUE),
    ASIMETRIA = skewness(., na.rm = TRUE),
    CV = (STDEV/MEDIA)*100,
    MIN = min(., na.rm = TRUE),
    MAX = max(., na.rm = TRUE)
  )
) %>% View

```

s\_n\_duplicados %>%

```

summarise_at( vars(FSCCN_9),
  funs (
    MEDIA= mean(., na.rm=TRUE),
    MEDIANA= median(., na.rm = TRUE),
    STDEV = sd(., na.rm = TRUE),
    KURTOSIS = kurtosis(., na.rm = TRUE),
    ASIMETRIA = skewness(., na.rm = TRUE),
    CV = (STDEV/MEDIA)*100,
    MIN = min(., na.rm = TRUE),
    MAX = max(., na.rm = TRUE)
  )
) %>% View

```

s\_n\_duplicados %>%

```

summarise_at( vars(FSCCN_10),
  funs (
    MEDIA= mean(., na.rm=TRUE),
    MEDIANA= median(., na.rm = TRUE),
    STDEV = sd(., na.rm = TRUE),
    KURTOSIS = kurtosis(., na.rm = TRUE),
    ASIMETRIA = skewness(., na.rm = TRUE),
    CV = (STDEV/MEDIA)*100,
    MIN = min(., na.rm = TRUE),
    MAX = max(., na.rm = TRUE)
  )
) %>% View

```

s\_n\_duplicados %>%

```

summarise_at( vars(FSCCN_12),
  funs (
    MEDIA= mean(., na.rm=TRUE),
    MEDIANA= median(., na.rm = TRUE),
    STDEV = sd(., na.rm = TRUE),
    KURTOSIS = kurtosis(., na.rm = TRUE),
    ASIMETRIA = skewness(., na.rm = TRUE),
    CV = (STDEV/MEDIA)*100,
    MIN = min(., na.rm = TRUE),

```



```

MAX = max(., na.rm = TRUE)
)
) %>% View
s_n_duplicados %>%
summarise_at( vars(FSCCN_19),
  funs (
    MEDIA= mean(., na.rm=TRUE),
    MEDIANA= median(., na.rm = TRUE),
    STDEV = sd(., na.rm = TRUE),
    KURTOSIS = kurtosis(., na.rm = TRUE),
    ASIMETRIA = skewness(., na.rm = TRUE),
    CV = (STDEV/MEDIA)*100,
    MIN = min(., na.rm = TRUE),
    MAX = max(., na.rm = TRUE)
  )
) %>% View

```

```

s_n_duplicados %>%
summarise_at( vars(FSCCN_22),
  funs (
    MEDIA= mean(., na.rm=TRUE),
    MEDIANA= median(., na.rm = TRUE),
    STDEV = sd(., na.rm = TRUE),
    KURTOSIS = kurtosis(., na.rm = TRUE),
    ASIMETRIA = skewness(., na.rm = TRUE),
    CV = (STDEV/MEDIA)*100,
    MIN = min(., na.rm = TRUE),
    MAX = max(., na.rm = TRUE)
  )
) %>% View

```

```

s_n_duplicados %>%
summarise_at( vars(FIPA_2),
  funs (
    MEDIA= mean(., na.rm=TRUE),
    MEDIANA= median(., na.rm = TRUE),
    STDEV = sd(., na.rm = TRUE),
    KURTOSIS = kurtosis(., na.rm = TRUE),
    ASIMETRIA = skewness(., na.rm = TRUE),
    CV = (STDEV/MEDIA)*100,
    MIN = min(., na.rm = TRUE),
    MAX = max(., na.rm = TRUE)
  )
) %>% View

```

#### VISUALIZACION DE VARIABLES CATEGORICAS

Gráficos de las variables categóricas en función de la variable dependiente

```

#FBCP1_4
ggplot(s_n_duplicados, aes(x = FBCP1_4, fill = dcronica)) +
  geom_bar(position = position_dodge(preserve = "single")) +
  theme_gray()
prop.table(table(s_n_duplicados$FBCP1_4))*100
(table(s_n_duplicados$FBCP1_4,s_n_duplicados$dcronica))

# FBCP1_6
ggplot(s_n_duplicados,na.rm = TRUE,aes(x = FBCP1_6, fill = dcronica)) +
  geom_bar(position = position_dodge(preserve = "single")) +
  theme_gray()
prop.table(table(s_n_duplicados$FBCP1_6))*100
(table(s_n_duplicados$FBCP1_6,s_n_duplicados$dcronica))

