



Pontificia Universidad  
Católica del Ecuador

PONTIFICIA UNIVERSIDAD CATÓLICA DEL ECUADOR

FACULTAD DE INGENIERÍA

“ANÁLISIS EXPLORATORIO DE DATOS E IDENTIFICACIÓN DE AGENTES QUE INFLUYEN EN LA  
DESNUTRICIÓN CRÓNICA DE NIÑOS MENORES A CINCO AÑOS DEL ECUADOR MEDIANTE LA  
APLICACIÓN DE TÉCNICAS DE CIENCIA DE DATOS.”

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL TÍTULO DE MÁSTER EN SISTEMAS DE  
INFORMACIÓN, MENCIÓN DATA SCIENCE

CATHERINE ELEANA YÁNEZ CARRERA

TUTOR: MSC. EDUARDO MONTERO BERMUDEZ

QUITO, 2023

## **AGRADECIMIENTO**

Agradezco infinitamente a Dios por darme la fortaleza de continuar cuando decaía y por poner de guías terrenales a dos seres humanos excepcionales; mis padres a quienes agradezco por siempre empujarme a lograr mis objetivos, por su ejemplo como seres humanos y profesionales, gracias por estar en todo momento y por ser firmes conmigo en este objetivo.

A mi director de tesis Msc. Eduardo Montero por compartir sus conocimientos y haber hecho este trabajo posible, y a todas las personas que de alguna u otra forma aportaron para que pueda culminar esta meta más en mi vida, gracias.

## **DEDICATORIA**

*A la bendición que no esperaba y que tanto necesitaba, mi hermana Nicole Mariana por ser mi mejor  
amiga y mi motivación para continuar.*

*A mi compañero de aventuras, mi hermano Eduardo Alexis a quien admiro por su perseverancia,  
fortaleza y por ser mi fuente de inspiración.*

*A ellos dedico mi esfuerzo y dedicación.*

## RESUMEN

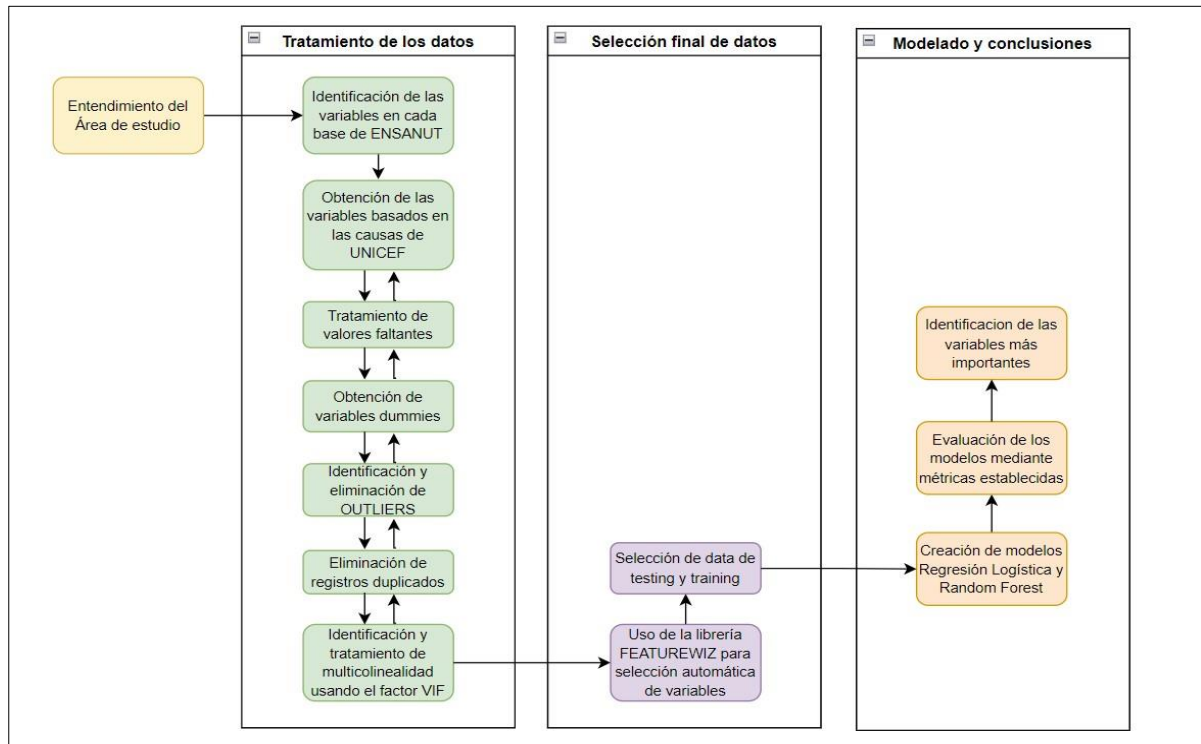
El presente estudio permitió determinar los factores más relevantes que influyen en la desnutrición crónica infantil en ecuatorianos menores de 5 años, para este propósito se usó la base de datos abierta obtenida de la Encuesta Nacional de Salud y Nutrición realizada por el Instituto Nacional de Estadísticas y Censos en el año 2018, donde constan 20.356 observaciones de niños menores de 5 años y el 28,73% pertenecen a niños con DCI. Después de un análisis exploratorio de datos se consiguió un total de 101 variables disponibles, de las cuales se seleccionó de forma automática diez de ellas, haciendo uso de la librería *featurewiz*. Con este nuevo grupo de variables se creó un modelo de Regresión Logística y un *Random Forest*, al medir el rendimiento de los modelos con la métrica *accuracy*, ambos obtuvieron más del 68%, mientras que, al medir la capacidad de diferenciar entre valores positivos y negativos mediante la métrica AUC, ambos modelos alcanzaron más de 0,6. Sin embargo según los valores de la matriz de confusión, el modelo de Regresión Logística es mejor para predecir falsos positivos, es decir predice mejor los casos sin DCI, tiene una precisión del 45,65% y predice correctamente el 97,7% de casos de menores sin DCI, mientras que el modelo *Random Forest* es mejor para predecir verdaderos positivos, es decir que predice mejor los casos con DCI, adicionalmente tiene una precisión de 68,2% y predice de forma correcta el 22,18% de casos de menores con DCI, bajo este criterio el modelo Random Forest es el modelo seleccionado que mejor explica la desnutrición crónica infantil. Las variables más importantes para construir los modelos fueron: segunda dosis de la vacuna contra el Neumococo: no, área donde habita el menor: rural, nivel de instrucción de la madre: Educación Básica, etnia del menor: indígena, material del techo de la vivienda del menor: palma/ paja/ hoja, baño de la vivienda del menor: no tiene, fuente del agua que bebe el menor: embotellada /envasada, material del piso de la vivienda del menor: tierra. Para conocer similitudes entre las categorías de algunas variables se realizó el análisis de correspondencia, para esto se dividió la data en observaciones con DCI = 1 y DCI = 0, entre las similitudes más relevantes están: “Nivel de instrucción de la madre: Educación Básica” con “Área donde habita el menor: rural”, “Etnia del menor: indígena” con “Baño de la vivienda del menor: no tiene”, “Baño de la vivienda del menor: tiene” con “Etnia del menor: diferente a indígena”, “Área donde habita el menor: urbana” con “Instrucción de la madre: diferente a Educación Básica” las categorías que no tienen ninguna similitud son: “Material de la vivienda del menor: palma /paja /hoja” y “Material del piso de la vivienda: tierra”.

## ***ABSTRACT***

The present study allowed determining the most important factors that influence chronic child malnutrition in Ecuadorians under five years of age, for this purpose, the open database obtained from the National Health and Nutrition Survey conducted by the National Institute of Statistics and Census in 2018 was used, where 20.356 observations of children under five years of age are recorded and 28,73% belong to children with ICD. After an exploratory data analysis, a total of 101 available variables were obtained, from which ten of them were automatically selected, making use of the featurewiz library, with this variables a Logistic Regression model and a Random Forest were created, when measuring the performance of the models with the accuracy metric, both obtained more than 68% and when measuring the ability to differentiate between positive and negative values using the AUC metric, both models reached more than 0,6 however according to the confusion matrix, the Logistic Regression model is better at predicting false positives, i.e., it predicts better the cases without ICD, has an accuracy of 45,65% and correctly predicts the 97,7% of cases of minors without ICD, while the Random Forest Model is better at predicting true positives, i.e. it predicts better the cases with ICD, has an accuracy of 68,2% and correctly predicts 22,18% of cases of minors with ICD, under this criterion, the Random Forest model is the selected model that best explains child chronic malnutrition. The most important variables for constructing the models were: the second dose of Pneumococcal vaccine: no, the area where the minor lives: rural, mother's level of education: basic education, ethnicity of the minor: indigenous, ceiling's material: palm/straw/leaf, bathroom in the minor's home: none, source of the water the minor drinks: bottled/package, floor material of the minor's home: dirt. To determine similarities between the categories of some variables, the correspondence analysis was carried out, for which the data was divided into observations with ICD=1 and ICD =0, among the most relevant similarities are: "Mother's level of education: basic education" with "Area where the minor lives: rural", "Ethnicity of the minor: indigenous" with "Bathroom of the minor's home: no has" with "Ethnicity of the minor: other than indigenous", "Area where the minor lives: urban" with "Mother's education: other than basic education" the categories that do not have any similarity are: "Material of the minor's dwelling: palm/straw/leaf" and "Material of the house floor: dirt",

## RESUMEN DE LAS ETAPAS EN LA IMPLEMENTACIÓN DE CIENCIA DE DATOS

Para un mejor entendimiento del estudio, se presenta un gráfico resumen de todas las etapas realizadas. El gráfico 1 muestra de forma detallada el paso a paso seguido para obtener los resultados mencionados anteriormente.



**Figura 1:** Gráfico resumen de la metodología de ciencia de datos.

**Fuente:** Gráfico obtenido mediante la herramienta viewer.diagrams.net

## TABLA DE CONTENIDOS

AGRADECIMIENTO.....	2
DEDICATORIA.....	¡Error! Marcador no definido.
RESUMEN .....	4
ABSTRACT .....	5
RESUMEN DE LAS ETAPAS EN LA IMPLEMENTACIÓN DE CIENCIA DE DATOS .....	6
ÍNDICE DE FIGURAS.....	10
ÍNDICE DE TABLAS .....	12
1. CAPITULO 1: SITUACIÓN ACTUAL .....	13
1.1. PROBLEMÁTICA .....	13
1.2. Antecedentes .....	14
1.3. Justificación e importancia .....	15
1.4. Objetivo General .....	17
1.5. Objetivos Específicos .....	17
2. Capítulo 2: Marco Teórico .....	18
2.1. Conceptos básicos de desnutrición crónica infantil .....	18
2.1.1. Desnutrición .....	18
2.1.2. Causas de la desnutrición crónica .....	18
2.2. Encuesta nacional de salud, salud reproductiva y nutrición (ENSANUT).....	19
2.2.1. Objetivo de la Encuesta ENSANUT .....	19
2.2.2. Marco conceptual y metodológico de ENSANUT .....	20
2.2.3. Delimitación del estudio .....	20
2.3. Teoría de ciencia de datos .....	20
2.3.1. Análisis exploratorio de datos (EDA) .....	20
2.3.2. Métodos para resolver el problema de valores faltantes .....	21
2.3.3. Eliminación de valores atípicos (outliers) .....	23
2.4. Bibliotecas de python para la selección de variables .....	23

2.5. Aprendizaje automático (Machine Learning) .....	24
2.5.1. Algoritmos de clasificación .....	25
2.5.2. Regresión Logística.....	25
2.5.3. Random Forest .....	26
2.6. Herramientas para visualizar la eficacia de un algoritmo de clasificación.....	28
2.6.1. Matriz de confusión .....	28
2.6.2. Curva ROC .....	29
2.7. ANÁLISIS DE CORRESPONDENCIAS .....	29
2.7.1. Ventajas del análisis de correspondencia .....	30
2.7.2. DESVENTAJAS DEL ANALISIS DE CORRESPONDENCIA .....	30
3. Capítulo 3: Marco Metodológico .....	31
3.1. Tipo de investigación.....	31
3.2. Fuente de información .....	31
3.3. Delimitación de la población.....	31
3.4. Dimensión de la muestra .....	31
3.5. Recolección de información .....	31
3.6. Determinación de la variable dependiente e independientes.....	34
3.7. Análisis y estudio estadístico.....	41
3.7.1. Análisis exploratorio de datos.....	41
3.7.2. Preprocesamiento de datos.....	42
3.7.3. Análisis de la multicolinealidad .....	47
3.7.4. SELECCIÓN DE VARIABLES INDEPENDIENTES .....	47
3.7.5. Desarrollo y evaluación de los modelos de clasificación .....	48
4. CAPÍTULO 4. Análisis de resultados.....	50
4.1. Análisis exploratorio de datos .....	50
4.2. Modelo de Regresión Logística .....	55
4.3 Modelo Random Forest .....	56
4.4. COEFICIENTE GINI DE LAS VARIABLES QUE INTERVIENEN EN LOS MODELOS .....	56



4.5. MODELO SELECCIONADO PARA LA IDENTIFICACIÓN DE LOS FACTORES QUE INFLUYEN EN LA DESNUTRICIÓN CRÓNICA INFANTIL .....	62
4.6. ANÁLISIS DE CORRESPONDENCIA DE LAS PRINCIPALES VARIABLES CATEGÓRICAS .....	62
4.6.1. ANÁLISIS DE CORRESPONDENCIA DE OBSERVACIONES DE NIÑOS MENORES DE 5 AÑOS CON DESNUTRICIÓN CRÓNICA INFANTIL .....	63
4.6.2. ANÁLISIS DE CORRESPONDENCIA DE OBSERVACIONES DE NIÑOS MENORES DE 5 AÑOS SIN DESNUTRICIÓN CRÓNICA INFANTIL .....	64
5. CAPITULO 5: CONCLUSIONES Y RECOMENDACIONES .....	66
5.1. CONCLUSIONES.....	66
5.2. RECOMENDACIONES .....	67
6. BIBLIOGRAFIA .....	68
ANEXO 2. CÓDIGO DE LAS FUNCIONES CREADAS PARA FACILITAR EL ANÁLISIS DE LOS DATOS .....	71
ANEXO3. MATRIZ DE CONFUSIÓN DEL MODELO REGRESIÓN LOGÍSTICA Y RANDOM FOREST.....	74

## ÍNDICE DE FIGURAS

<b>Figura 1:</b> Gráfico resumen de la metodología de ciencia de datos. ....	6
<b>Figura 2:</b> Porcentajes de Desnutrición Crónica Infantil del según las provincias del Ecuador .....	14
<b>Figura 3:</b> Clasificación de las causas de la Desnutrición Crónica Infantil. ....	19
<b>Figura 4:</b> Guía práctica para un análisis exploratorio de datos. ....	21
<b>Figura 5:</b> Estructura de un modelo Randon Forest. ....	26
<b>Figura 6:</b> Matriz de confusión. ....	29
<b>Figura 7:</b> ETL de la fuente de datos ENSANUT. ....	41
<b>Figura 8:</b> Diagrama de caja de la variable tiempo_control_posparto. ....	44
<b>Figura 9:</b> Diagrama de caja de la variable primer_control_medico. ....	45
<b>Figura 10:</b> Diagrama de caja de la variable edad_fin_lactancia.....	45
<b>Figura 11:</b> Gráfico del top 10 de variables que hacen un modelo óptimo. ....	48
<b>Figura 12:</b> Porcentaje de DCI por Etnia. ....	50
<b>Figura 13:</b> Porcentaje de DCI por el estado civil de los padres. ....	51
<b>Figura 14:</b> Porcentaje de DCI por el baño en la vivienda. ....	52
<b>Figura 15:</b> Porcentaje de DCI por la instrucción de la madre. ....	52
<b>Figura 16:</b> Porcentajes de DCI por nutrientes recibidos durante el embarazo .....	53
<b>Figura 17:</b> Porcentaje de DCI si el menor recibió o no desparasitante.....	53
<b>Figura 18:</b> Porcentaje de DCI si el menor recibió o no micronutrientes. ....	54
<b>Figura 19:</b> Porcentaje de DCI de menores donde el padre está presente. ....	54
<b>Figura 20:</b> Porcentaje de DCI en menores donde hubo o no hubo control médico en los últimos 30 días.....	55
<b>Figura 21:</b> Curva ROC para el modelo de Regresión Logística. ....	55
<b>Figura 22:</b> Curva ROC para el modelo de Random Forest. ....	56
<b>Figura 23:</b> Gráfico bivariado de la variable 'tierra' con respecto a la variable dependiente DCI. ....	58
<b>Figura 24:</b> Gráfico bivariado de la variable 'Educación Básica con respecto a la variable dependiente DCI.....	58
<b>Figura 25:</b> Gráfico bivariado de la variable 'indígena' con respecto a la variable dependiente DCI....	59
<b>Figura 26:</b> Gráfico bivariado de la variable 'BANIONo tiene' con respecto a la variable dependiente DCI.....	59
<b>Figura 27:</b> Gráfico bivariado de la variable 'AGUA_BEBERagua embotellada' con respecto a la variable dependiente DCI. ....	59
<b>Figura 28:</b> Gráfico bivariado de la variable 'rural' con respecto a la variable dependiente DCI.....	60

<b>Figura 29:</b> Gráfico bivariado de la variable 'TECHOpalma/paja/hoja' con respecto a la variable dependiente DCI.....	60
<b>Figura 30:</b> Gráfico bivariado de la variable 'NEUMOCC2_NO' con respecto a la variable dependiente DCI.....	61
<b>Figura 31:</b> Porcentajes de niños con y sin DCI agrupados en niños menores a 30 meses y mayores a 30 meses.....	61
<b>Figura 32:</b> Porcentajes de niños con y sin DCI por el número de hijos que viven en casa.....	<b>¡Error!</b>
<b>Marcador no definido.</b>	
<b>Figura 33:</b> Análisis de correspondencia de las principales variables categóricas con la data de menores ecuatorianos de 5 años con DCI. ....	63
<b>Figura 34:</b> Análisis de correspondencia de las principales variables categóricas con la data de menores ecuatorianos de 5 años sin DCI. ....	64
<b>Figura 35:</b> Matriz de confusión del modelo Regresión Logística .....	74
<b>Figura 36:</b> Matriz de confusión del modelo Random Forest .....	74

## ÍNDICE DE TABLAS

<b>Tabla 1:</b> Encuesta Nacional de Salud y Nutrición: Formulario 1 - Hogar. ....	32
<b>Tabla 2:</b> Encuesta Nacional de Salud y Nutrición: Formulario 2 - MEF.....	32
<b>Tabla 3:</b> Encuesta Nacional de Salud y Nutrición: Formulario 3 - Salud sexual y reproductiva en hombres. ....	33
<b>Tabla 4:</b> Encuesta Nacional de Salud y Nutrición: Formulario 4 - Salud sexual y reproductiva en hombres. ....	33
<b>Tabla 5:</b> Encuesta Nacional de Salud y Nutrición: Formulario 5 – Desarrollo infantil. ....	33
<b>Tabla 6:</b> Variables independientes considerando las causas clasificadas por UNIECF. ....	34
<b>Tabla 7:</b> Diagrama de caja de las variables independientes. ....	46
<b>Tabla 8:</b> Variables seleccionadas mediante la biblioteca featurewiz ....	48
<b>Tabla 9:</b> Porcentaje y cantidad de menores de 5 años vivos y fallecidos con y sin desnutrición crónica infantil. ....	50
<b>Tabla 10:</b> Variables con los 10 coeficientes de significancia más altos, del modelo: Regresión Logística.....	56
<b>Tabla 11:</b> Variables con los 10 coeficientes de significancia más altos, del modelo: Random Forest. ....	57
<b>Tabla 12:</b> Variables que intervienen en el análisis de correspondencia .....	63

## **1. CAPITULO 1: SITUACIÓN ACTUAL**

### **1.1. PROBLEMÁTICA**

La desnutrición crónica infantil (DCI) forma parte del grupo de problemas sociales que afecta a la humanidad, en todo el mundo más de 50 millones de niños menores de cinco años están afectados y según la Organización Mundial de la Salud (OMS), la desnutrición es causante de un tercio de todas las muertes infantiles, la situación es realmente grave en América Latina y el Caribe, en el caso de Ecuador se registra el 27% de niños y niñas menores a 2 años con desnutrición crónica (DCI) y se constituye como el segundo país de América Latina con el mayor índice de desnutrición infantil. (Instituto Nacional de Estadísticas y Censos, 2018)

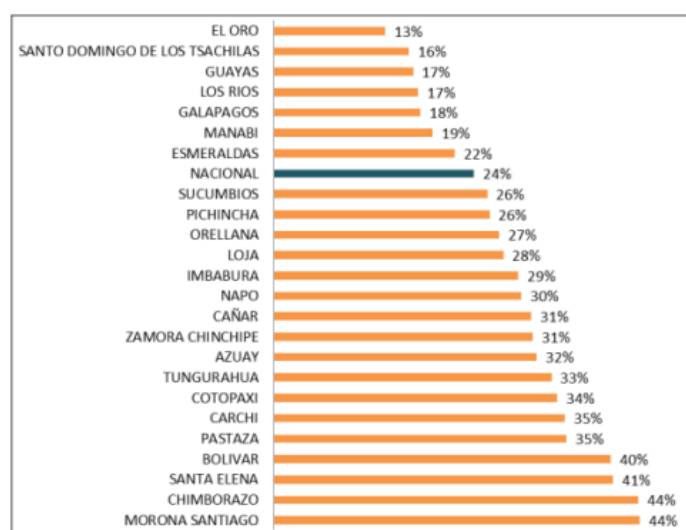
Varias organizaciones nacionales e internacionales han implementado planes activos para nutrición y asistencia alimentaria con el fin de erradicar la desnutrición infantil en Ecuador, sin embargo, la curva de DCI no ha cambiado. (UNICEF, 2019)

## 1.2. ANTECEDENTES

Según la UNICEF en 54 países de ingresos bajos y medios, en el periodo de gestación, es donde empieza las deficiencias de crecimiento y se prolongan hasta cerca de los 24 meses de edad. La probabilidad de reparar este daño en los próximos meses de vida es casi nula; se calcula que entre el 60 y 80% de neonatos fallecidos presentaron bajo peso al nacer.

En todo el mundo existen ocho millones de niños menores a 5 años que corren el riesgo de morir por bajo peso en comparación a su altura, si no reciben de manera inmediata tratamiento y alimentación (UNICEF, 2022). Según el Comité Español de la agencia de la ONU para los refugiados “la región más afectada por el hambre actualmente es el sur de Asia, donde el 34,4% de los menores de cinco años sufre desnutrición crónica. África mantiene sus índices en el 33%”. (Agencia de la ONU para los Refugiados, 2020). Las comunidades más pobres son las que sufren mayor desnutrición infantil. En su gran mayoría las muertes que se dan por este problema son en países de ingresos bajos y medianos.

Existe un estudio realizado por el MIES donde se menciona que el índice de la DCI pasó de 28,9% a 25,3% y del 2013 al 2014 el índice baja al 23,9%, en la figura 2 es visible que las provincias Morona Santiago, Chimborazo, Santa Elena y Bolívar es donde se concentran las cifras más altas. (Malo, Mejía, & Vinuesa, 2015)



**Figura 2:** Porcentajes de Desnutrición Crónica Infantil del según las provincias del Ecuador

**Fuente:** Ministerio de Inclusión Económica y Social, 2015

El Informe de Impacto Social y Económico de la Malnutrición reflejó que aproximadamente 27 mil estudiantes en el Ecuador repiten el año escolar debido a motivos relacionados con la DCI. El costo estimado anual es de USD 27 millones. Uno de los motivos de la repetición escolar es que hubo un crecimiento anormal en los primeros años de infancia, “el cerebro de las personas se forma hasta un 90% durante los primeros años de vida, aquí se sientan los cimientos para el potencial de aprendizaje

de una persona por el resto de su vida” (UNICEF, 2019). A nivel nacional la DCI en infantes menores de 5 años se presenta en 413.913 niños y en el rango de 5 a 11 años llega a los 340.000 niños siendo una edad avanzada en niños para tratar la DCI, la consecuencia en este último grupo es un retraso del 15% en su crecimiento (Ministerio de Salud Pública, 2019)

En un artículo de UNICEF del 2021, se menciona que la DCI “No solo afecta a la persona con desnutrición crónica, sino también tiene un impacto negativo en el desarrollo económico y social de los países. Los gastos derivados de la malnutrición en ámbitos de salud, educación y baja productividad equivalen al 4,3% del producto interno bruto ecuatoriano.” (UNICEF, 2019). En el 2021 en varias comunidades de Chunchi según los reportes médicos de la Unidad de Salud de ese sector, casi todos los niños menores de 5 años presentan un alarmante retraso de peso y talla. Este problema es bastante habitual, tal es el caso de 59 comunidades de Guamote, cantón de la provincia de Chimborazo. (Márquez, 2021)

En informes no oficiales después de la pandemia, se conoce que la tasa de DCI ya llega a los 30 puntos, así expuso a Ecuavisa el Secretario Técnico de Ecuador CRECE SIN DESNUTRICIÓN INFANTIL, Erwin Ronquillo, "no es un problema de los últimos 10 años tenemos décadas así", también reconoció que la mayor tasa de DCI se evidencia en provincias de la Sierra y la Amazonía, destaca que es una problemática presente en todo el país y no solo en los estratos económicos bajos, insiste en que el problema de la DCI no se reduce a la mala alimentación, sino que se trata de un problema estructural en el que influyen aspectos como la pobreza, el desempleo, la desigualdad y hasta la formación académica de los padres. (Ronquillo, 2021; citado por Bermeo).

### **1.3. JUSTIFICACIÓN E IMPORTANCIA**

A pesar de los planes que se han activado para erradicar la destrucción crónica infantil (DCI), según las Naciones Unidas, la curva de la DCI en menores de cinco años casi no se ha movido. Entre los años 2014 y 2018, aumentó de 24,8% a 27,2% en niños menores de dos años y se concluyó que uno de cada cuatro ecuatorianos menores de cinco años tiene DCI. Por todo lo expuesto es necesario analizar con profundidad los datos existentes, recolectados por la Encuesta Nacional de Salud, Salud Reproductiva y Nutrición y mediante la aplicación de técnicas de ciencia de datos, identificar los factores que contribuyen a que la curva de la DCI en Ecuador se mantenga.

En un artículo titulado: “Ecuador el alto costo del hambre y la desnutrición crónica”, se menciona que el hambre en el Ecuador tiene una relación directamente proporcional con la DCI y es el resultado de la dificultad económica para acceder a una alimentación suficiente. Ecuador es el segundo país con los peores índices del continente, uno de cada cuatro niños ecuatorianos tendrá incapacidad. Si se logra disminuir considerablemente la DCI, el hambre ya no tendrá costo a largo plazo y el desarrollo social

en el país mejorará. Considerando que la DCI es un problema social multifactorial, es necesario realizar análisis y estudios de todas las posibles causas y proyectarlas en visualizaciones apropiadas para que se entienda el nivel de la problemática y posibles soluciones.

En la encuesta realizada por ENSANUT en el 2018, se refleja que existe varias fuentes con datos sobre el entorno, condiciones de vivienda, lactancia y salud del niño, estos datos pueden ser el punto de partida para encontrar los principales factores que influyen en la DCI, y permitirnos así obtener conclusiones que posiblemente no han sido detectadas. La desnutrición crónica infantil se ha tomado vidas humanas y ha empeorado la calidad de vida de muchos niños ecuatorianos, las cifras no son alentadoras ya que según medios no oficiales se presume que aumentó del 27% en 2018 al 30% en 2020, para disminuirla es importante analizar a profundidad los factores que influyen en la DCI, existen varias investigaciones sobre la DCI en Ecuador, así como también varias organizaciones que han planteado programas para reducirla como es el caso de UNICEF que actualmente apoya al MSP en la normativa Establecimientos Amigos de la Madre y del Niño, con asistencia técnica, fomentando y apoyando las capacidades de profesionales de la salud, también apoya a las Juntas Administradoras de Agua Potable para abastecer de agua potable a familias de zonas de Imbabura y Pichincha (UNICEF, 2019), pese a todos los esfuerzos realizados las cifras no mejoran y hoy por hoy el porcentaje de DCI en Ecuador se ha mantenido.

Aportar con un estudio sobre uno de los problemas sociales más significativos del país ayudará de una u otra manera a encontrar soluciones, y dará luz verde a futuras políticas para que inviertan en un plan que contrarreste la DCI en el país.



#### **1.4. OBJETIVO GENERAL**

Identificar factores que influyen en la desnutrición crónica infantil de niños menores a cinco años en el Ecuador.

#### **1.5. OBJETIVOS ESPECÍFICOS**

- Realizar el análisis exploratorio y limpieza de datos de la base abierta proporcionada por ENSANUT 2018.
- Aplicar técnicas de ciencia de datos, para obtener los factores influyentes en la desnutrición crónica infantil en el Ecuador.
- Aplicar técnicas de visualización de datos que permitan una mejor comprensión y difusión de los resultados del análisis.
- Obtener conclusiones relevantes para que futuras políticas ayuden a contrarrestar la desnutrición crónica infantil.

## **2. CAPÍTULO 2: MARCO TEÓRICO**

### **2.1. CONCEPTOS BÁSICOS DE DESNUTRICIÓN CRÓNICA INFANTIL**

#### **2.1.1. DESNUTRICIÓN**

La desnutrición crónica infantil (DCI) es un problema social grave que se define como el bajo peso en bebés y niños pequeños, debido principalmente a la falta de nutrientes y a una alimentación inadecuada. En un gran porcentaje se presenta desde la edad de gestación y tiene consecuencias de por vida. La DCI debe ser tratada antes del retraso en el crecimiento, este factor afecta el desarrollo y crecimiento físico y cognoscitivo de una persona, se considera uno de los mayores problemas de salud pública en los países más pobres del mundo, las causas y consecuencias de la DCI son pluridimensionales, razón por la cual es bastante complicado entender esta condición y por ende aún más complicado encontrar soluciones sociales y políticas a este problema. Se necesita una planificación acompañada de estrategias y la colaboración de múltiples sectores para mejorar la nutrición infantil, por lo que son necesarias intervenciones en las áreas de salud, medio ambiente, agricultura, agua y saneamiento, infraestructura, educación, etc. (Acción contra el hambre, 2020)

#### **2.1.2. CAUSAS DE LA DESNUTRICIÓN CRÓNICA**

La desnutrición crónica es un problema multifactorial, es decir, su origen se debe a múltiples causas, se pueden agrupar en tres: causas inmediatas, subyacentes y básicas (Acción contra el hambre, 2020):

- *Causas inmediatas:* El consumo de alimentos de mala calidad y en una inapropiada cantidad, además de los riesgos por una mala higiene de los alimentos por el entorno. Las infecciones y las diarreas, que son muy comunes en los niños de países pobres, provocan que sus organismos absorban pobremente los nutrientes que consumen. Estos padecimientos derivan, entre otros, de la escasez de agua.
- *Causas subyacentes:* Una causa subyacente basada en las causas descritas anteriormente son los factores sociales y económicos, políticos e institucionales y factores del medio ambiente. Las causas subyacentes de la desnutrición crónica en adultos y niños son aquellos factores que aumentan las posibilidades de que ocurran las causas inmediatas, como el inadecuado acceso a comidas nutritivas que tienen millones de hogares en el mundo, además del uso adecuado y manipulación que puedan darle.
- *Causas básicas:* Son los factores sociales obvios como pobreza, desigualdad, cultura, educación, entre otras.

Una causa que se ha evidenciado es la inadecuada protección y cuidado de los niños, por ejemplo, la falta de la leche materna en los primeros meses de vida considerando que esta es vital para un recién nacido, en la figura 3 se muestra la clasificación de las causas de la DCI:



**Figura 3:** Clasificación de las causas de la Desnutrición Crónica Infantil.  
**Fuente:** (Ramírez, 2014)

## 2.2. ENCUESTA NACIONAL DE SALUD, SALUD REPRODUCTIVA Y NUTRICIÓN (ENSANUT)

### 2.2.1. OBJETIVO DE LA ENCUESTA ENSANUT

La Encuesta Nacional de Salud y Nutrición – ENSANUT 2018 es un instrumento estadístico que tiene como objetivo principal establecer factores sobre las más relevantes problemáticas sociales y el estado de la salud de los ecuatorianos, estas consideraciones sirven para evaluar y crear políticas públicas referentes a la salud y nutrición.

La necesidad de disponer de datos sobre salud, condiciones de vida y calidad alimenticia de la población fueron la base para la creación de ENSANUT, su objetivo es subsanar los problemas sociales mediante programas que ayuden a controlar problemas sociales y de salud, y de esta forma lograr reducirlos a niveles que fomenten calidad y apropiada salud pública. La primera encuesta de ENSANUT se dio en el año 2012, recopiló datos de la situación nutricional de la población ecuatoriana y las prácticas de lactancia materna, así como las condiciones de salud sexual y reproductiva de las mujeres en edad fértil. En el 2018 fue necesario evaluar la situación de salud de la población ecuatoriana para conocer la evolución de los principales problemas de salud que aborda esta encuesta. Es necesidad del Estado Ecuatoriano contar con información actualizada que permita evaluar las políticas asociadas a la erradicación de la DCI en menores de 5 años. (Instituto Nacional de Estadísticas y Censos, 2018)

### **2.2.2. MARCO CONCEPTUAL Y METODOLÓGICO DE ENSANUT**

ENSANUT fundamenta sus más importantes indicadores en los conceptos y métodos sobre lactancia materna y malnutrición desarrollados por la Organización Mundial de la Salud (OMS), así como la normalización de hitos basados en los picos de crecimiento considerados por este organismo. Se siguen también algunas recomendaciones de las variables utilizados en la Encuesta de Indicadores Múltiples por Conglomerados de UNICEF, como es el caso de las preguntas incluidas sobre disciplina y desarrollo infantil.

### **2.2.3. DELIMITACIÓN DEL ESTUDIO**

La ENSANUT, tiene como punto de partida las encuestas de demografía y salud, en individuos ecuatorianos seleccionados de forma *random*, enfocando el análisis en mujeres en edad reproductiva y niños menores de 5 años, que son los sectores más vulnerables y con edades en las que se puede actuar para resolver problemas estructurales, se realizó un listado previo de los sectores muestrales, del cual se obtuvo certeza de la población a investigar. Según la Organización Mundial de la Salud, el término malnutrición se refiere a la falta, y el desequilibrio del consumo de cantidad de calorías y nutrientes de un ser humano. Abarca tres grandes grupos de afecciones:

- La desnutrición, que incluye un peso insuficiente con respecto a la talla, el retraso del crecimiento y un peso bajo para la edad del individuo.
- La malnutrición está relacionada con los micronutrientes, que incluye la falta de vitaminas o minerales importantes o el exceso de micronutrientes.
- El sobrepeso, la gordura y las enfermedades derivadas de la mala calidad en la alimentación como enfermedades del corazón, sangre y algunos tipos de cáncer.

La talla insuficiente respecto de la edad se denomina retraso del crecimiento, es resultado de una DCI frecuentemente asociada a circunstancias socioeconómicas escasas, una malnutrición y mala salud de la madre, presencia constante de enfermedades y/o a una mala nutrición o malos cuidados para el niño pequeño. La falta de un crecimiento normal hace que los niños tengan retrasos en su desarrollo físico y cognitivo mientras que los niños que tienen un peso bajo en comparación a su edad sufran insuficiencia ponderal. Si un niño tiene este problema el daño colateral es un retraso del crecimiento (Organización Mundial de la Salud, 2021).

## **2.3. TEORÍA DE CIENCIA DE DATOS**

### **2.3.1. ANÁLISIS EXPLORATORIO DE DATOS (EDA)**

Es el proceso mediante el cual se usan sinopsis numéricas y gráficas para explorar los datos y reconocer vínculos entre variables. Mediante este análisis exploratorio se puede encontrar datos defectuosos,

como *outliers* proporcionados por observaciones con valores diferentes a los normales, además podemos encontrar similitudes y relaciones entre unas variables y otras, se puede hallar métricas estadísticas principales que ayudarán en el conocimiento y unificación de los datos.

En la figura 4 se muestra una guía práctica para realizar un correcto análisis exploratorio de datos donde el objetivo principal es buscar patrones que puedan guiar al reconocimiento de las probables causas del problema en este caso de la DCI en niños ecuatorianos menores de cinco años. Se exploran las variables de manera individual, luego en pares, y luego un conjunto de variables a la vez. (Statistical Discovery, 2022)



**Figura 4:** Guía práctica para un análisis exploratorio de datos.  
**Fuente:** (Ministerio de Asuntos Económicos y Transformación Digital, 2021)

Los objetivos principales del análisis exploratorio de datos son:

- Conocer la distribución de las variables disponibles para el estudio
- Conocer y entender las relaciones entre las variables del conjunto de datos
- Conocer los valores faltantes en la recopilación y saber cómo subsanar este problema.
- Conocer *outliers* en los datos, y saber cómo actuar dependiendo del valor atípico identificado.
- Conocer tipos de datos y hacer conversiones en caso de ser necesario considerando el objetivo del estudio.