# FBCP1_7
ggplot(s_n_duplicados, aes(x = FBCP1_7, fill = dcronica)) +
  geom_bar(position = position_dodge(preserve = "single")) +
  theme_gray()
prop.table(table(s_n_duplicados$FBCP1_7))*100
(table(s_n_duplicados$FBCP1_7,s_n_duplicados$dcronica))

# FBCP1_8
ggplot(s_n_duplicados, aes(x = FBCP1_8, fill = dcronica)) +
  geom_bar(position = position_dodge(preserve = "single")) +
  theme_gray()
prop.table(table(s_n_duplicados$FBCP1_8))*100
(table(s_n_duplicados$FBCP1_8,s_n_duplicados$dcronica))

# FBCH_1
ggplot(s_n_duplicados, aes(x = FBCH_1, fill = dcronica)) +
  geom_bar(position = position_dodge(preserve = "single")) +
  theme_gray()
prop.table(table(s_n_duplicados$FBCH_1))*100
(table(s_n_duplicados$FBCH_1,s_n_duplicados$dcronica))

# FBCH_2
ggplot(s_n_duplicados, aes(x = FBCH_2, fill = dcronica)) +
  geom_bar(position = position_dodge(preserve = "single")) +
  theme_gray()
prop.table(table(s_n_duplicados$FBCH_2))*100
(table(s_n_duplicados$FBCH_2,s_n_duplicados$dcronica))

# FBCH_3
ggplot(s_n_duplicados, aes(x = FBCH_3, fill = dcronica)) +
  geom_bar(position = position_dodge(preserve = "single")) +
  theme_gray()
prop.table(table(s_n_duplicados$FBCH_3))*100
(table(s_n_duplicados$FBCH_3,s_n_duplicados$dcronica))

```

## ANEXO D: Código en R para preprocesado de datos

```
library(caret)
library(tidyverse)
library(MASS)
library(dummies)
library(VIM)
library(DMwR)
library(funModeling)
library(rpart)
library(ggplot2)
library(ggpubr)
library(PASWR)
library(data.table)
library(ROSE)

set.seed(98561)
```

```
BASECOMPLETA <- read.csv("BASE_DEPURADA.csv")
```

### *Preprocesado de datos*

*Dividimos el conjunto de datos en un conjunto de entrenamiento y validación*

```
index <- createDataPartition(BASECOMPLETA$dchronica, p = 0.70, list = FALSE)
train <- BASECOMPLETA[index,]
test <- BASECOMPLETA[-index,]
```

### *Detección de valores ausentes*

#### *Conjunto de entrenamiento*

```
df_status(train)
```

#### *Conjunto de validación*

```
df_status(test)
```

Elimino todas las variables que tengan más del 50% de Nas

```
train$FBCP1_2 <- NULL
train$FIPA_7 <- NULL
train$FIPA_8 <- NULL
train$FIPA_9 <- NULL
train$FIPA_10 <- NULL
train$FIPA_16 <- NULL
train$FIPA_18 <- NULL
train$FIPA_23 <- NULL
train$FIPA_25 <- NULL
train$FIPA_27 <- NULL
train$FIPA_29 <- NULL
train$FIPA_31 <- NULL
train$FSCN_3 <- NULL
train$FSCN_4 <- NULL
```

```
train$FSCN_10 <- NULL
train$FSCN_12 <- NULL
train$FSCN_14 <- NULL
train$FSCCN_20 <- NULL
train$FIEI_2 <- NULL
train$FIEI_3 <- NULL
train$FIEI_5 <- NULL
train$FIEI_6 <- NULL
train$FIEI_7 <- NULL
```

*Imputación con el método k vecinos más cercanos*

*Conjunto de entrenamiento*

```
cleantrain <- knnImputation(train)
summary(cleantrain)
anyNA(cleantrain)
```

*Conjunto de validación*

```
cleantest <- knnImputation(test)
summary(cleantest)
anyNA(cleantest)
```

*Centrado y escalado de datos*

*Conjunto de entrenamiento*

```
train <- preProcess(cleantrain,method = c("center","scale"))
cleantrain <- predict(pre_ce, newdata = cleantrain)
summary(cleantrain)
```

*Conjunto de validación*