Los objetivos mencionados anteriormente se alcanzan mediante la implementación de estrategias, métodos y técnicas de ciencia de datos.

### **2.3.2. MÉTODOS PARA RESOLVER EL PROBLEMA DE VALORES FALTANTES**

Los valores faltantes son datos no existentes en la base de datos, es el problema más común en la recolección de datos. Hay varias causas por las que existen datos faltantes, de estas causas depende la forma de solucionar este problema, el manejo y solución correcto de este problema es

indispensable para la creación de un modelo útil (Dagnino, 2014), a continuación, se detallan las dos técnicas más frecuentes usadas para afrontar esta problemática:

- *Eliminar muestras o variables que tienen datos faltantes:* Es la eliminación de las variables que abarcan datos faltantes, el principal problema de esta técnica es que al eliminar toda la variable también se pierde una gran cantidad de datos útiles para el estudio, es recomendable aplicar esta técnica solo si la cantidad de valores nulos supera el 50% de las observaciones y previo a un análisis adecuado.
- *Imputar los valores perdidos (reemplazarlos por estimaciones):* El objetivo de esta técnica es sostener la mayor cantidad de datos para el estudio, a continuación, enumeramos las principales formas para hacer imputación y retener la mayor cantidad de datos válidos:
  1. Media, Mediana y Moda: Consiste en reemplazar los valores faltantes por valores estadísticos obtenidos a partir de los valores válidos: media, mediana o moda. La media es una estimación lógica para una observación elegida aleatoriamente de una distribución normal, la mediana es utilizada en el caso de que la variable tenga una distribución sesgada mientras que la moda consiste en reemplazar los datos faltantes por el valor más común.
  2. Última observación llevada hacia adelante: Esta técnica es usada en datos de series temporales, consiste en reemplazar un valor faltante por el último valor observado.
  3. Siguiente observación realizada hacia atrás: Es un método contrario al anterior, toma la siguiente observación después del valor faltante para sustituirlo por este.
  4. Imputación de punto común: Esta técnica es frecuente en una escala de clasificación parecida al valor medio y más habitual para datos ordinales. Consiste en sustituir el valor faltante por el valor más común.
  5. Añadir una categoría para capturar el N/A: Es la técnica más utilizada en variables categóricas, consiste en crear una nueva categoría (faltante) en la variable, los valores faltantes se agrupan bajo una nueva etiqueta.
  6. Imputación de categorías frecuentes: Es una técnica para tratar los valores faltantes en variables categóricas, que consiste en reemplazar todos los valores faltantes de una variable por la categoría más común de los valores válidos.
  7. Imputación de valores arbitrarios: Consiste en sustituir un valor faltante de una variable por un valor arbitrario, este debe ser diferente a la media, mediana o moda.
  8. Añadir una variable para capturar N/A: Consiste en captar la importancia de los faltantes creando una variable adicional que indique si los datos faltan para esa observación (1) o no (0). La variable añadida solo tomará los valores 0 y 1.

9. Imputación por muestreo aleatorio: Consiste en escoger una observación *random* del conjunto de datos válidos de la variable y utilizar ese valor *random* para llenar la N/A. (Roy, 2020)

### **2.3.3. ELIMINACIÓN DE VALORES ATÍPICOS (OUTLIERS)**

Los *outliers* o valores atípicos son datos que difieren numéricamente a los demás valores de la variable, son aquellos valores fuera del rango donde se encuentran la mayor cantidad de observaciones. Los valores atípicos se pueden dar por varias razones la principal razón, es por error humano, es decir que el encuestador se equivocó en la recolección de datos mostrando así un comportamiento de la variable distinto al normal.

La existencia de *outliers* afecta de manera significativa a la construcción de un modelo de *machine learning*. La eliminación o no de los *outliers* depende de tres factores:

1. *ERROR*: Un ejemplo práctico es si tenemos una muestra de un grupo de niños y tenemos una variable *ALTURA* con un valor de tres metros, este es un error en la recolección de datos (ningún niño mide 3 metros de altura), en este caso se debe eliminar los valores fuera del rango.
2. *LIMITES*: En otros casos, podemos tener valores que salen del “grupo medio”, pero es necesario mantener este valor para que no se afecte el aprendizaje del modelo.
3. *PUNTO DE INTERÉS*: Este caso es cuando los valores fuera de rango son el objeto de estudio, por lo que se decide mantenerlos. (Detección de outliers en Python, 2020)

### **2.4. BIBLIOTECAS DE PYTHON PARA LA SELECCIÓN DE VARIABLES**

Existe una gran cantidad de bibliotecas de *Python* que permiten una selección automática de las mejores variables independientes para la creación de un modelo con buen rendimiento, una de ellas es *sklearn.feature\_selection*, esta es una librería útil para la selección de características y para disminuir la dimensionalidad, tiene como objetivo mejorar el rendimiento del modelo, esta librería permite eliminar variables bajo algunos contextos: (Desarrolladores de scikit learn, 2023)

- Eliminación de entidades con baja varianza: En este caso se eliminan las variables independientes cuya varianza tiende a cero.
- Selección de características univariantes: Consiste en seleccionar variables haciendo uso de pruebas estadísticas.
- Selección de variables usando *SelectFromModel*: El *SelectFromModel* se puede usar junto a cualquier estimador que asigne importancia a cada variable mediante un atributo específico como *coef\_* o *feature\_importances\_*.

Otra de las librerías de *Python* para la selección automática de variables es *featurewiz*, esta librería permite la selección automática de variables para optimizar el rendimiento del aprendizaje automático con un mínimo esfuerzo, entre algunas de las ventajas de esta librería están:

1. Proporciona uno de los mejores algoritmos de selección automática de variables independientes: selección de características de redundancia mínima. (Comunidad de *Python*, 2023)
2. Elige la cantidad adecuada de variables no correlacionadas que aportan de manera significativa al *target*.
3. Su uso es fácil, y dispone de varias funciones útiles para la selección de características.
4. Es una librería que se actualiza constantemente por lo que se han resultado la mayor cantidad de errores en su implementación.

## 2.5. APRENDIZAJE AUTOMÁTICO (MACHINE LEARNING)

Es una ciencia de la inteligencia artificial mediante la cual las computadoras desarrollan la habilidad de un aprendizaje automático y, partiendo de este aprendizaje realizan predicciones. Dentro de este aprendizaje existen dos tipos:

- Aprendizaje Supervisado

Es uno de los métodos en que las máquinas “aprenden”, es aplicado cuando tenemos una variable que queremos predecir y explicar, este aprendizaje usa un conjunto de datos categorizados para entrenar un modelo. Este aprendizaje usa los datos disponibles recopilados a través del tiempo para que el algoritmo aprenda de estos y pueda predecir el valor de una variable específica.

Dentro de este tipo de aprendizaje tenemos dos tipos de algoritmos: clasificación y regresión, siendo la principal diferencia, la variable a predecir, en el primer caso es categórica, mientras que en el caso de regresión la variable a predecir es de tipo numérica continua.

Los algoritmos más representativos bajo esta categoría son:

- Regresión Logística.
- Árboles de decisión.
- *Random Forest*.
- Regresión por mínimos cuadrados.
- *Support Vector Machines* (SVM).

- Aprendizaje no Supervisado

También es un método en que una máquina “aprende”, a diferencia del aprendizaje supervisado, este permite encontrar nuevas características, en este caso los datos disponibles no están categorizados,



el objetivo del modelo es dar sentido a los datos disponibles a través de nuevas características o patrones, el aprendizaje no supervisado usa el agrupamiento. (TIBCO, 2023)

### **2.5.1. ALGORITMOS DE CLASIFICACIÓN**

Los algoritmos de clasificación predicen el valor de una variable categórica (característica), es utilizado para identificar una categoría de un conjunto de nuevas observaciones, se entrena al algoritmo a partir del set de datos disponibles de *training*, con este aprendizaje el algoritmo clasifica las observaciones dadas.

El primer paso es dividir el total del conjunto de datos disponibles en dos subconjuntos: entrenamiento (*training*) y pruebas (*test*), los datos de *training* se usan para preparar y precisar los parámetros del modelo de clasificación, mientras que los datos de *test* son usados para pruebas del funcionamiento y comportamiento correcto del modelo creado. Lo ideal es que los datos de *test* no sean los mismos datos de *training*. (Lantz, Métodos de clasificación., 2019)

### **2.5.2. REGRESIÓN LOGÍSTICA**

Es un modelo usado para establecer la probabilidad de que un evento suceda. Este modelo es semejante a la regresión lineal, la diferencia es que en este caso la variable a predecir es binaria (1 o 0).

La regresión logística se puede usar para medir dos tipos de variables: variable explicativa (característica) y variable binaria (objetivo). Un claro ejemplo es al predecir la presencia de una enfermedad, los síntomas son las características y la variable objetivo es la presencia o no de la enfermedad. Es uno de los algoritmos más usados para la clasificación de dos clases.

#### Consideraciones para la regresión logística

La regresión logística da por hecho los siguientes puntos para el correcto funcionamiento:

- La variable objetivo es binaria.
- La variable objetivo puede tener dos valores únicamente: 1 deseado y 0 no deseado.
- Las variables independientes que intervienen en el modelo deben ser significativas.
- Las variables independientes deben ser independientes entre sí, deben tener poca o ninguna multicolinealidad.
- Debe usarse para muestras masivas.

#### Ventajas y desventajas de la Regresión Logística

##### Ventajas

- Es eficiente y necesita pocos recursos computacionales.
- Es fácilmente interpretable.
- Es fácil de regular por lo que los resultados están correctamente calculados.
- Funciona de forma más eficiente si las variables independientes involucradas están relacionadas con la variable objetivo.
- Es fácil de implementar y entrenar, por lo que es una excelente base para calcular el desempeño de otros algoritmos más complejos

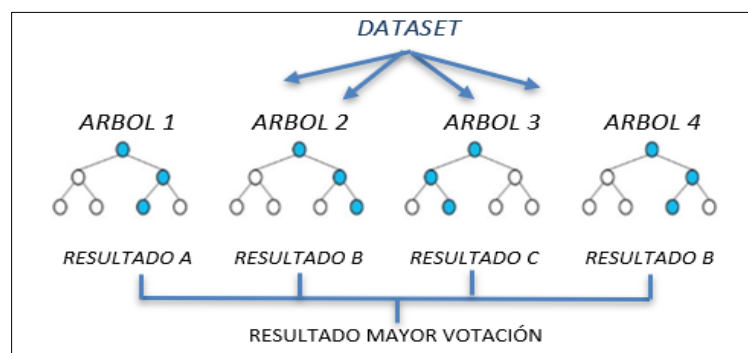
#### Desventajas

- No es utilizable en problemas no lineales.
- Actualmente hay varias opciones que predicen mejor y son más eficaces.
- Es usada únicamente para predecir una variable objetivo-categorica.
- Es un algoritmo en el que hay que cuidar el sobreajuste. (TIBCO, 2023)

#### **2.5.3. RANDOM FOREST**

Es un algoritmo de aprendizaje supervisado usado tanto en problemas de regresión como en problemas de clasificación, más comúnmente usado en este último caso, la propiedad más significativa de este algoritmo es que maneja datos con variables numéricas continuas y variables categóricas, estas últimas son el caso de clasificación.

Este algoritmo está estructurado por subconjuntos de árboles de decisión, tal como se muestra en la figura 5 cada subconjunto está entrenado con muestras *random* y diferentes, por esta razón es un algoritmo más eficiente que el árbol de decisión, ya que en este caso cada grupo aporta una predicción y la predicción final es el resultado de la combinación de las predicciones de cada árbol.



**Figura 5:** Estructura de un modelo Random Forest.

**Fuente:** Figura creada para representar la estructura de Random Forest

La implementación de un *Random Forest* la podemos resumir en el siguiente listado de pasos realizados automáticamente:

- Se elige un subconjunto de observaciones y un subconjunto de variables independientes para construir cada árbol de decisión.
- Para cada subconjunto creado en el primer paso se crea un árbol de decisión.
- Cada uno de los árboles creado en el paso dos genera una salida.
- El resultado final es determinado por la mayor votación para la clasificación. (E., 2023)

#### Consideraciones para el algoritmo Random Forest

En este algoritmo al usar múltiples árboles de decisión, existe la posibilidad de que algunos árboles predigan correctamente, mientras que otros no, sin embargo, el número total de árboles como un todo deben predecir un resultado correcto. Para una predicción eficaz este algoritmo asume los dos siguientes puntos:

- Debe existir observaciones reales en la variable dependiente para que así se prediga de manera precisa.
- Las correlaciones de las predicciones de cada uno de los árboles de decisión deben ser bajas.

A continuación, se enumeran los pasos para aplicación del algoritmo *Random Forest*:

1. Conjunto de datos preprocesado adecuadamente.
2. Algoritmo *Random Forest* ajustado al conjunto de *training*.
3. Predicción del resultado en los datos de *testing*.
4. Creación de la matriz de confusión para conocer la eficacia de predicción.
5. Proyección de los resultados en el subconjunto de *testing*. (Random Forest Algorithm, s.f.)

#### Ventajas y desventajas del algoritmo Random Forest

##### Ventajas

- Tiene un buen funcionamiento aún sin considerar el ajuste de hiper-parámetros.
- Es usado con buenos resultados tanto en algoritmos de clasificación como en algoritmos de regresión.
- Uno de los principales problemas del algoritmo árbol de decisión es el *overfitting*, en el algoritmo *Random Forest* disminuye significativamente la posibilidad de sobreajuste ya que se usa múltiples árboles.
- Al cambiar las muestras de entrada se mantiene estable ya que al tener varios árboles se mantiene el promedio de sus votaciones.

##### Desventajas

- Múltiples árboles de decisión son más costosos de construir que un solo árbol de decisión.

- Dependiendo el caso de estudio puede requerir tiempo considerable de entrenamiento.
- Este algoritmo no tiene un correcto funcionamiento con conjuntos de datos pequeños.
- Si un usuario requiere la explicación del comportamiento de un algoritmo *Random Forest* es bastante complicado interpretar los múltiples árboles creados. (Random Forest, el poder del Ensamble, 2019)

## 2.6. HERRAMIENTAS PARA VISUALIZAR LA EFICACIA DE UN ALGORITMO DE CLASIFICACIÓN

Una vez construido el modelo de aprendizaje automático, es necesario obtener una medida que indique qué tan eficaz es el modelo. Para los algoritmos de aprendizaje supervisado como sabemos son entrenados con datos de *training*, de estos datos el algoritmo aprende y por ende sabemos el resultado a obtener, basado en este concepto medir la eficiencia del algoritmo no tiene sentido, porque se pasan los mismos datos y se obtiene el resultado óptimo lo cual no conduce a nada. El objetivo de medir la eficiencia es ver la capacidad que tiene el algoritmo para una vez entrenada, prediga de manera correcta datos no vistos, a priori, si la predicción es correcta podemos decir que el algoritmo es eficaz.

### 2.6.1. MATRIZ DE CONFUSIÓN

Muestra la cantidad de aciertos o desaciertos de cada una de las posibles respuestas: “cada columna en la matriz muestra la cantidad de predicciones de cada clase, y cada fila de la matriz representa las instancias en la clase real”, el principal beneficio de una matriz de confusión es que permite conocer si el algoritmo está confundido entre las dos posibles respuestas.

Los resultados de la matriz de confusión se pueden observar en la figura 6, para entender las cuatro opciones de la matriz de confusión es necesario un ejemplo:

Necesitamos predecir si una persona tiene lupus o no:

- Verdaderos positivos: Persona que tiene lupus y el modelo lo clasificó como lupus positivo (VP).
- Verdaderos negativos: Persona que no tiene lupus y el modelo lo clasificó como lupus negativo (VN).
- Falsos positivos: Persona que no tiene lupus y el modelo lo clasificó como lupus positivo (FP)
- Falsos negativos: Persona que tiene lupus y el modelo lo clasificó como lupus negativo (FN)

		VALORES REALES	
VALORES PREDICCIÓN		Verdaderos positivos (VP)	Falsos positivos (FP)
		Falsos Negativos (FN)	Verdaderos Negativos (VN)

**Figura 6:** Matriz de confusión.

**Fuente:** Figura creada para fines explicativos de los valores resultados de la matriz de confusión.

Existe una métrica: *accuracy* (exactitud) que permite conocer el porcentaje de predicciones acertadas por el modelo, es decir el número de predicciones correctas y se obtiene juntamente con la matriz de confusión. (Sancho, 2021)

### 2.6.2. CURVA ROC

Es una visualización gráfica que permite conocer de manera fácil con qué valor de umbral se obtiene un mejor resultado, los diferentes valores de umbrales cambian dos factores del modelo: la sensibilidad y la especificidad, haciendo que con cada variación del umbral los resultados sean diferentes. La curva ROC hace posible conocer los modelos óptimos, descartando así los demás modelos. Al hablar de curva ROC es indispensable mencionar el área bajo la curva (AUC), cuanto más alta sea la curva, el rendimiento del modelo es mejor para diferenciar entre valores positivos y negativos, a continuación, se establece el rendimiento de un modelo de acuerdo con el valor de AUC:

- AUC igual a 1: El modelo tiene la capacidad de diferenciar de manera totalmente correcta entre valores positivos y negativos;
- AUC menor a 1 y mayor a 0,5: El modelo tiene una probabilidad alta de diferenciar entre valores positivos y negativos, teniendo mejor rendimiento al diferenciar los valores verdaderos positivos y verdaderos negativos.
- AUC igual a 0,5: El modelo no es capaz de distinguir entre valores positivos y negativos, los está prediciendo de manera *random*. (Lantz, Estadística y Machine Learning con R, 2019)

### 2.7. ANÁLISIS DE CORRESPONDENCIAS

Es un método estadístico de visualización que tiene como objetivo principal encontrar y presentar la relación entre múltiples variables categóricas. El análisis de correspondencia toma una tabla de datos y la transforma en comparaciones significativas que ayudan a obtener conclusiones sobre las variables categóricas. En el gráfico del análisis de correspondencia, la distancia entre dos puntos es la fuerza de la relación de las variables que estos dos puntos representan. (TIBCO CLOUD, 2023)

### **2.7.1. VENTAJAS DEL ANÁLISIS DE CORRESPONDENCIA**

- Descubre y muestra la fuerza de la relación entre las variables categóricas.
- Permite conocer la relación no solo entre dos variables sino también entre múltiples variables categóricas.
- Es una visualización que no usa valores reales, usa valores calculados en relación con otros resultados por lo que no supone, es objetivo.
- Es un método que traduce una tabla con múltiples variables categóricas y múltiples registros y los transforma en una visualización simple y fácil de comprender

### **2.7.2. DESVENTAJAS DEL ANALISIS DE CORRESPONDENCIA**

- Debido a que es un gráfico con información relativa se malinterpreta fácilmente.
- Es indispensable que en el conjunto de datos no exista datos nulos, negativos o con escala diferente.
- El análisis de correspondencia únicamente muestra una relación mas no una significación estadística.

### **3. CAPÍTULO 3: MARCO METODOLÓGICO**

#### **3.1. TIPO DE INVESTIGACIÓN**

El presente estudio es una investigación aplicada, explicativa, de campo. De acuerdo con su objetivo es aplicada ya que mediante conocimientos de ciencia de datos se va a determinar los factores que influyen en la DCI en menores de 5 años. De acuerdo con los medios usados para su desarrollo es de campo ya que la información usada proviene de la encuesta ENSANUT 2018. Finalmente, de acuerdo con el nivel de conocimientos adquiridos es explicativa ya que el objetivo es conocer las causas de la DCI en menores de 5 años en el país, a más de describir este problema.

#### **3.2. FUENTE DE INFORMACIÓN**

La información para el presente estudio es obtenida por el Instituto Nacional de Estadísticas y Censos de la última Encuesta Nacional de Salud y Nutrición realizada en el año 2018

#### **3.3. DELIMITACIÓN DE LA POBLACIÓN**

La encuesta Nacional de Salud y Nutrición ENSANUT 2018 recolectó información de manera aleatoria a nivel nacional de ecuatorianos entre 0 a 49 años, para nuestro estudio la población se limitará a niños ecuatorianos menores a 5 años.

#### **3.4. DIMENSIÓN DE LA MUESTRA**

Este estudio tiene como objetivo niños ecuatorianos menores de 5 años de los que tenemos información en la encuesta Nacional de Salud y Nutrición ENSANUT 2018, tenemos un total de 20.356 individuos, de este total para el modelado se tomó 70% para entrenamiento (14.249), y 30% para pruebas (6.107).

#### **3.5. RECOLECCIÓN DE INFORMACIÓN**

ENSANUT recolectó, información mediante cinco formularios los cuales constan de las siguientes secciones:

- Formulario 1 - Hogar: En la tabla 1 se presentan las secciones del formulario Hogar, corresponde a la información de todos los miembros del hogar, dentro de esta información está: información de vivienda, información social y económica de los miembros del hogar, consumo de servicios básicos, egresos de salud y alimentación.

**Tabla 1:** Encuesta Nacional de Salud y Nutrición: Formulario 1 - Hogar.

Sección	Descripción
Sección 1.	Información de la vivienda
Sección 2.	Información de los individuos del hogar
Sección 3.	Información de la economía del hogar
Sección 4.	Datos del uso de servicios y salud
Sección 5.	Datos de servicios y salud de mujeres del hogar en edad fértil
Sección 6.	Datos de seguridad e higiene de los alimentos
Sección 7.	Antropometría
Sección 8.	Información de etiquetado de alimentos procesados

**Fuente:** Instituto Nacional de Estadística y Censos, 2018

- Formulario 2 - MEF: La tabla 2 contiene las secciones referentes a la información de mujeres en edad fértil, salud de la niñez, lactancia materna, historial de embarazos, planificación de embarazos.

**Tabla 2:** Encuesta Nacional de Salud y Nutrición: Formulario 2 - MEF.

Sección	Descripción
Sección 1.	Información general
Sección 2.	Información histórica de embarazos y nacimientos
Sección 3.	Datos de lactancia materna
Sección 4.	Datos de la salud y la niñez
Sección 5.	Información de servicios médicos para salud materna
Sección 6.	Datos de planificación familiar
Sección 7.	Datos de selecciones reproductivas
Sección 8.	Información de salud reproductiva
Sección 9.	Información del estado civil
Sección 10.	Datos de enfermedades de transmisión sexual

**Fuente:** Instituto Nacional de Estadística y Censos, 2018

- Formulario 3 - Salud sexual y reproductiva en hombres: La tabla 3 tiene las secciones que abarcan la información de salud sexual y reproducción en hombres de 12 años en adelante, planificación en el hogar, información de enfermedades de transmisión sexual.



**Tabla 3:** Encuesta Nacional de Salud y Nutrición: Formulario 3 - Salud sexual y reproductiva en hombres.

Sección	Descripción
Sección 1.	Información general
Sección 2.	Información de actividad sexual y reproductiva
Sección 3.	Datos de la planificación familiar
Sección 4.	Información de enfermedades de transmisión sexual

**Fuente:** Instituto Nacional de Estadística y Censos, 2018

- Formulario 4 - Factores de riesgo: La tabla 4 contempla información de factores de riesgo en individuos entre 0 a 18 años.

**Tabla 4:** Encuesta Nacional de Salud y Nutrición: Formulario 4 - Salud sexual y reproductiva en hombres.

Sección	Descripción
Sección 1.	Información general del individuo
Sección 2.	Información de la salud oral
Sección 3.	Datos de la actividad física
Sección 4.	Datos de alimentación y nutrición
Sección 5.	Información del consumo de bebidas alcohólicas
Sección 6.	Información del consumo de tabaco

**Fuente:** Instituto Nacional de Estadística y Censos, 2018

- Formulario 5 - Desarrollo infantil: En la tabla 5 constan las secciones correspondientes a la información de desarrollo infantil para niños menores de 5 años. (Instituto Nacional de Estadística y Censos, 2018)

**Tabla 5:** Encuesta Nacional de Salud y Nutrición: Formulario 5 – Desarrollo infantil.

Sección	Descripción
Sección 1.	Información general del niño
Sección 2.	Información de educación de primera infancia
Sección 3.	Información de juegos
Sección 4.	Información de disciplina infantil
Sección 5.	Información de desarrollo y educación infantil
Sección 6.	
Sección 7.	Lenguaje
Sección 8.	

Sección 9.	
Sección 10.	
Sección 11.	Datos del inventario hogar
Sección 12.	Información de la motricidad
Sección 13.	Información de la madurez emocional
Sección 14.	Datos de depresión y sociabilización

**Fuente:** Instituto Nacional de Estadística y Censos, 2018

La información recolectada con estos formularios está disponible en la página oficial de ENSANUT, los datos están recolectados en nueve bases de datos abiertas y disponibles en la página oficial de INEC: <https://www.ecuadorencifras.gob.ec/salud-salud-reproductiva-y-nutricion/>

### 3.6. DETERMINACIÓN DE LA VARIABLE DEPENDIENTE E INDEPENDIENTES

Variable dependiente: En el presente estudio, la variable dependiente es *dcronica* es una variable binaria que toma dos valores 1 si tiene desnutrición crónica y 0 si no tiene desnutrición crónica.

Variables independientes: Las bases de datos a utilizar para este estudio son:

- 1\_BDD\_ENS2018\_f1\_personas (264 variables)
- 2\_BDD\_ENS2018\_f1\_hogar (120 variables)
- 4\_BDD\_ENS2018\_f2\_mef (536 variables)
- 5\_BDD\_ENS2018\_f2 lactancia (88 variables)
- 6\_BDD\_ENS2018\_f2\_salud\_ninez (350 variables)

No se consideraron el total de variables en cada base, para seleccionar las variables independientes se tomó en cuenta las causas clasificadas por la Organización del Fondo de las Naciones Unidas para la Infancia (UNICEF), la misma que clasifica estas causas en tres grupos, básicas, subyacentes e inmediatas, de acuerdo con estos tres grupos se obtuvo las variables independientes para este estudio, las mismas que se describen en la tabla 6:

**Tabla 6:** Variables independientes considerando las causas clasificadas por UNICEF.

<b>BASE ORIGEN</b>	<b>VARIABLE</b>	<b>DESCRIPCIÓN</b>	<b>TIPO</b>
1_BDD_ENS2018_f1 _personas	<u>id_viv</u>	Identificador de vivienda	Numérico
	<u>id_hogar</u>	Identificador del hogar	Numérico
	<u>id_per</u>	Identificador de la persona	Numérico

	<i>f1_s2_9</i>	Cómo se Identifica, según su cultura y costumbres	Categórica
	<i>f1_s2_14</i>	El padre vive en este hogar	Categórica
	<i>f1_s2_15</i>	La madre vive en este hogar	Categórica
	<i>f1_s4_41</i>	Atención de salud preventiva, si se hizo chequear en los últimos 30 días	Categórica
	<i>f1_s7_4_1</i>	Peso 1	Numérico
	<i>f1_s7_4_2</i>	Peso 2	Numérico
	<i>f1_s7_4_3</i>	Peso 3	Numérico
	<i>f1_s7_5_1</i>	Longitud 1	Numérico
	<i>f1_s7_5_2</i>	Longitud 2	Numérico
	<i>f1_s7_5_3</i>	Longitud 3	Numérico
	<i>f1_s7_6_1</i>	talla 1	Numérico
	<i>f1_s7_6_2</i>	talla 2	Numérico
	<i>f1_s7_6_3</i>	talla 3	Numérico
TOTAL: 16 variables			
	<u>id_viv</u>	Identificador de vivienda	Numérico
	<u>id_hogar</u>	Identificador del hogar	Numérico
	<i>f1_s1_3</i>	Material predominante del techo de la vivienda	Categórica
	<i>f1_s1_4</i>	Material predominante del piso de la vivienda	Categórica
	<i>f1_s1_5</i>	Material predominante de las paredes de la vivienda	Categórica
	<i>f1_s1_13</i>	Tipo de servicio higiénico en el hogar	Categórica
2_BDD_ENS2018_f1_hogar	<i>f1_s1_25</i>	De donde proviene principalmente, el agua que se usa para beber en este hogar	Categórica
	<i>f1_s1_27</i>	Tiempo que se demora en llegar a la fuente para obtener agua para beber	Numérico
	<i>f1_s1_28</i>	Pudo obtener las cantidades necesarias de agua para beber en las últimas 2 semanas	Categórica

	<i>f1_s6_1_6</i>	Se quedaron sin alimentos	Categórica
	TOTAL: 10 variables		
	<i>id_viv</i>	Identificador de vivienda	Numérico
	<i>id_hogar</i>	Identificador del hogar	Numérico
	<i>id_per</i>	Identificador de la madre	Numérico
	<i>id_hijo</i>	Identificador del hijo	Numérico
	<i>f2_s3a_302</i>	Al nacer su último hijo/a le dio el seno	Categórica
	<i>f2_s3a_303a</i>	Aunque no haya sido amamantado al nacer por usted, recibió leche materna de otra madre, del banco de leche, extraída o de otra forma	Categórica
	<i>f2_s3a_304</i>	A qué tiempo después del nacimiento empezó a mamar, lactar o recibir leche materna	Numérico
	<i>f2_s3c_307_2</i>	Tiempo que le dio solamente pecho, sin ningún otro líquido o complemento alimenticio (meses)	Numérico
	<i>f2_s3d_311f_2</i>	Cuántas veces consumió el día o la noche de ayer yogurt	Numérico
	<i>f2_s3d_311b_2</i>	Cuántas veces consumió el día o la noche de ayer leche de fórmula	Numérico
	<i>f2_s3d_311c_2</i>	Cuántas veces consumió el día o la noche de ayer leche en polvo	Numérico
	<i>f2_s3d_311d_2</i>	Cuántas veces consumió el día o la noche de ayer jugos naturales	Numérico
5_BDD_ENS2018_f2	<i>f2_s3d_311e_2</i>	Cuántas veces consumió el día o la noche de ayer sopa	Numérico
_lactancia	<i>f2_s3d_311g_2</i>	Cuántas veces consumió el día o la noche de ayer colada	Numérico
	<i>f2_s3d_312</i>	Comió algún alimento sólido o semisólido, durante el día o la noche de ayer	Categórica

	<i>f2_s3d_313_1</i>	Comió todo el día de ayer: colada espesa de harina de trigo o cebada, pan, arroz, fideos u otro alimento	Categórica
	<i>f2_s3d_315</i>	Cuántas veces comió el alimento sólido, semisólido o suave que no durante el día o la noche de ayer	Numérico
	<i>f2_s3_322</i>	Recibió al menos dos tomas de leche artificial, leche de vaca, u otra leche animal el día o noche de ayer	Categórica
	<i>f2_s3c_307_1</i>	Tiempo que le dio solamente pecho, sin ningún otro líquido o complemento alimenticio (años)	Numérico
	<i>f2_s3c_307_3</i>	Tiempo que le dio a solamente pecho, sin ningún otro líquido o complemento alimenticio (días)	Numérico
	<i>f2_s3d_311a_1</i>	Consumió durante el día o la noche de ayer agua pura	Categórica
	<i>f2_s3d_311a_2</i>	Cuántas veces consumió el día o la noche de ayer agua pura	Numérico
TOTAL: 22 variables			
	<i>id_viv</i>	Identificador de vivienda	Numérico
	<i>id_hogar</i>	Identificador del hogar	Numérico
	<i>id_per</i>	Identificador de la persona	Numérico
	<i>f2_s1_101</i>	Años cumplidos	Numérico
	<i>f2_s2_208_3</i>	Cuántos hijos viven actualmente con usted (total hijos/as en casa)	Numérico
	<i>f2_s2_211_3</i>	Total, hijo/as que murieron	Numérico
	<i>f2_s2_212_2</i>	Cuántos hijos /as fallecieron antes de nacer (mortinato)	Numérico
4_BDD_ENS2018_f2 _mef	<i>f2_s2a_224</i>	Número de sesiones demostrativas sobre alimentación complementaria de su niña/niño del personal de salud o de otras instituciones	Numérico

<i>f2_s5_504j_1</i>	La información que le dio el personal de salud sobre su salud o la de su bebé	Categórica
<i>f2_s5_513_1</i>	Hinchazón de manos, pies o cara	Categórica
<i>f2_s5_513_2</i>	Desmayos	Categórica
<i>f2_s5_513_4</i>	Convulsiones	Categórica
<i>f2_s5_513_8</i>	Preclamsia/Eclampsia (Producida por la presión arterial alta)	Categórica
<i>f2_s5_513_9</i>	Infección a las vías urinarias	Categórica
<i>f2_s5_516_7</i>	Infección generalizada/ sepsis	Categórica
<i>f2_s5_513_1</i>	Hinchazón de manos, pies o cara	Categórica
TOTAL: 16 variables		
<i>area</i>	Área	
<i>id_viv</i>	Identificador de vivienda	Numérico
<i>id_hogar</i>	Identificador del hogar	Numérico
<i>id_per</i>	Identificador de la madre	Numérico
<i>id_hijo</i>	Identificador del hijo	Numérico
<i>f2_s4a_402_</i>	Está vivo	Categórica
<i>f2_s4b_406_</i>	Tuvo algún control prenatal cuando estaba embarazada	Categórica
<i>f2_s4b_408_</i>	Consumió algún micronutriente durante el embarazo	Categórica
<i>f2_s4b_418_</i>	Cuántas veces le vacunaron contra el tétanos	Numérico
<i>f2_s4b_419a_</i>	Durante el control del embarazo, recibió consejería o asesoría sobre Lactancia materna	Categórica
<i>f2_s4b_419d_</i>	Recibió consejería o asesoría sobre, higiene en preparación de alimentos	Categórica
<i>f2_s4b_419b_</i>	Recibió consejería o asesoría sobre el uso de micronutrientes (hierro, ácido fólico)	Categórica

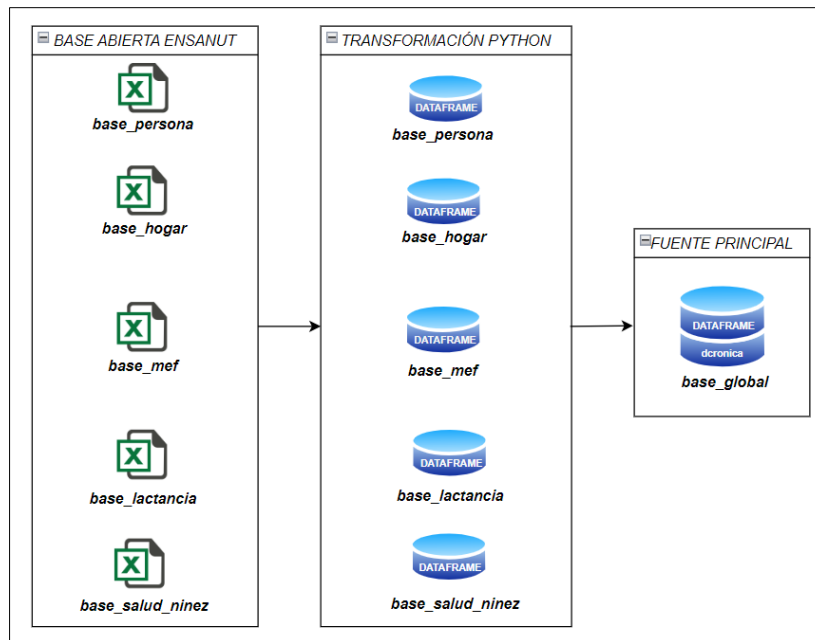
6_BDD_ENS2018_f2 _salud_ninez	<i>f2_s4e_441a_</i>	Cuánto tiempo después del parto tuvo su primer control post parto (días)	Numérico
	<i>f2_s4e_441b_</i>	Cuánto tiempo después del parto tuvo su primer control post parto (semanas)	Numérico
	<i>f2_s4e_441c_</i>	Cuánto tiempo después del parto tuvo su primer control post parto (meses)	Numérico
	<i>f2_s4d_433b_</i>	Cuántos puntos de curva de crecimiento se registraron en el carnet	Numérico
	<i>f2_s4f_449dias_</i>	Que tiempo después de nacido, le llevó al control médico por primera vez (días)	Numérico
	<i>f2_s4f_449semanas_</i>	Que tiempo después de nacido, le llevó al control médico por primera vez (semanas)	Numérico
	<i>f2_s4f_449meses_</i>	Que tiempo después de nacido, le llevó al control médico por primera vez (meses)	Numérico
	<i>f2_s4f_451num_</i>	Número de controles en el carnet del último nacido vivo	Numérico
	<i>f2_s4f_454anios_</i>	Edad que le dio el seno (leche materna) en años	Numérico
	<i>f2_s4f_454meses_</i>	Edad que le dio el seno (leche materna) en meses	Numérico
	<i>f2_s4f_454dias_</i>	Edad que le dio el seno (leche materna) en días	Numérico
	<i>f2_s4g_458_</i>	Cuántos días le duró la diarrea	Numérico
	<i>f2_s4g_461_</i>	La diarrea tenía sangre	Categórica
	<i>f2_s4h_474_</i>	Días que estuvo enfermo	Numérico
	<i>f2_s4i_482_</i>	Le dio algún desparasitante durante los últimos 6 meses	Categórica