```
test <- preProcess(cleantest,method = c("center","scale"))
cleantest <- predict(pre_cv, newdata = cleantest)
summary(cleantest)
```

*Variables Dummys*

*Conjunto de entrenamiento*

```
train_d <- dplyr::select(train,-id_viv,-id_hogar,-id_per)
dummy <- dummyVars(~.-dcronica, data = train_d)
```

*#Predecimos sobre el conjunto de datos*

```
train_d_t <- as.data.frame(predict(dummy,train_d))
```

*#Añadimos la variable dependiente*

```
train_d_t$dcronica <- train_d$dcronica
summary(train_d_t)
```

*Conjunto de validación*

```
test_d <- dplyr::select(test,-id_viv,-id_hogar,-id_per)
dummy <- dummyVars(~.-dcronica, data = test_d)
```

*#Predecimos sobre el conjunto de datos*

```
test_d_t <- as.data.frame(predict(dummy,test_d))
```

*#Añadimos la variable dependiente*

```
test_d_t$dchronica <- test_d$dchronica
```

```
summary(test_d_t)
```

*Variables con varianza cero o próxima a cero*

*Conjunto de validación*

```
nearZeroVar(train_d_t, saveMetrics = TRUE)
```

Eliminación de variables con varianza cero o próxima a cero

```
ntrain <- train_d_t %>%
```

```
dplyr::select(-"FBCEP1_6.afroecuatoriano/afrodescendiente?",-"FBCEP1_6.blanco/a?",-"FBCEP1_6.montuvio/a?",-"FBCEP1_6.mulato/a?",-"FBCEP1_6.negro/a?",-"FBCEP1_6.otra, cuál? (especifique)",-"FBCEP1_7.divorciado",-"FBCEP1_7.unión de hecho",-"FBCEP1_7.viudo",-"FBCEH_1.otro, cuál?",-"FBCEH_1.río/mar",-"FBCEH_2.choza",-"FBCEH_2.covacha",-"FBCEH_2.cuarto/s en casa de inquilinato",-"FBCEH_2.otra, cuál?",-"FBCEH_3.otro, cuál?",-"FBCEH_3.palma/paja/hoja?",-"FBCEH_4.adobe/tapia?",-"FBCEH_4.bahareque (caña, carrizo revestido)",-"FBCEH_4.caña o estera?",-"FBCEH_4.otra, cuál?",-"FBCEH_5.caña?",-"FBCEH_5.mármol/marmetón?",-"FBCEH_5.otro, cuál?",-"FBCEH_5.tierra?",-"FBCEH_6.carro repartidor/triciclo?",-"FBCEH_6.otro, cuál?",-"FBCEH_6.pila o llave pública?",-"FBCEH_7.por tubería fuera del edificio, lote o terreno?",-"FBCEH_8.letrina?",-"FBCEH_9.empresa eléctrica pública?",-"FBCEH_9.ninguno?",-"FBCEH_9.planta eléctrica privada?",-"FBCEH_9.vela, candil, mechero, gas?",-"FBCEH_10.botan a la calle/ quebrada/ río?",-"FBCEH_10.contratan el servicio?",-"FBCEH_10.la entierran?",-"FBCEH_10.otra, cuál?",-"FBCEH_13.no sabe",-"FBCEH_13.otro tratamiento?",-"FBCEH_14.electricidad? (inducción)",-"FBCEH_14.no cocina",-"FBCEH_18.anticresis y arriendo?",-"FBCEH_18.otra, cuál?",-"FBCEH_18.recibida por servicios?",-"FBCEP1_8.Ninguno o Centro de Alfabetización",-"FIPA_13.no sabe / no responde",-"FIPA_15.no sabe / no responde",-"FIPA_17.no sabe / no responde",-"FIPA_20.no sabe / no responde",-"FIPA_22.no sabe / no responde",-"FIPA_26.no sabe / no responde",-"FIPA_30.no",-"FIPA_30.no sabe / no responde",-"FIPA_30.si",-"FIPA_34.no sabe / no responde",-"FIPA_35.no sabe / no responde",-"FBCEP1_3.no quería tener hijos?",-"FSCE_1.no",-"FSCE_1.si",-"FSCE_2.consejo provincial/unidad municipal de salud",-"FSCE_2.fundación/ ong",-"FSCE_2.hospital ff.aa/policía",-"FSCE_2.junta de beneficencia",-"FSCE_2.otro, cuál?",-"FSCE_2.partera",-"FSCE_2.seguro social campesino",-"FSCE_5.no sabe / no responde",-"FSCE_6.consejo provincial/unidad municipal de salud",-"FSCE_6.fundación/ ong",-"FSCE_6.hospital ff.aa/ policía",-"FSCE_6.junta de beneficencia",-"FSCE_6.otro, cuál?",-"FSCE_6.seguro social campesino",-"FSCE_7.aux. enfermería",-"FSCE_7.comadróna o partera",-"FSCE_7.enfermera",-"FSCE_7.familiar",-"FSCE_7.otro, cuál?",-"FSCE_7.usted misma",-"FSCE_9.no sabe",-"FSCE_9.posmaduro",-"FSCE_14.fundación/ ong",-"FSCE_14.hospital ff.aa/policía",-"FSCE_14.hospital/clínica/dispensario del iess",-"FSCE_14.junta de beneficencia",-"FSCE_14.no recuerda",-"FSCE_14.otro, cuál?",-"FSCE_14.partera",-"FSCE_14.seguro social campesino",-"FSCN_1.no",-"FSCN_1.si",-"FSCN_5.muy pequeño",-"FSCN_5.no sabe",-"FSCCN_1.no",-"FSCCN_1.si",-"FSCCN_5.no recuerda",-"FSCCN_6.consejo provincial/unidad municipal de salud",-"FSCCN_6.fundación/ ong",-"FSCCN_6.hospital ff.aa/policía",-"FSCCN_6.junta de beneficencia",-"FSCCN_6.no recuerda",-"FSCCN_6.otro, cuál?",-"FSCCN_6.seguro social campesino",-"FSCCN_7.no",-"FSCCN_7.si",-"FSCCN_21.36-42",-"FSCCN_21.43-47",-"FSCCN_21.48-59",-"FIEI_1.no sabe / no responde")
```

*Balanceo de clases: Conjunto de entrenamiento*