<i>f2_s4i_483_</i>	Recibió del personal de Salud, hierro en polvo como micronutrientes (chispas) para prevenir la anemia últimos 12 meses	Categórica
<i>f2_s4j_4871a1_</i>	BCG según carnet	Categórica
<i>f2_s4j_4872a1_</i>	HEPATITIS B según carnet	Categórica
<i>f2_s4j_4875a1_1</i>	PENTAVALENTE1 según carnet	Categórica
<i>f2_s4j_4876a1_</i>	PENTAVALENTE2 según carnet	Categórica
<i>f2_s4j_4877a1_</i>	PENTAVALENTE 3 según carnet	Categórica
<i>f2_s4j_4878a1_</i>	OPV 1 según carnet	Categórica
<i>f2_s4j_4879a1_</i>	OPV 2 según carnet	Categórica
<i>f2_s4j_48710a1_</i>	OPV 3 según carnet	Categórica
<i>f2_s4j_48711a1_</i>	NEUMOCOCO1 según carnet	Categórica
<i>f2_s4j_48712a1_</i>	NEUMOCOCO2 según carnet	Categórica
<i>f2_s4j_48713a1_</i>	NEUMOCOCO3 según carnet	Categórica
<i>f2_s4j_48714a1_</i>	SRP 1 según carnet	Categórica
<i>f2_s4j_48715a1_</i>	SRP 2 según carnet	Categórica
<i>Edadmeses</i>	Edad calculada en meses	Numérico
<i>nivins_mef</i>	Nivel de instrucción de la MEF	Categórica
<u><i>dcronica</i></u>	Desnutrición crónica infantil	Numérico
<i>f2_s4b_420_</i>	Cuántas semanas de embarazo tenía cuando le hicieron el primer control	Numérico
<i>f2_s4b_421_</i>	Cuántos controles antes del parto	Numérico
TOTAL: 46 variables		

**Fuente:** Base abierta de ENSANUT, 2018.

Las fuentes de datos se visualizan en la figura 7 a continuación:





**Figura 7:** ETL de la fuente de datos ENSANUT.

**Fuente:** Gráfico realizado con drawio online por Catherine Yáñez Carrera.

### 3.7. ANÁLISIS Y ESTUDIO ESTADÍSTICO

La herramienta usada para el análisis exploratorio de datos, modelamiento y obtención de los factores que influyen en la desnutrición crónica infantil fue *Python 3.9.3* mediante el editor de código *Visual Studio Code 1.75*. Dentro de *Python* para el manejo de los datos se usaron las siguientes librerías: *pandas*, *numpy*, *math*, *matplotlib.pyplot*, *seaborn*, *sklearn.model\_selection*, *statsmodels.stats.outliers\_influence*, *sklearn.preprocessing*, *sklearn.metrics*, entre otras.

#### 3.7.1. ANÁLISIS EXPLORATORIO DE DATOS

- Se leyeron los cinco archivos.csv, *base\_personas*, *base\_hogar*, *base\_mef*, *base\_lactancia*, *base\_salud\_ninez*.
- Creamos cinco *dataframes* (uno por cada uno de los leídos en el paso anterior) únicamente con las variables útiles para este estudio: *df\_persona\_ninio*, *df\_hogar*, *df\_mef*, *df\_lactancia*, *df\_salud\_ninez*, los mismos que se indican en la tabla 6.
- Obtenemos la base principal donde está la variable dependiente: *dcronica*, sólo las observaciones donde la variable ‘edadmeses’ (*edad en meses*) sea menor o igual a 71 es decir 5 años o menos y ‘f2\_s4a\_402\_’ (individuo vivo o no) sea igual a ‘SI’, llegando a tener 20.356 observaciones
- Unificamos los cinco *dataframes* usando la función *merge* y haciendo uso de las variables únicas en cada *dataframe*: *id\_hogar*, *id\_viv*, *id\_per*, *id\_hijo*. El *dataframe* unificado tiene 45 variables.
- Se crearon ocho funciones, su código se visualiza en el Anexo 2, las mismas que facilitan el preprocesamiento de datos que realizaremos más adelante:

- Función para reemplazar los valores nulos de un *dataframe* por la media de los valores válidos.
- Función para reemplazar los valores nulos de un *dataframe* por el valor predominante,
- Función para convertir a *booleano*, una variable categórica con valores sí y no a valores 0 y 1.
- Función para eliminar *outliers*.
- Función para la obtención de variables *dummies*.
- Función para graficar barras con porcentajes considerando la variable de clasificación a *dcronica*
- Función para realizar gráficos bivariados con cantidad y porcentaje en cada barra.
- Función para exportar un *dataset*

### 3.7.2. PREPROCESAMIENTO DE DATOS

En este punto realizamos la depuración de la base unificada: tratamiento de datos faltantes, análisis de *outliers*, obtención de variables *dummies*, conversión de tipos de datos y eliminación de observaciones duplicadas.

Tratamiento de datos faltantes: Mediante la función *isnull().sum()* se obtuvo el total de datos nulos de cada variable y siguiendo las recomendaciones del marco teórico se procedió a eliminar las variables cuyo total de datos nulos supera el 50% de las observaciones, estas variables son: 'f2\_s4g\_458\_': número de días con diarrea, 'f2\_s4g\_461\_': la diarrea tenía sangre, 'f2\_s4h\_474\_': número de días que estuvo enfermo, 'f2\_s4d\_433b\_':, 'f1\_s7\_4\_3': peso, tercera toma, 'f1\_s7\_5\_1': longitud, primera toma, 'f1\_s7\_5\_2': longitud, segunda toma, 'f1\_s7\_5\_3': longitud, tercera toma, 'f1\_s7\_6\_3': talla, tercera toma, 'f1\_s1\_27': tiempo que se demora en llegar la fuente para obtener agua para beber, 'f2\_s3a\_303a': aunque no haya sido amamantado al nacer, recibió leche materna del banco de leche, 'f2\_s3a\_304': a que tiempo después del nacimiento empezó a lactar o recibir leche materna, 'f2\_s3c\_307\_2': tiempo que le dio solamente pecho, sin ningún otro líquido o suplemento, 'f2\_s3d\_311f\_2': cuántas veces consumió el día o noche de ayer yogurt, 'f2\_s3d\_311b\_2': cuántas veces consumió el día o la noche de ayer leche en fórmula, 'f2\_s3d\_311c\_2': cuántas veces consumió el día o la noche de ayer leche en polvo, 'f2\_s3d\_311d\_2': cuántas veces consumió el día o la noche de ayer jugos naturales, 'f2\_s3d\_311e\_2': cuántas veces consumió el día o la noche de ayer sopa, 'f2\_s3d\_311g\_2': cuántas veces consumió el día o la noche de ayer colada, 'f2\_s3d\_313\_1': comió todo el día de ayer colada espesa de harina de trigo o cebada, pan, arroz, fideos u otro alimento, 'f2\_s3d\_315': cuántas veces consumió el alimento sólido, semisólido o suave durante el día o la noche de ayer, 'f2\_s3c\_307\_1': tiempo que le dio solamente pecho sin ningún suplemento en meses,

*'f2\_s3c\_307\_3': tiempo que le dio solamente pecho sin ningún suplemento en días, 'f2\_s3d\_311a\_1': consumió durante el día o la noche de ayer agua pura, 'f2\_s2\_212\_2': cuántos hijos murieron antes de nacer, 'f2\_s2a\_224': número de sesiones demostrativas sobre alimentación complementaria de su niño del personal de salud o de otras instituciones, 'f2\_s5\_504j\_1': el personal de salud le dio información sobre su salud o la de su bebé, 'f2\_s2\_211\_3': total de hijos que murieron.*

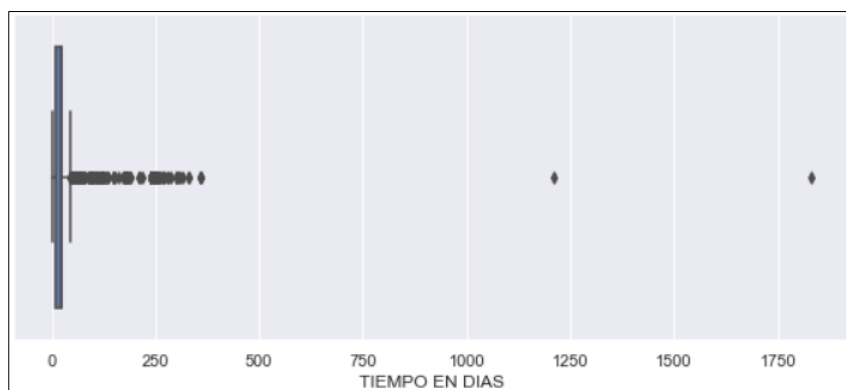
En el caso de las demás variables con un porcentaje menor al 50% en valores nulos se procede a usar métodos menos invasivos, en las variables numéricas se reemplaza por la media de los valores válidos (usando una función) y en las variables categóricas se reemplazan por el valor más frecuente, este valor se obtiene mediante *values\_count*, para este reemplazo se usó una de las funciones creadas inicialmente, en algunas variables categóricas existe el valor “no sabe / no responde”, en este caso se reemplazó por el valor más predominante de la variable, esto en las siguientes variables: *'f2\_s5\_513\_1': hinchazón de manos pies o cara de la madre, 'f2\_s5\_513\_2': la madre tuvo desmayos, 'f2\_s5\_513\_4': la madre tuvo convulsiones, 'f2\_s5\_513\_8': la madre tuvo preclamsia, 'f2\_s5\_513\_9': la madre tuvo infecciones de vías urinarias, 'f2\_s5\_516\_7': la madre tuvo sepsia, 'f1\_s1\_28': pudo obtener las cantidades necesarias de agua para beber en las últimas dos semanas.*

*Análisis y creación de variables Dummies:* La mayoría de los modelos de *machine learning* usan variables numéricas, por lo que el tratamiento de variables categóricas es muy importante en cualquier conjunto de datos, antes del modelado. Las variables *dummies* o indicadores son aplicables cuando la variable categórica tiene dos o más categorías o niveles diferentes, se crea una variable por cada categoría, donde se asigna 1 en la variable que cumple la condición y 0 en el resto de variables, esto se aplicó en los siguientes casos: *'f2\_s4b\_406\_': tuvo algún control prenatal cuando estaba embarazada, 'f2\_s4b\_419a\_': durante el control del embarazo recibió consejería o asesoría sobre lactancia materna, 'f2\_s4b\_419b\_': recibió consejería o asesoría sobre el uso de micronutrientes, 'f2\_s4b\_419d\_': recibió consejería o asesoría sobre la higiene en preparación de alimentos, 'f2\_s4i\_482\_': le dio algún desparasitante durante los últimos 6 meses, 'f2\_s4i\_483\_', 'f2\_s4j\_4871a1\_': recibió del personal de salud, hierro en polvo como micronutrientes para prevenir anemias durante los últimos 12 meses, 'f2\_s4j\_4872a1\_': tiene dosis de hepatitis B según carnet, 'f2\_s4j\_4875a1\_': tiene primera dosis de pentavalente según carnet, 'f2\_s4j\_4876a1\_': tiene segunda dosis de pentavalente según carnet, 'f2\_s4j\_4877a1\_': tiene tercera dosis de pentavalente, 'f2\_s4j\_4878a1\_': tiene primera dosis de OPV según carnet, 'f2\_s4j\_4879a1\_': tiene segunda dosis de OPV según carnet, 'f2\_s4j\_48710a1\_': tiene tercera dosis de OPV según carnet, 'f2\_s4j\_48711a1\_': tiene primera dosis de neumococo según carnet, 'f2\_s4j\_48712a1\_': tiene segunda dosis de neumococo según carnet, 'f2\_s4j\_48713a1\_': tiene tercera dosis de neumococo según carnet, 'f2\_s4j\_48714a1\_': tiene primera dosis contra sarampión según carnet, 'f2\_s4j\_48715a1\_': tiene*

segunda dosis contra sarampión según carnet, 'f1\_s2\_14': el padre vive en este hogar, 'f1\_s2\_15': la madre vive en este hogar, 'f1\_s4\_41': se hizo chequear en los últimos 30 días, atención preventiva, 'f1\_s1\_28': pudo obtener las cantidades necesarias de agua para beber en las últimas dos semanas, 'f1\_s6\_1\_6': se quedaron sin alimento en los últimos 6 meses, 'f2\_s5\_513\_1': hinchazón de manos, pies o cara, 'f2\_s5\_513\_2': desmayos de la madre, 'f2\_s5\_513\_4': convulsiones de la madre, 'f2\_s5\_513\_8': preclamsia de la madre, 'f2\_s5\_513\_9': infección de vías urinarias de la madre, 'f2\_s5\_516\_7': sepsis de la madre, 'f2\_s3a\_302': al nacer su último hijo, le dio el seno, 'f2\_s3d\_312': comió algún alimento sólido o semisólido, durante el día o noche de ayer, 'f2\_s3\_322': recibió al menos dos tomas de leche artificial, leche de vaca u otra leche animal el día o noche de ayer , en total son 146 variables, incluidas las variables *dummies*.

Análisis de outliers: Siendo un caso de estudio para determinar factores que influyen en la DCI, es necesario tener cuidado en la detección y posible eliminación de estos valores atípicos, para este análisis se tomó las variables numéricas y se usó el diagrama de caja para una mejor visualización.

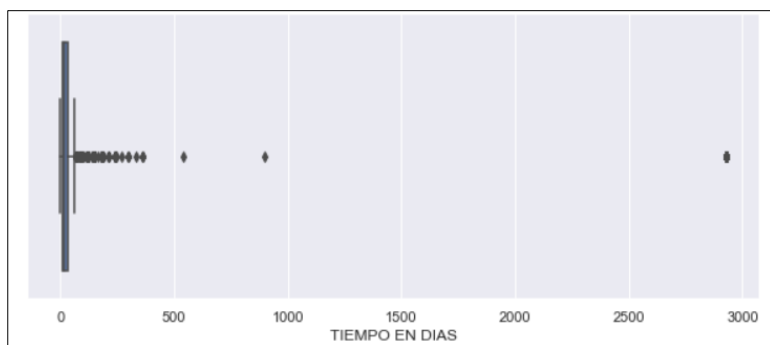
Unificamos las variables: 'f2\_s4e\_441a\_', 'f2\_s4e\_441b\_', 'f2\_s4e\_441c\_' en una sola: *tiempo\_control\_posparto* ya que las tres hacen referencia a la misma característica, pero en diferentes unidades de tiempo (meses, días y años) la nueva variable está en días, en la figura 8 podemos ver el diagrama de caja de la variable, el mismo que indica los *outliers*.



**Figura 8:** Diagrama de caja de la variable *tiempo\_control\_posparto*.

**Fuente:** Gráfico realizado con Python en *Visual Studio Code*.

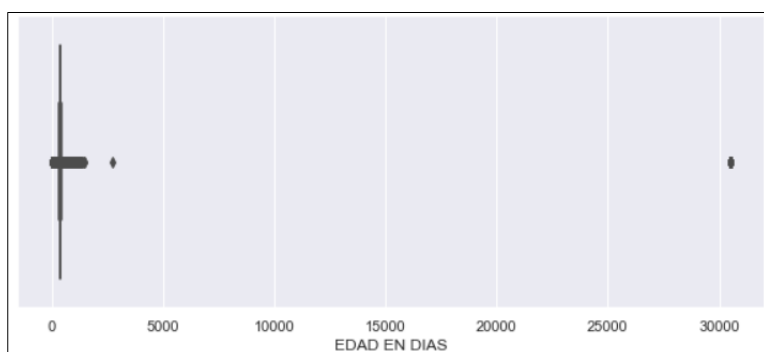
De igual manera tratamos las variables: 'f2\_s4f\_449dias\_', 'f2\_s4f\_449semanas', 'f2\_s4f\_449meses\_' las unificamos en la variable *primer\_control\_medico* en días, en la figura 9 consta el diagrama de caja, esta variable hace referencia al tiempo que se demoró una mujer en hacerse el primer control durante el embarazo.



**Figura 9:** Diagrama de caja de la variable `primer_control_medico`.

**Fuente:** Gráfico realizado con Python en *Visual Studio Code*.

La variable `edad_fin_lactancia` es la suma de las variables `'f2_sf_454dias_'`, `'f2_sf_454meses'` y `'f2_sf_454años_'`, previamente convertidas a días y representa la edad del menor a la que terminó su periodo de lactancia.

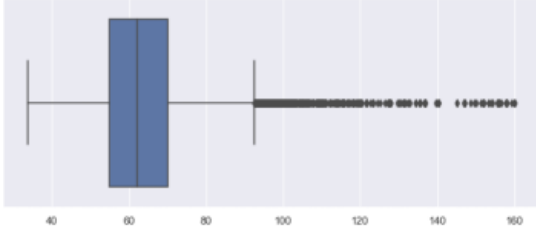
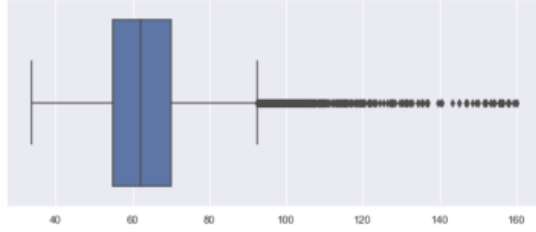
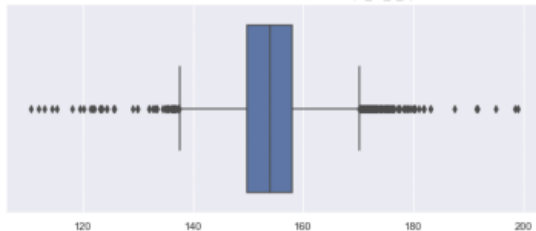
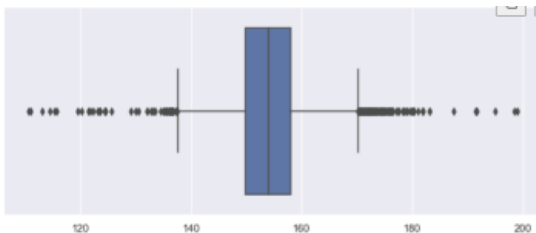
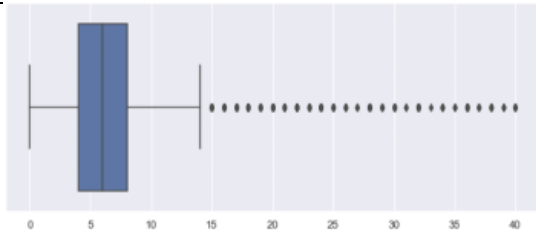



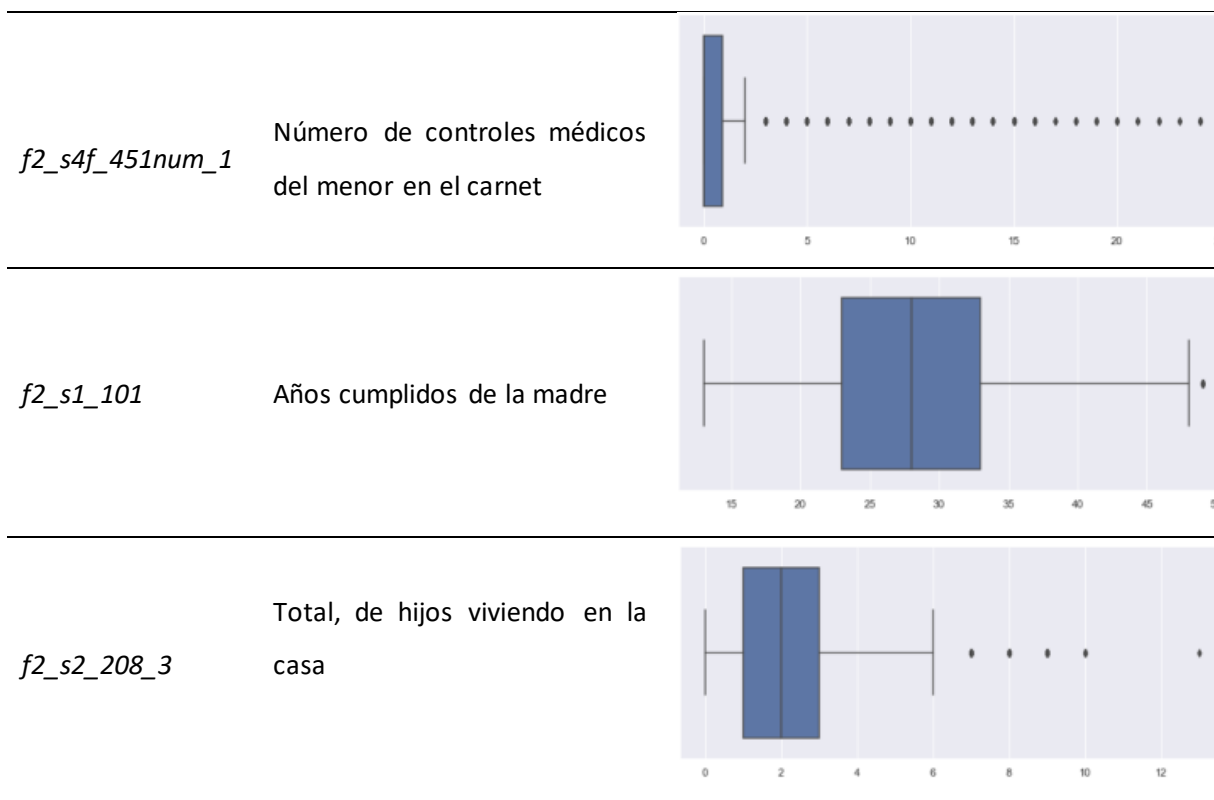
**Figura 10:** Diagrama de caja de la variable `edad_fin_lactancia`.

**Fuente:** Gráfico realizado con Python en *Visual Studio Code*.

Las dos variables: `tiempo_control_posparto`, `primer_control_medico` representan los días que una mujer tardó en hacerse el primer control después del parto y días que una mujer tardó en hacerse el primer control médico durante el embarazo respectivamente, tenemos varios valores atípicos, sin embargo, en este caso no se eliminan ya que son variables que no tienen un rango estándar, dependen de diferentes factores: geográficos, económicos, sociales, etc. La variable `edad_fin_lactancia` de igual forma tiene valores atípicos que en este caso se eliminan (usando una función) ya que existen observaciones que en esta variable tienen valores altos y anormales, por ejemplo, no existen niños de 13 años o más que aún estén en etapa de lactancia. Las demás variables se visualizan en la tabla 7, de las cuales los *outliers* no se eliminaron, debido a que pueden ser el punto de interés para el presente estudio.

**Tabla 7:** Diagrama de caja de las variables independientes.

Variable	Descripción	Outliers (Diagrama de caja)
<i>f1_s7_4_1</i>	Toma uno del peso del menor en kilogramos.	
<i>f1_s7_4_2</i>	Toma dos del peso del menor en kilogramos.	
<i>f1_s7_6_1</i>	Toma uno del tamaño del menor en centímetros	
<i>f1_s7_6_2</i>	Toma dos del tamaño del menor en centímetros	
<i>f2_s4b_420_</i>	Número de semanas de embarazo al realizarse el primer control médico.	
<i>f2_s4b_421_</i>	Número de controles que se hizo la mujer durante el embarazo	



**Fuente:** Gráficos obtenidos con la herramienta Python en Visual Studio Code.

### 3.7.3. ANÁLISIS DE LA MULTICOLINEALIDAD

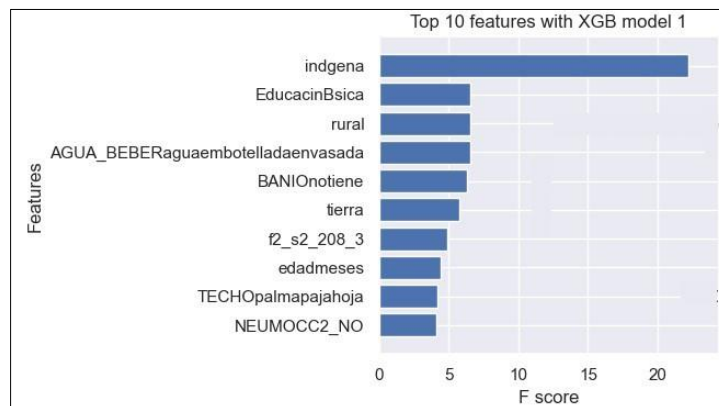
Una de las condiciones para realizar un buen modelo en regresión logística es que las variables que tomamos como independientes no tengan multicolinealidad, es decir que no tengan una alta colinealidad entre sí.

La multicolinealidad se puede detectar usando el factor de inflación de varianza (VIF), para ello obtendremos las variables cuantitativas, usamos la función *variance\_inflation\_factor* de la librería *statsmodels.stats.outliers\_influence*, en este caso las variables con alta multicolinealidad son *Educación Básica*, *Educación Media/Bachillerato* y *Superior*, las tres variables corresponden a la educación de la madre, de estas eliminamos *Superior* y observamos claramente que VIF baja en las otras dos variables, el mismo caso es de las variables '*f1\_s7\_4\_2*': *peso del menor, segunda toma* y '*f1\_s7\_6\_1*': *talla del menor, primera toma*, procedemos a eliminar la variable '*f1\_s7\_6\_1*' y observamos que el valor de VIF baja significativamente, entendiéndose que ya no existe multicolinealidad entre las variables independientes.

### 3.7.4. SELECCIÓN DE VARIABLES INDEPENDIENTES

Una vez analizado el VIF es necesario seleccionar las variables para que el modelo tenga el mejor rendimiento posible, *Python* tiene algunas bibliotecas que permiten seleccionar las variables óptimas

y conseguir un buen modelo, en este caso usaremos la librería: *featurewiz*, en la figura 11 y tabla 8, a continuación, se visualizan las variables obtenidas mediante esta librería:



**Figura 11:** Gráfico del top 10 de variables que hacen un modelo óptimo.

**Fuente:** Gráfico realizado con Python en *Visual Studio Code*.

**Tabla 8:** Variables seleccionadas mediante la biblioteca *featurewiz*

Variable	Descripción
'indígena'	Etnia del menor
'Educación Básica'	Nivel de instrucción de la madre
'rural'	Área donde se ubica la vivienda
'AGUA_BEBERagua embotellada /envasada'	Fuente del agua que bebe el menor
'BANIONo tiene'	La vivienda del menor no tiene baño
'tierra?'	El piso del menor es de tierra
'f2_s2_208_3'	Total, de hijos que habitan la vivienda
'edadmeses'	Edad en meses del menor
'TECHOpalma/paja/hoja'	El material del techo de la vivienda es paja o palma
'NEUMOCC2_NO'	El menor no tiene la segunda dosis de la vacuna de NEUMOCOCO.

**Fuente:** Variables obtenidas con Python en *Visual Studio Code*

### 3.7.5. DESARROLLO Y EVALUACIÓN DE LOS MODELOS DE CLASIFICACIÓN

Para este estudio se eligieron dos modelos de clasificación: Regresión logística y *Random Forest*, antes de empezar con el modelado se dividió a los datos en dos subconjuntos, el 70% para entrenamiento y el 30% para pruebas, se usó la librería: *train\_test\_split*.

Una vez seleccionadas las variables se cambió a una misma escala de valores, haciendo uso de la librería *StandardScaler*, el proceso de escalabilidad se realiza eliminando la media y escalando a la varianza de la unidad.



Para la implementación del algoritmo Regresión Logística se consideró la librería: *linea\_model.LogisticRegression*, con esta función lo primero que se realiza es la creación de un objeto *LogisticRegression* con la propiedad *random\_state* para que la selección de las observaciones sea de manera *random*, a continuación se hace un ajuste del modelo usando *fit()*, todo esto se realiza con los datos de entrenamiento, además se usó la función *predict()* con el subconjunto de prueba para la predicción.

Se importó la librería *classification\_report* la misma que permite obtener un reporte del algoritmo de regresión logística: *precision*, *recall*, *f1-score*, estas variables permiten tener noción del rendimiento del modelo.

En el caso del algoritmo *Random Forest* se usó la librería *RandomForestClassifier*, y para ajustar el modelo se usó los principales hiper-parámetros, en este caso son los siguientes:

- *n-estimators*: Es la cantidad de árboles de decisión en el modelo, mientras más elevado sea este número, el rendimiento del modelo será mejor a cambio de un aumento en el costo computacional de su implementación, para este estudio se definió un valor de 45 para este hiper-parámetro
- *random-state*: Este hiper-parámetro reproducirá los mismos resultados cada vez que se ejecute, el valor por *default* es *none* pero en este caso queremos mantener los mismos resultados, en cada ejecución.
- *bootstrap*: Hace referencia a los diferentes tamaños de las muestras para *training* para utilizar en la fase de *training*. En este caso usamos *False* ya que usaremos el conjunto completo de observaciones de *training*.
- *Max-features*: Es la forma de elegir el número máximo de características para cada árbol de decisión del modelo, en este caso se define: *sqrt* (Cardellino, 2021)

Mediante la característica *feature\_importances\_* se obtuvieron la importancia y significancia de cada variable independiente en el modelo. Tanto para la regresión logística como para *random forest* se usa la matriz de confusión para el análisis del rendimiento de modelo, para el algoritmo de regresión logística también se usó la función *roc\_curve* la misma que permite obtener el valor del área bajo la curva ROC gráfica y numéricamente. (Pramoditha, 2022)

El código consta en el archivo Jupyter adjunto al documento Anexo 1, llamada: *Anexo1, Análisis Desnutrición Crónica infantil basado en ENSANUT 2018*.

## 4. CAPÍTULO 4. ANÁLISIS DE RESULTADOS

### 4.1. ANÁLISIS EXPLORATORIO DE DATOS

El presente estudio se realizó sobre un total de 20.356 individuos que es el total de niños ecuatorianos menores de 5 años vivos, de los cuales 28,73% presentan desnutrición crónica infantil, además 153 niños constan como fallecidos de los cuales el 99,35%, presentaban DCI, las cantidades y porcentajes se pueden visualizar en la tabla 9.

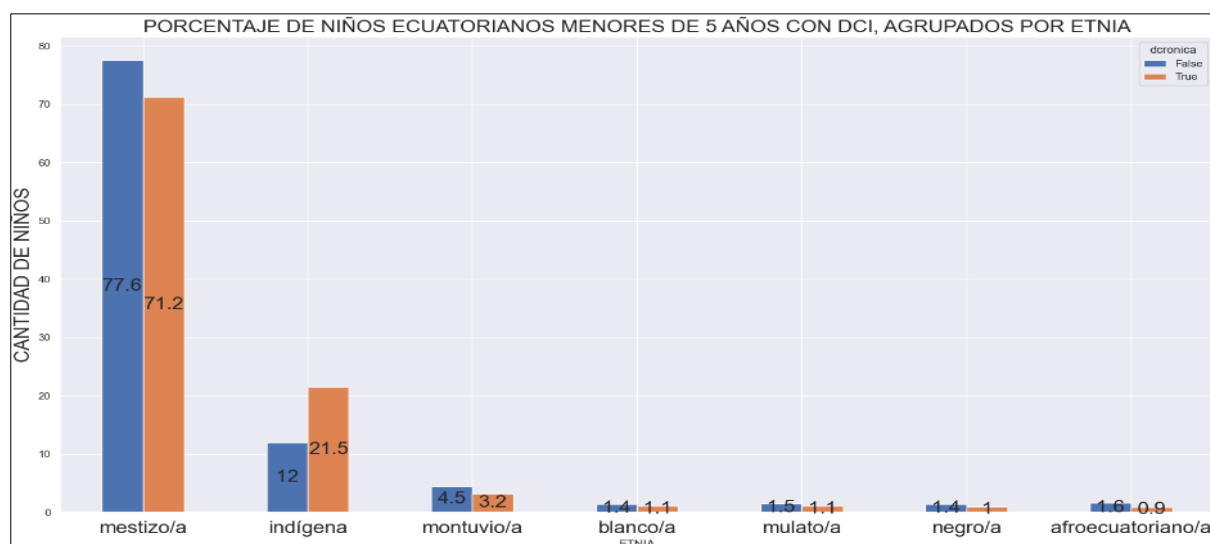
**Tabla 9:** Porcentaje y cantidad de menores de 5 años vivos y fallecidos con y sin desnutrición crónica infantil.

VIVO	DCI	CANTIDAD	PORCENTAJE
NO	TRUE	152	99,35%
	FALSE	1	0,65%
SI	FALSE	14.507	71,27%
	TRUE	5.849	28,73%

**Fuente:** Tabla obtenida en Python con Visual Studio Code.

A continuación, se muestra un análisis de algunas de las variables a intervenir en el estudio con respecto a la variable dependiente: dcronica.

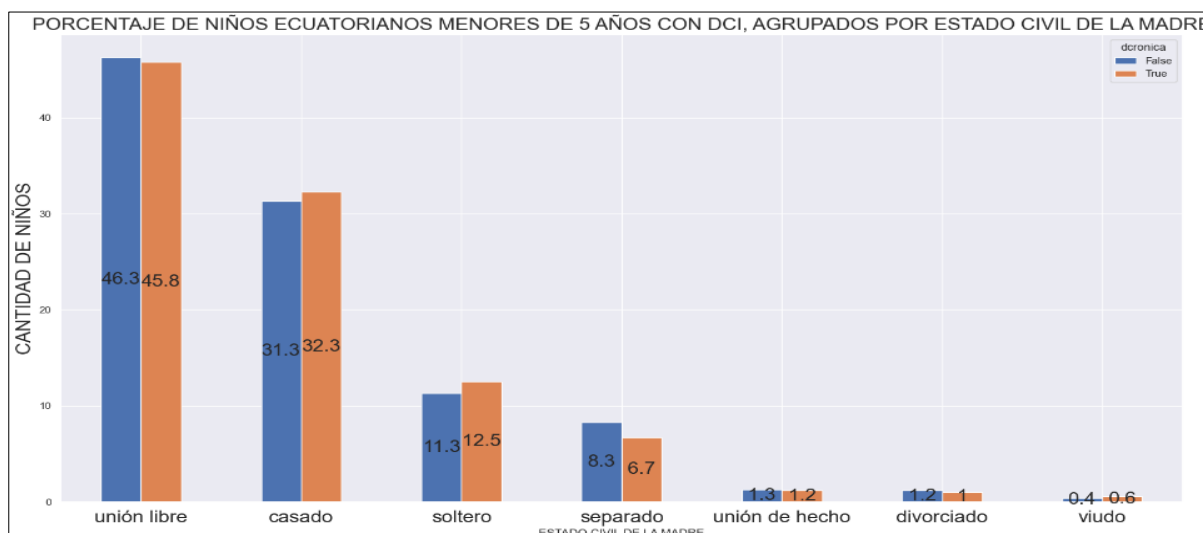
*f1\_s2\_9:* Analizando la distribución de los niños sin DCI vemos que las categorías con más concentración de casos son mestizos/a con 77,5% e indígena con un 12%. Para los niños con DCI la tendencia es similar, sin embargo, la concentración de casos baja a 71,2% para mestizo/a y sube a 21,5% para el caso de indígena. A continuación, podemos visualizar como está distribuida la variable con respecto a la variable dependiente:



**Figura 12:** Porcentaje de DCI por Etnia.

**Fuente:** Gráfico obtenido usando Python en Visual Studio Code, a partir de la base abierta de ENSANUT

*f2\_s9\_900*: Considerando la distribución del total de niños ecuatorianos menores de 5 años sin DCI de acuerdo al estado civil de los padres, se puede notar que los casos donde se centraliza es donde los padres están en unión libre con 46,3%, padres casados con 31,3% y padre o madre soltero/a con 11,3%. En el caso de los menores ecuatorianos con DCI, la inclinación es parecida, pero en el caso donde los padres están en unión libre el porcentaje baja a 45,8%, en el caso de los padres casados sube a 32,3% de igual forma el caso donde padre o madre están soltero/a sube a 12,5%, en la figura 13, podemos visualizar como está distribuida la variable con respecto a la variable dependiente:

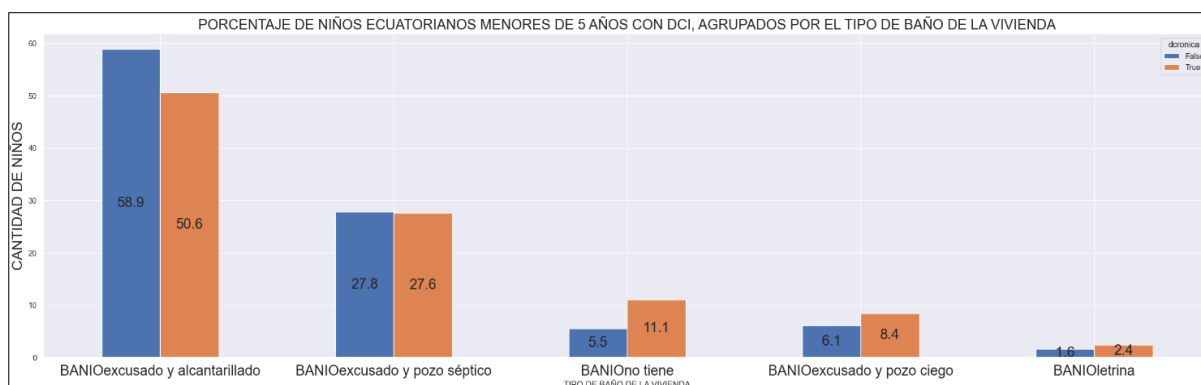


**Figura 13:** Porcentaje de DCI por el estado civil de los padres.

**Fuente:** Gráfico obtenido usando Python en Visual Studio Code, a partir de la base abierta de ENSANUT.

*f1\_s1\_25*: Del total de individuos ecuatorianos menores de 5 años, en el caso donde la fuente de agua para beber es río o acequia y es agua lluvia el porcentaje de niños con DCI es casi igual al total de niños sin DCI, la figura de esta variable con respecto a la variable dependiente está en el Anexo 1.