```
dchronica_sub <- ovun.sample(dchronica~.,  
  data = cleantrain,  
  method = "under",  
  seed = 98561,  
  N=4564)$data  
table(dchronica_sub$dchronica)
```

## ANEXO E: Código en R utilizado para construir y evaluar el modelo de árboles de decisión

### *Creamos el modelo*

```
model_dcronica <- rpart(dcronica ~ .,  
  data = ntrain,  
  method = "class",  
  control = rpart.control(cp = 0.001))
```

### *Representación gráfica del modelo*

```
rpart.plot(model_dcronica, extra = 1, type = 2, digits = 2)
```

### *Predicción sobre el árbol creado*

```
predict <- predict(model_dcronica,  
  newdata = ntest,  
  type = "class")
```

### *Matriz de confusión*

```
confusionMatrix(predict, ntest$dcronica, positive = "1")
```

### *Poda del árbol utilizando la CP óptima*

```
printcp(model_dcronica)  
  
model_dcronica$cptable[5,]  
  
model_dcronica_opcp <- prune(model_dcronica, cp = 0.0046468)
```

### *Representación gráfica del árbol podado*

```
rpart.plot(model_dcronica_opcp, extra = 1, type = 2, digits = 2)
```

### *Predicción sobre el árbol podado*

```
predict_opcp <- predict(model_dcronica_opcp, ntest, type = "class")
```

### *Matriz de confusión*

```
confusionMatrix(predict_opcp, ntest$dcronica, positive = "1")
```

### *Curva ROC*

```
predict_model_prob <- predict(model_dcronica_opcp, newdata = ntest, type = "prob")[,2]  
pr <- prediction(predict_model_prob, ntest$dcronica)  
perf <- performance(pr, measure = "tpr", x.measure = "fpr")  
plot(perf, colorize = T, main = "Curva ROC Arbol de decision")
```

### *AUC*

```
auc(ntest$dcronica, predict_model_prob)
```

## ANEXO F: Código en R utilizado para construir y evaluar el modelo Gradient Boosting

```
grid <- expand.grid(n.trees = 1000,
  shrinkage = 0.1,
  interaction.depth = 3:5,
  n.minobsinnode = 10)
control <- trainControl(method = "cv",
  number = 3)
model_grid_depth <- train(dcronica~.,
  data = ntrain,
  distribution = "bernoulli",
  method = "gbm",
  tuneGrid = grid,
  trControl = control,
  verbose = F)
plot(model_grid_depth)

ntrain$dcronica <- as.numeric(ntrain$dcronica)-1
model_gbm <- gbm(dcronica~.,
  data = ntrain,
  distribution = "bernoulli",
  n.trees = 3000,
  cv.folds = 3,
  shrinkage = model_grid$bestTune$shrinkage,
  n.cores = NULL,
  interaction.depth = model_grid_depth$bestTune$interaction.depth)
gbm.perf(model_gbm)

summary(model_gbm,
  method = relative.influence,
  plotit = F)
```

### *Predicciones*

```
predicciones <- predict(model_gbm,ntest,type = "response")
```

### *Matriz de confusión*

```
predicciones_class <- ifelse(predicciones > 0.5, 1, 0)
confusionMatrix(factor(predicciones_class),factor(ntest$dcronica), positive = "1")
```

### *Curva ROC*

```
roc.curve(ntest$dcronica, predicciones,plotit = T)
```

## ANEXO G: Código en R utilizado para construir y evaluar el modelo de Regresión logística

### *Modelo con la base completa*

```
mod_logit_1 <- glm(dcronica~.,data=datos_train,family=binomial(link="logit"))
```

```
summary(mod_logit_1)
```

### *Modelo con las variables significativas*

#### *Modelo logit por pasos hacia atrás*

```
mod_logit_2 <- glm(dcronica~FBCP1_6.indígena. +FBCP1_7.separado+FBCH_17.no+FBCH_4.asbesto.cemento..fi  
brolit..+FBCH_12+FBCH_15.no+FBCP1_1+FBCP1_5+FBCP1_8.Educación.Media.Bachillerato+FBCP1_8.Superio  
r+FIPA_21+FIPA_33.no.sabe...no.responde+FSCE_3+FSCE_9.prematuro+FSCE_12+FSCN_2.gramos+FSCN_2.kil  
os.gramos+FSCN_2.libras.onzas+FSCN_5.más.grande+FSCN_5.pequeño+FSCN_13.no+FSCCN_21.19.23+FSCCN  
_21.31.35+FIPA_2+FIEI_1.no+FIEI_1.si,data=datos_train,family=binomial(link="logit"))
```

```
summary(mod_logit_2)
```

### *Modelo con las variables significativas*

```
mod_logit_b <- glm(dcronica~FBCP1_6.indígena. +FBCP1_7.separado+FBCH_17.no+FBCH_12+ FBCP1_1+FBC  
P1_5+FBCP1_8.Educación.Media.Bachillerato+FBCP1_8.Superior+FIPA_33.no.sabe...no.responde+FSCE_3+FSCE  
_9.prematuro+FSCE_12+FSCN_5.más.grande+FSCN_5.pequeño+FSCCN_21.19.23+FSCCN_21.31.35+FIEI_1.no+  
FIEI_1.si,data=datos_train,family=binomial(link="logit"))
```

```
summary(mod_logit_b)
```

### *Prueba Omnibus*

```
with(mod_logit_b, null.deviance-deviance)
```

### *Valor p del estadístico de prueba*

```
with(mod_logit_b, pchisq(null.deviance-deviance,df.null-df.residual,lower.tail = F))
```

### *Odds y odds-ratio*

```
exp(coef(mod_logit_b))
```

### *Matriz de confusión*

```
predicted_value <- predict(mod_logit_b, datos_test,type = "response")
```

```
predicted_class <- ifelse(predicted_value > 0.5, "1", "0")
```

```
performance_data <- data.frame(observed = datos_test$dcronica,  
predicted = predicted_class)
```

```
positive <- sum(performance_data$observed=="1")
```

```
negative <- sum(performance_data$observed=="0")
```

```
predicted_positive <- sum(performance_data$predicted=="1")
```

```
predicted_negative <- sum(performance_data$predicted=="0")
```

```
total <- nrow(performance_data)
```

```
data.frame(positive, negative, predicted_positive, predicted_negative)
```



```
VP <- sum(performance_data$observed=="1" & performance_data$predicted=="1")
VN <- sum(performance_data$observed=="0" & performance_data$predicted=="0")
FP <- sum(performance_data$observed=="0" & performance_data$predicted=="1")
FN <- sum(performance_data$observed=="1" & performance_data$predicted=="0")
data.frame(VP,VN,FP,FN)
```

*Medidas de bondad de ajuste*