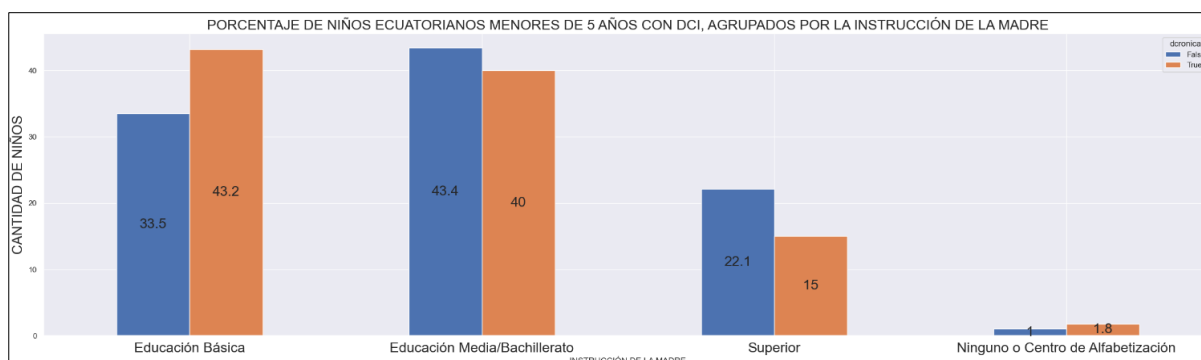
*f1\_s1\_13*: Al visualizar la distribución de los niños menores de 5 años sin DCI, se nota claramente que el porcentaje más alto está en niños que habitan viviendas con excusado y alcantarillado con un porcentaje de 58,9% y menores que habitan en viviendas con excusado y pozo séptico con 27,8%, en el caso de los niños menores de 5 años con DCI la curva de tendencia es parecida sin embargo el porcentaje que habitan en viviendas con baño excusado y alcantarillado baja a 50,6%, los menores que habitan en viviendas con excusado y pozo séptico tienen prácticamente el mismo porcentaje 27,6%, en el caso de los menores que habitan viviendas que no tienen baño tienen un porcentaje de 11,1%, casi el doble del porcentaje de niños sin DCI que es el 5,5%, en la figura 14 se visualiza esta distribución:



**Figura 14:** Porcentaje de DCI por el baño en la vivienda.

**Fuente:** Gráfico obtenido usando Python en Visual Studio Code, a partir de la base abierta de ENSANUT.

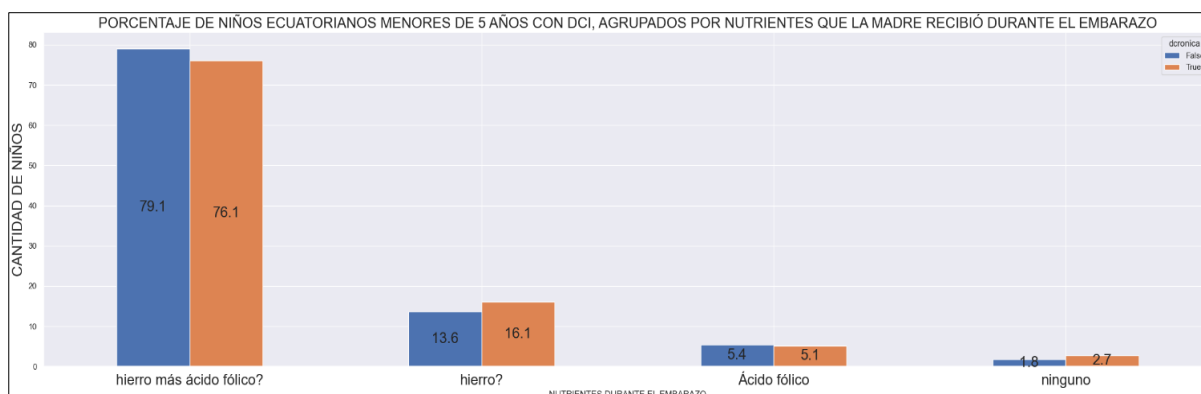
*nivins\_mef:* Analizando la distribución de menores de 5 años sin DCI, se observa que donde se concentran la mayor cantidad de casos es donde la madre tiene Educación Básica como nivel de instrucción con un porcentaje de 33,5% y en los casos donde la madre tiene como nivel de instrucción Educación Media/Bachillerato con un porcentaje de 43,4%, en el caso de la distribución de menores con DCI la tendencia es parecida, donde la madre tiene Educación Básica como nivel de instrucción el porcentaje sube a 43,2% y en el caso donde la madre tenga como nivel de instrucción Media/Bachillerato baja a 40%, en la figura 15 podemos visualizar cómo está distribuida la variable con respecto a la variable dependiente.



**Figura 15:** Porcentaje de DCI por la instrucción de la madre.

**Fuente:** Gráfico obtenido usando Python en Visual Studio Code, a partir de la base abierta de ENSANUT.

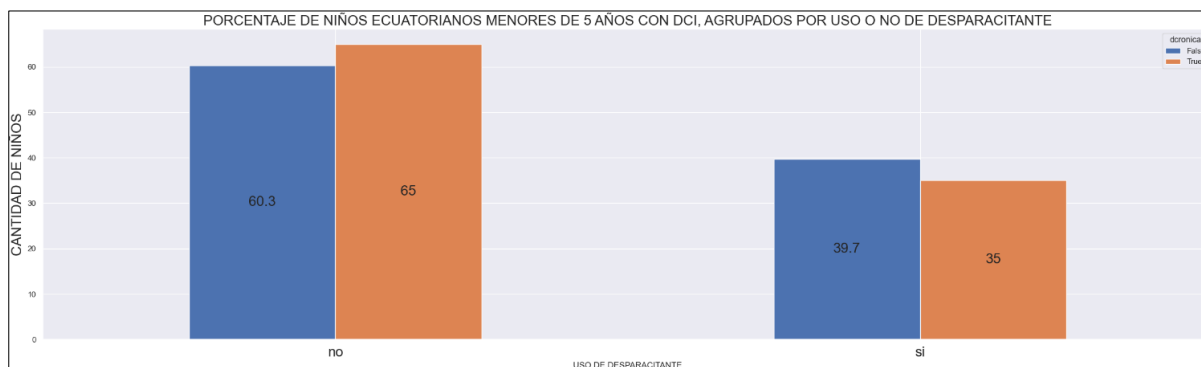
*f2\_s4b\_408:* En la distribución de menores sin DCI podemos deducir que la mayor cantidad de casos es en madres que recibieron ácido y hierro durante el embarazo con el 79,1% y en el caso de madres que recibieron únicamente hierro durante el embarazo con un porcentaje de 13,6%, en el caso de la distribución de menores con DCI, la inclinación es parecida, sin embargo el porcentaje de casos donde la madre recibió ácido y hierro como micronutrientes baja a 76,1% y donde la madre recibió únicamente hierro sube a 16,1%, en la figura 16, podemos visualizar como está distribuida la variable con respecto a la variable dependiente.



**Figura 16:** Porcentajes de DCI por nutrientes recibidos durante el embarazo

**Fuente:** Gráfico obtenido usando Python en Visual Studio Code, a partir de la base abierta de ENSANUT.

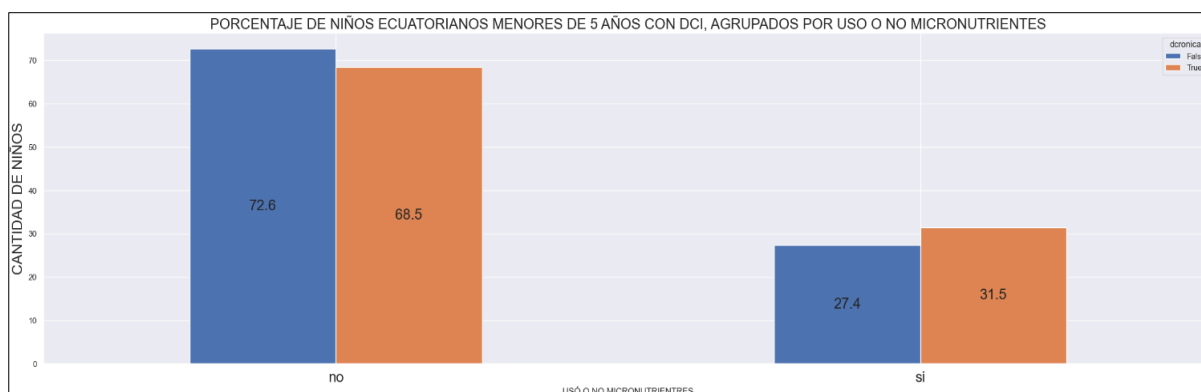
f2\_s4i\_482\_: De la distribución de niños sin DCI, el porcentaje donde se concentran los casos es en menores que en los últimos 6 meses no se desparasitaron con un 60,3%, la diferencia pertenece a niños que se desparasitaron en los últimos 6 meses es decir el 39,7%, la tendencia de menores con DCI es parecida sin embargo sube en los menores que no se desparasitaron en los últimos 6 meses con un porcentaje del 65% y la diferencia pertenece a los niños que se desparasitaron en los últimos 6 meses es decir el 35%, en la figura 17 podemos visualizar como está distribuida la variable con respecto a la variable dependiente.



**Figura 17:** Porcentaje de DCI si el menor recibió o no desparasitante.

**Fuente:** Gráfico obtenido usando Python en Visual Studio Code, a partir de la base abierta de ENSANUT.

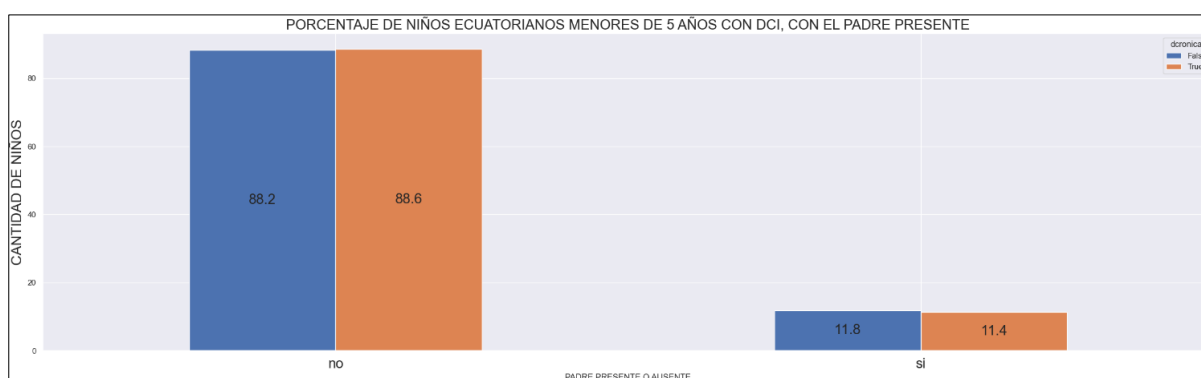
f2\_s4i\_483\_: Según la distribución de los niños sin DCI, se centraliza la mayor cantidad de casos en menores que no recibieron micronutrientes por parte del personal médico de la salud con un porcentaje de 72,6%, la diferencia es de niños que recibieron micronutrientes del personal médico con un porcentaje de 27,4%, la tendencia es similar en niños con DCI pero el porcentaje baja en menores que no recibieron micronutrientes por el personal médico a 68,5%, y en el caso donde el menor si recibió micronutrientes por parte del personal aumenta a 31,5%, en la figura 18 podemos visualizar como está distribuida la variable con respecto a la variable dependiente.



**Figura 18:** Porcentaje de DCI si el menor recibió o no micronutrientes.

**Fuente:** Gráfico obtenido usando Python en *Visual Studio Code*, a partir de la base abierta de ENSANUT.

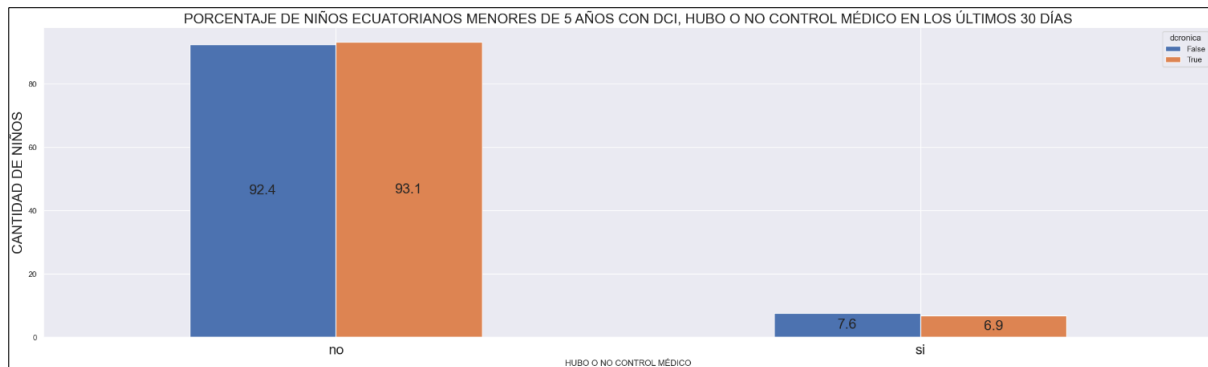
f1\_s2\_14: De la distribución de niños sin DCI, la mayor cantidad se concentra en niños que no tienen un padre presente con un porcentaje de 88,2% y la diferencia en casos de niños con el padre presente es decir 11,8%, la tendencia es la misma en niños con DCI, el porcentaje aumenta en niños donde el padre no está presente con un porcentaje de 88,6% y ligeramente menos en niños donde el padre está presente con un porcentaje de 11,4%, en la figura 19 podemos visualizar como está distribuida la variable con respecto a la variable dependiente.



**Figura 19:** Porcentaje de DCI de menores donde el padre está presente.

**Fuente:** Gráfico obtenido usando Python en *Visual Studio Code*, a partir de la base abierta de ENSANUT.

f1\_s4\_41: De la distribución de niños sin DCI, la mayor cantidad se concentra en niños que no tienen madre presente con un porcentaje de 92,4% y la diferencia en casos de niños con la madre presente es decir 7,6%, la tendencia es la misma en niños con DCI, el porcentaje aumenta en niños donde la madre no está presente con un porcentaje de 93,1% y ligeramente menos en niños donde la madre está presente con un porcentaje de 6,9%, como se visualiza en la figura 20.

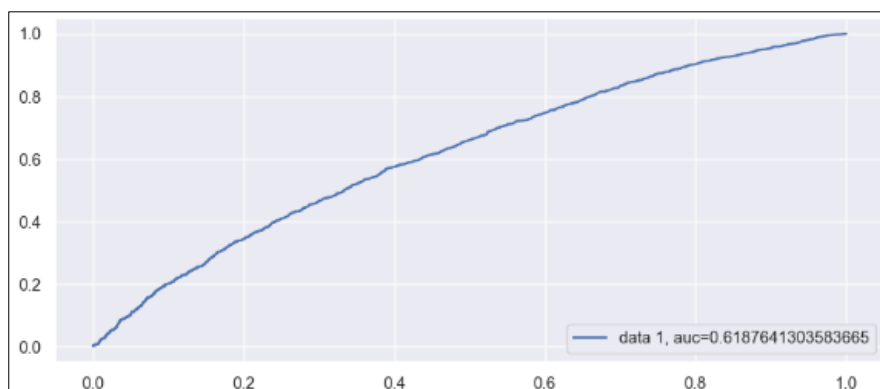


**Figura 20:** Porcentaje de DCI en menores donde hubo o no hubo control médico en los últimos 30 días.

**Fuente:** Gráfico obtenido usando Python en *Visual Studio Code*, a partir de la base abierta de ENSANUT.

## 4.2. MODELO DE REGRESIÓN LOGÍSTICA

Una vez implementado el modelo de Regresión Logística ( $P(Y) = \frac{1}{1+e^{-(b_0+b_1X_1+b_2X_2+\dots+b_nX_n)}}$ ) con el conjunto de datos de *testing*, para medir su rendimiento se obtuvo el área bajo la curva figura 21, la misma que da un valor de 0,618764 lo que indica que el modelo tiene una probabilidad alta de diferenciar entre valores positivos y negativos



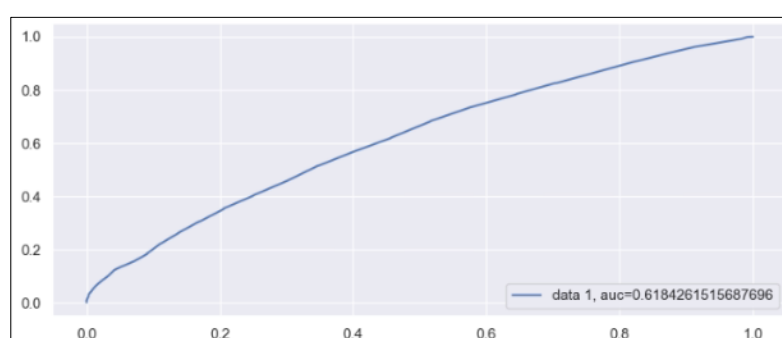
**Figura 21:** Curva ROC para el modelo de Regresión Logística.