```
tasa_error <- (FP+FN)/total
```

*Curva ROC y AUC*

```
roc.curve(datos_test$dchronica, predicted_value, plotit = T)
```

## ANEXO H: Código en R utilizado para construir la aplicación web, pasos detallados para su publicación e Interfaz de usuario de la aplicación web interactiva

### *Código*

```
library(shiny)

library(shinydashboard)

library(DT)

library(rpart)

library(rpart.plot)

library(data.table)

header <- dashboardHeader(

  title = "ENSANUT 2018",

  tags$li(a(href = 'http://shinyapps.com',

            icon("power-off"),

            title = "Back to Apps Home"),

          class = "dropdown" )

sidebar <- dashboardSidebar(

  sidebarMenu(

    menuItem("Modelo logit", icon = shiny::icon("line-chart"),

             menuSubItem("Datos", tabName = "data_1", icon = shiny::icon("table")),

             menuSubItem("Resultados", tabName = "rest_1", icon = shiny::icon("th-list")),

             menuSubItem("Predicciones", tabName = "pred_1", icon = shiny::icon("area-chart"))),

    menuItem("Arboles de decision", icon = shiny::icon("sitemap"),

             menuSubItem("Datos", tabName = "data", icon = shiny::icon("table")),

             menuSubItem("Resultados", tabName = "rest", icon = shiny::icon("th-list")),

             menuSubItem("Grafica", tabName = "graph", icon = shiny::icon("sitemap")),

             menuSubItem("Predicciones", tabName = "pred", icon = shiny::icon("area-chart"))

    ) ) )

body <- dashboardBody(

  tabItems(

    tabItem("data_1",
```

```

      dataTableOutput(outputId = "tabla_1")),
  tabItem("rest_1",
    verbatimTextOutput(outputId = "res_1")),
  tabItem("pred_1",
    verbatimTextOutput(outputId = "predict_1") ),
  tabItem("data",
    dataTableOutput(outputId = "tabla")),
  tabItem("rest",
    verbatimTextOutput(outputId = "res")),
  tabItem("graph",
    plotOutput(outputId = "graf")),
  tabItem("pred",
    verbatimTextOutput(outputId = "predict") ) ) )
ui <- dashboardPage(skin = "purple",header, sidebar, body)
server <- function(input, output){
  datapath_1 <- paste(getwd(), "data", "base_entrenamiento_balaceada.csv", sep="/")
  datareact_1 <- reactive({
    input$go
    isolate(
      read.table(datapath_1, sep=",", header = T) )
  })
  output$tabla_1 <- renderDataTable({
    datareact_1()
  })
  output$res_1 <- renderPrint({
    input$go_1
    isolate(
      readRDS("mod_logit_b.rds") )
  })
  datapath <- paste(getwd(), "data", "base_entrenamiento.csv", sep="/")

```

```

datareact <- reactive({

input$go

isolate(

  read.table(datapath, sep=",", header = T ) )

})

output$tabla <- renderDataTable({

  datareact()

})

output$res <- renderPrint({

  input$go

isolate(

  readRDS("model_tree.rds") )

})

output$graf <- renderPlot({

  rpart.plot(readRDS("model_tree.rds"),extra = 1, type = 2, digits = 2)

})

}

```

### *Pasos de la publicación de la aplicación*

1. Ir a <https://www.shinyapps.io>
2. Crearse una cuenta
  - 2.1. Instalar en R el package **rsconnect**
  - 2.2. Conectarse: crear una clave, copiar el código y ejecutarlo en un R script
3. Ejecutar el código:
 

```

library(rsconnect)

rsconnect::deployApp('D:/La/Ruta/hasta/su/app')

```
4. Cada publicación que se realice de una app, se crea un directorio en la carpeta llamada **rsconnect**

## Interfaz de usuario

C:\Users\jorg\Desktop\APLICACION TEST GORGIA/DASHBOARD\_2 - Shiny

https://127.0.0.1:7712/ Open in browser Admin

### ENSANUT 2016

Show 25 entries Search:

FBCP1_4.no	FBCP1_4.si	FBCP1_6.indigena	FBCP1_6.mestizo.a	FBCP1_7.casado	FBCP1_7.separado	FBCP1_7.soltero	FBCP1_7.viudo
0	1	0	1	0	0	1	0
1	0	0	0	0	0	0	1
1	0	1	0	1	0	0	0
1	0	0	1	0	1	0	0
1	0	0	1	0	0	0	1
1	0	0	0	0	0	0	1
0	1	0	1	1	0	0	0
1	0	0	0	1	0	0	0
0	1	1	0	0	1	0	0
0	1	1	0	0	1	0	0
0	1	0	1	0	1	0	0
1	0	0	1	0	0	0	1
1	0	0	0	0	0	0	1

C:\Users\jorg\Desktop\APLICACION TEST GORGIA/DASHBOARD\_2 - Shiny

https://127.0.0.1:7712/ Open in browser Admin

### ENSANUT 2016

```
Call: glm(formula = dcronica ~ FBCP1_6.indigena + FBCP1_7.separado +
  FBCP1_7.no + FBCP1_7.si + FBCP1_7.viudo + FBCP1_8.educación.Medio.Bachillerato +
  FBCP1_8.Superior + FBCP1_8.no.sabe...no.responde + FBCP1_8 +
  FBCP1_9.prematuro + FBCP1_9 + FBCP1_9.más.grande + FBCP1_9.pequeño +
  FBCP1_9.19.20 + FBCP1_9.21.25 + FBCP1_9.no + FBCP1_9.si,
  family = binomial(link = "logit"), data = datos_train)
```

Coefficients:

(Intercept)	-0.02540	FBCP1_6.indigena	0.47556
FBCP1_7.separado	0.10324	FBCP1_7.no	0.30667
FBCP1_7.si	-0.07380	FBCP1_7.viudo	-0.01378
FBCP1_7.viudo	0.09479	FBCP1_8.educación.Medio.Bachillerato	-0.14417
FBCP1_8.Superior	-0.27693	FBCP1_8.no.sabe...no.responde	-0.32971
FBCP1_8.no.sabe...no.responde	0.01594	FBCP1_9.prematuro	0.17886
FBCP1_9.prematuro	0.10926	FBCP1_9.más.grande	-0.08126
FBCP1_9.más.grande	0.27129	FBCP1_9.pequeño	0.43828
FBCP1_9.pequeño	0.33943	FBCP1_9.19.20	0.76528
FBCP1_9.19.20	0.04182	FBCP1_9.21.25	
FBCP1_9.21.25		FBCP1_9.no	
FBCP1_9.no		FBCP1_9.si	
FBCP1_9.si			

Number of Observations: 4563 Total (1) = 4563 (1) = 4563

C:\Users\jorg\Desktop\APLICACION TESS GORDA\DASHBOARD\_2 - Rely

https://127.0.0.1:7711/ Open in Browser

### ENSANUT 2016

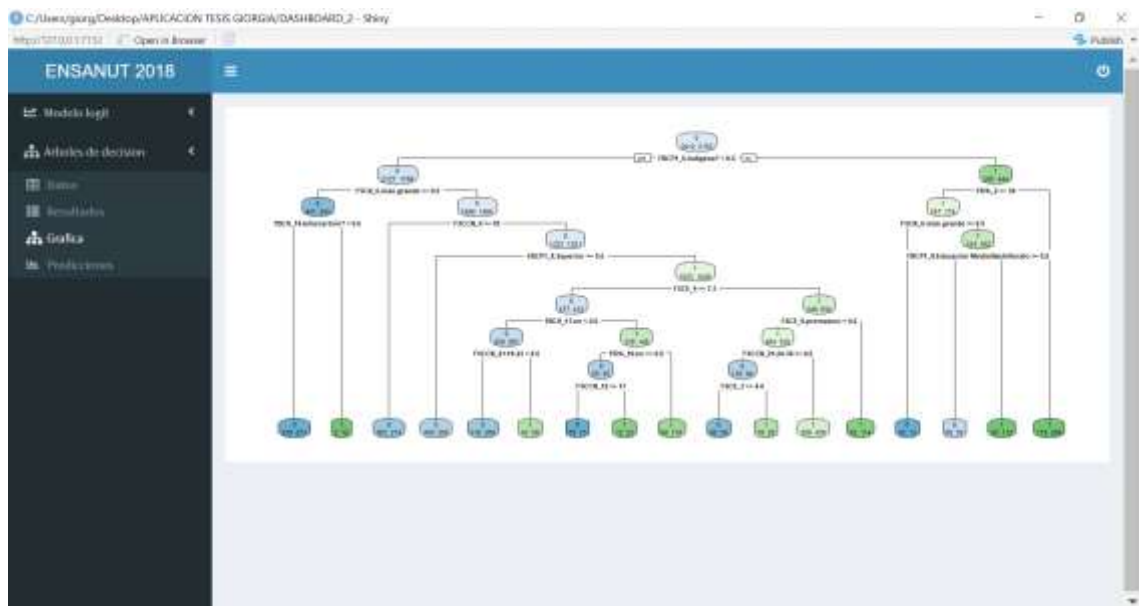
- Modelo logit
- Artículos de decisión
- Inicio
- Resultados
- Gráfica
- Producciones

```

n= 4564
node). split, n, loss, yval, (yprob)
* detotas terminal node

1) root 4564 2152 0 (0.0286838 0.4717162)
2) FSCPI_5<=1.021947 0.5 1825 1708 0 (0.0544204 0.6053716)
4) FSCQ_5<=0.5 684 243 0 (0.6447368 0.3512632)
8) FBOH_14<=0.070 231 0 (0.6152229 0.3847771) *
9) FBOH_14<=0.070 231 2 1 (0.1428871 0.8571129) *
5) FSCQ_5<=0.5 2739 1465 0 (0.5059642 0.4940358)
10) FSCCN_0<=17.51732 377 234 0 (0.6291161 0.3708839) *
11) FSCCN_0<=17.51732 377 234 0 (0.5119809 0.4880191) *
22) FBCPL_8<=0.5 189 285 0 (0.3940594 0.6059406) *
23) FBCPL_8<=0.5 189 285 1 (0.6044418 0.3955582)
46) FSCX_4<=7.008945 889 412 0 (0.3205579 0.6794421)
92) FBOH_17<=0.5 826 287 0 (0.5734824 0.4265176)
184) FSCCN_21<=23 0.5 324 288 0 (0.5890534 0.4109466) *
185) FSCCN_21<=23 0.5 324 288 1 (0.4109466 0.5890534) *
93) FBOH_17<=0.5 826 287 1 (0.4408992 0.5591008)
186) FBOH_19<=0.5 85 35 0 (0.5862553 0.4137447)
372) FSCCN_23<=16 92189 51 13 0 (0.7458988 0.2541012) *
373) FSCCN_23<=16 92189 51 13 1 (0.2541012 0.7458988) *
187) FBOH_19<=0.5 85 35 0 (0.5862553 0.4137447) *
47) FSCX_4<=7.008945 1386 546 1 (0.4627139 0.5372861)
94) FSCX_9<=0.5 1894 484 1 (0.4839717 0.5160283)
188) FSCCN_21<=24 0.5 380 34 0 (0.5800000 0.4200000)
376) FSCX_3<=4.446887 154 96 0 (0.4363034 0.5636966) *
377) FSCX_3<=4.446887 154 96 1 (0.5636966 0.4363034) *
189) FSCCN_21<=24 0.5 380 34 1 (0.4177154 0.5822846) *
95) FSCX_9<=0.5 176 42 1 (0.3522727 0.6477273) *

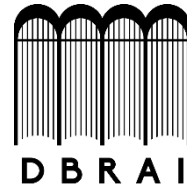
```





**ESCUELA SUPERIOR POLITÉCNICA DE CHIMBORAZO**

**DIRECCIÓN DE BIBLIOTECAS Y RECURSOS PARA EL  
APRENDIZAJE Y LA INVESTIGACIÓN**



**UNIDAD DE PROCESOS TÉCNICOS**  
**REVISIÓN DE NORMAS TÉCNICAS, RESUMEN Y BIBLIOGRAFÍA**

**Fecha de entrega:** 12 / 01 / 2021

<b>INFORMACIÓN DEL AUTOR/A (S)</b>
<b>Nombres – Apellidos:</b> Giorgia Nohelia Congacha Ortega
<b>INFORMACIÓN INSTITUCIONAL</b>
<b>Facultad:</b> Ciencias
<b>Carrera:</b> Ingeniería en Estadística informática
<b>Título a optar:</b> Ingeniera en Estadística informática
<b>f. Analista de Biblioteca responsable:</b> Lic. Luis Caminos Vargas Mgs.



0539 -DBRAI-UPT-2020