**Fuente:** Gráfico obtenido usando Python en *Visual Studio Code*, a partir de la base abierta de ENSANUT.

Conjuntamente, se obtuvo la matriz de confusión, que consta en el Anexo 3. la misma que permitió conocer métricas importantes con respecto al modelo, se determinó que el modelo realiza predicciones acertadas en un 71,16%, del total de predicciones, el 98% del total de casos de niños que no tienen DCI fueron predichos correctamente por el modelo es decir se predijeron correctamente 4.261 de los 4.361 que no tienen DCI, mientras que el 5% del total de niños con DCI se predijo correctamente por el modelo es decir 84 de los 1.745 que tienen DCI se predijeron correctamente.

### 4.3 MODELO RANDOM FOREST

El modelo de *Random Forest* fue evaluado mediante la curva AUC figura 22 donde se obtuvo un valor de 0,618426 nos indica que el modelo tiene una probabilidad alta de diferenciar entre valores positivos y negativos, también se obtuvo la matriz de confusión: Anexo 3, la misma que da como resultado un 68,20% de predicciones acertadas del total de predicciones, el 86,61% del total de casos de niños que no tienen DCI fueron predichos correctamente por el modelo es decir en 3.777 de los 4.361 que no tienen DCI se predijo correctamente mientras que el 22,18% del total de niños con DCI se predijo correctamente por el modelo es decir 387 de los 1.745 que tienen DCI se predijeron correctamente.



**Figura 22:** Curva ROC para el modelo de Random Forest.

**Fuente:** Gráfico obtenido usando Python en Visual Studio Code, a partir de la base abierta de ENSANUT.

### 4.4. COEFICIENTE GINI DE LAS VARIABLES QUE INTERVIENEN EN LOS MODELOS

En la tabla 10, constan las diez variables independientes con los *betas* obtenidas en el modelo de Regresión Logística, ordenados ascendentemente, este coeficiente sirve para determinar la significancia de cada variable en el modelo.

**Tabla 10:** Variables con los 10 coeficientes de significancia más altos, del modelo: Regresión Logística.

VARIABLE	DESCRIPCIÓN	COEFICIENTE
<u>tierra?</u>	Material del piso donde vive el menor	0,481422
<u>indígena</u>	Etnia del menor: indígena	0,390932
<u>BANIO no tiene</u>	Baño de la vivienda del menor: no tiene	0,228765
<u>Educación Básica</u>	Nivel de instrucción de la madre	0,227992
<u>AGUA BEBERagua embotellada</u> <u>/envasada</u>	El agua que bebe el menor es agua embotellada	0,181703
<u>rural</u>	El área donde vive el menor	0,120226
<u>TECHO palma/paja/hoja</u>	Material del techo de la vivienda	0,069338



<u>f2_s2_208_3</u>	Cuántos hijos viven actualmente con usted (total hijos/as en casa)	0,066672
<u>NEUMOCC2_NO</u>	El menor no tiene la segunda dosis contra el neumococo	0,064469
<u>edadmeses</u>	Edad en meses del menor	0,009803

**Fuente:** Tabla obtenida en Python con Visual Studio Code.

En el caso del modelo Random Forest se utiliza el coeficiente *gini* para medir la significancia de las variables, estos coeficientes se muestran en la tabla 11, las variables están ordenadas ascendentemente.

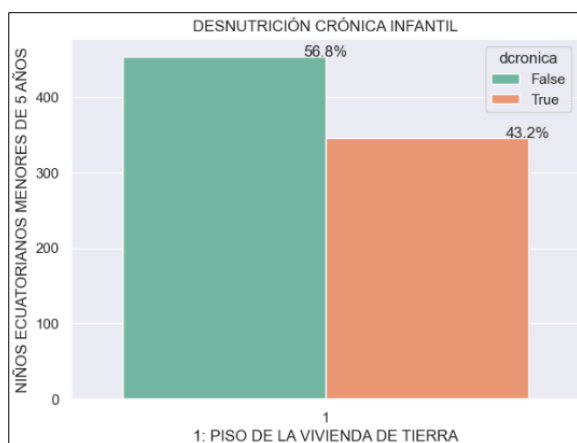
**Tabla 11:** Variables con los 10 coeficientes de significancia más altos, del modelo: Random Forest.

VARIABLE	DESCRIPCIÓN	COEFICIENTE
<u>edadmeses</u>	Edad en meses del menor	0,616805
<u>f2_s2_208_3</u>	Cuántos hijos viven actualmente con usted (total hijos/as en casa)	0,197654
<u>NEUMOCC2_NO</u>	El menor no tiene la segunda dosis contra el neumococo	0,031208
<u>indígena</u>	Etnia del menor: indígena	0,029906
<u>rural</u>	El área donde vive el menor	0,026239
<u>Educación Básica</u>	Nivel de instrucción de la madre	0,024773
<u>AGUA BEBERagua embotellada</u> <u>/envasada</u>	El agua que bebe el menor es agua embotellada	0,023311
<u>tierra?</u>	Material del piso donde vive el menor	0,021065
<u>BANIONo tiene</u>	Baño de la vivienda del menor: no tiene	0,020384
<u>TECHOpalma/paja/hoja</u>	Material del techo de la vivienda	0,008656

**Fuente:** Tabla obtenida en Python con Visual Studio Code.

A continuación, mediante gráficos bivariados encontramos la distribución de las variables que intervienen en los modelos, con respecto a la variable dependiente:

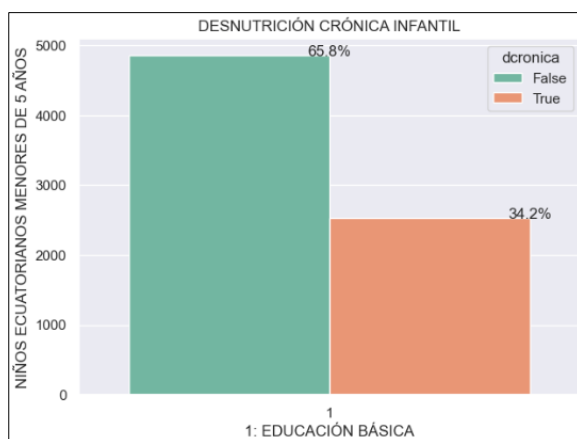
*tierra:* El 43,2% de los menores que habitan una vivienda con piso de tierra tienen desnutrición crónica infantil, como se visualiza en la figura 23.



**Figura 23:** Gráfico bivariado de la variable 'tierra' con respecto a la variable dependiente DCI.

**Fuente:** Gráfico obtenido usando Python en Visual Studio Code, a partir de la base abierta de ENSANUT.

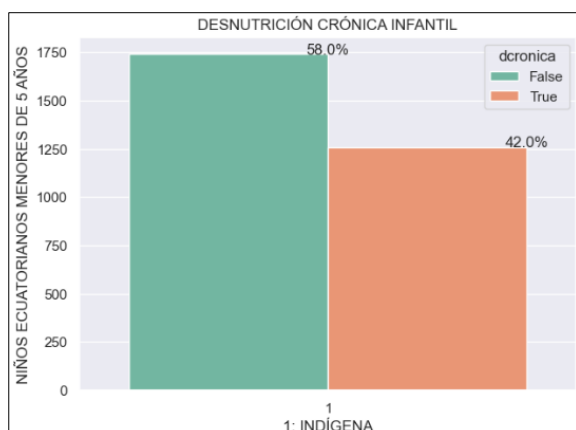
*Educación Básica:* En la figura 24 se visualiza que, del total de niños, cuya madre tiene como instrucción educación básica, el 34,2% tienen desnutrición crónica infantil.



**Figura 24:** Gráfico bivariado de la variable 'Educación Básica' con respecto a la variable dependiente DCI.

**Fuente:** Gráfico obtenido usando Python en Visual Studio Code, a partir de la base abierta de ENSANUT

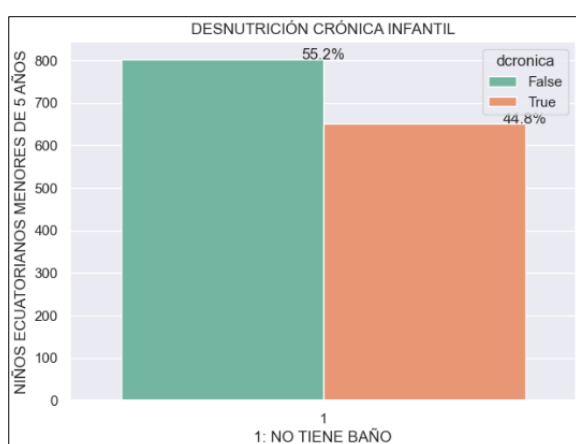
*indígena:* El 42% de niños de etnia indígena tienen desnutrición crónica infantil, tal como se visualiza en la figura 25



**Figura 25:** Gráfico biviado de la variable 'indígena' con respecto a la variable dependiente DCI.

**Fuente:** Gráfico obtenido usando Python en Visual Studio Code, a partir de la base abierta de ENSANUT.

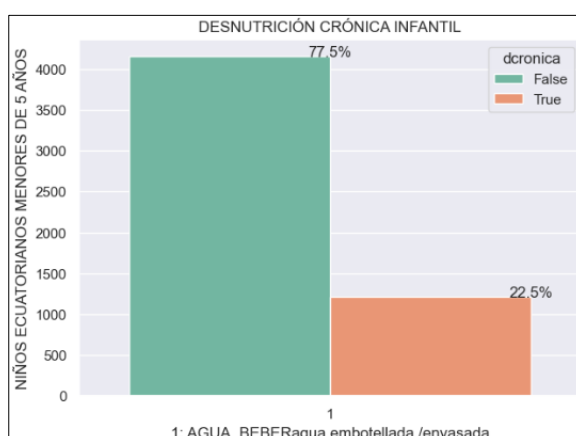
**BANIO**no tiene: El 44,8% de los menores que habitan una vivienda sin baño, tienen desnutrición crónica infantil, como se muestra en la figura 26.



**Figura 26:** Gráfico biviado de la variable 'BANIO'no tiene' con respecto a la variable dependiente DCI.

**Fuente:** Gráfico obtenido usando Python en Visual Studio Code, a partir de la base abierta de ENSANUT.

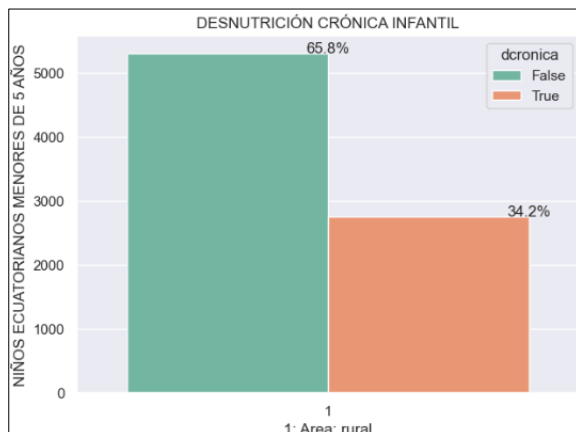
**AGUA\_BEBE**Ragua embotellada /envasada: El 22,5% de los menores que beben agua embotellada o envasada, tienen desnutrición crónica como se visualiza en la figura 27.



**Figura 27:** Gráfico biviado de la variable 'AGUA\_BEBE'Ragua embotellada' con respecto a la variable dependiente DCI.

**Fuente:** Gráfico obtenido usando Python en Visual Studio Code, a partir de la base abierta de ENSANUT.

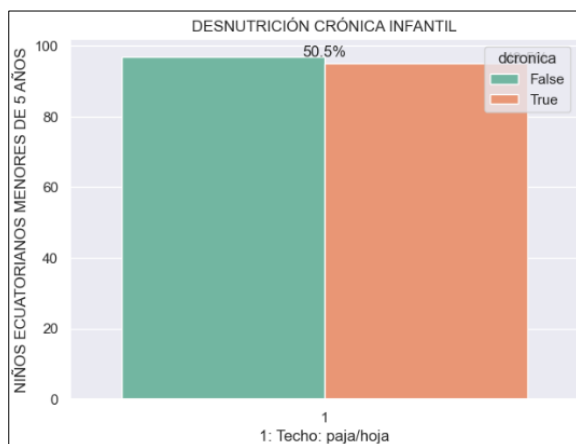
Rural: El 34,2% de los menores que viven en el área rural tienen desnutrición crónica, como se visualiza en la figura 28



**Figura 28:** Gráfico biviado de la variable 'rural' con respecto a la variable dependiente DCI.

**Fuente:** Gráfico obtenido usando Python en Visual Studio Code, a partir de la base abierta de ENSANUT.

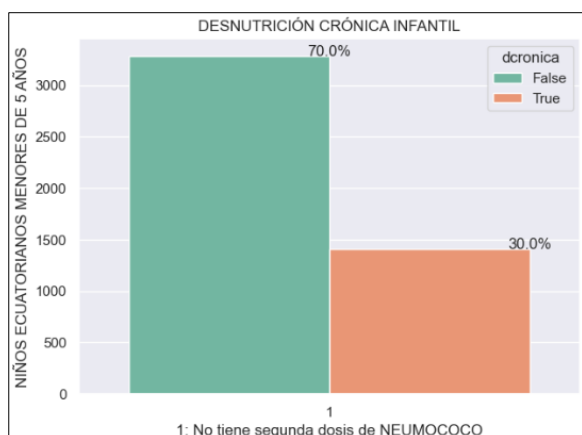
TECHOpalma/paja/hoja: El 49,5% de los menores cuya vivienda tiene un techo con material de palma/paja/hoja tienen desnutrición crónica, como se visualiza en la figura 29.



**Figura 29:** Gráfico biviado de la variable 'TECHOpalma/paja/hoja' con respecto a la variable dependiente DCI.

**Fuente:** Gráfico obtenido usando Python en Visual Studio Code, a partir de la base abierta de ENSANUT.

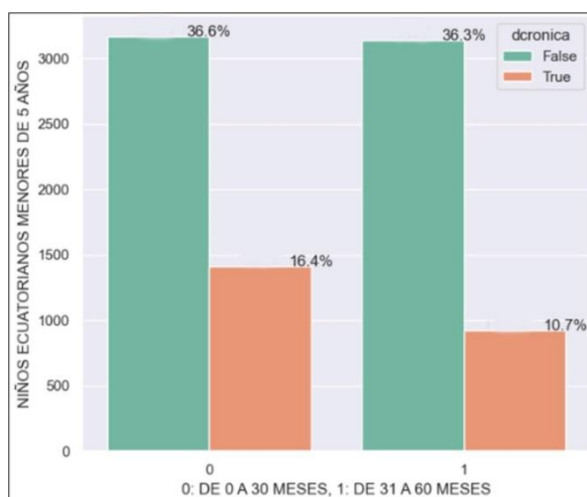
NEUMOCC2\_NO: El 30% de los menores ecuatorianos que no tienen segunda dosis de neumococo tienen desnutrición crónica infantil, como se muestra en la figura 30.



**Figura 30:** Gráfico bivariado de la variable 'NEUMOC2\_NO' con respecto a la variable dependiente DCI.

**Fuente:** Gráfico obtenido usando Python en Visual Studio Code, a partir de la base abierta de ENSANUT.

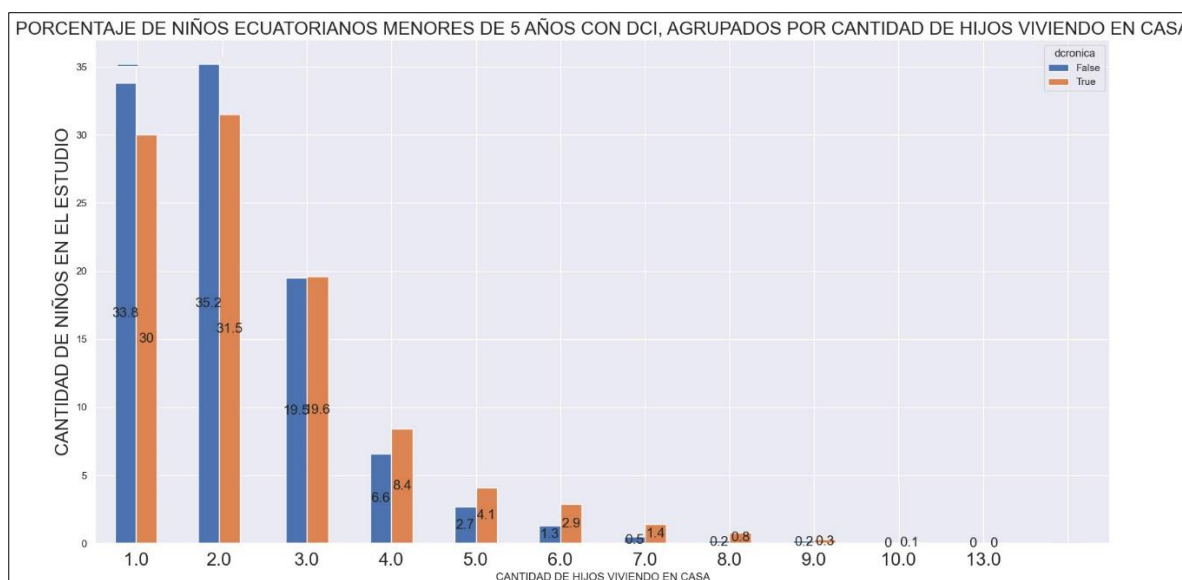
edadmeses: En la figura 31 se visualiza que la mayor cantidad de niños con desnutrición crónica está en el grupo de niños de 0 a 30 meses



**Figura 31:** Porcentajes de niños con y sin DCI agrupados en niños menores a 30 meses y mayores a 30 meses.

**Fuente:** Gráfico obtenido usando Python en Visual Studio Code, a partir de la base abierta de ENSANUT

f2\_s2\_208\_3: Según la figura 32 se visualiza que al aumentar el número de hijos viviendo en casa, el número de niños con DCI es igual o mayor que el número de niños sin DCI.



**Figura 32:** Porcentajes de niños con y sin DCI por el número de hijos que viven en casa.

**Fuente:** Gráfico obtenido usando Python en Visual Studio Code, a partir de la base abierta de ENSANUT.

#### 4.5. MODELO SELECCIONADO PARA LA IDENTIFICACIÓN DE LOS FACTORES QUE INFLUYEN EN LA DESNUTRICIÓN CRÓNICA INFANTIL

Una vez establecidos los modelos Regresión Logística y *Random Forest* y evaluados mediante las métricas de rendimiento, *recall*, matriz de confusión y curva ROC, es necesario establecer el modelo que mejor define la desnutrición crónica infantil. El modelo seleccionado es *Random Forest*, si bien ambos tienen un porcentaje similar de rendimiento, al igual que la curva ROC, en este caso el criterio de selección es el porcentaje de predicción de verdaderos positivos es decir el porcentaje que predijo correctamente la mayor cantidad de casos de menores con DCI, el modelo *Random Forest* predice correctamente un porcentaje de 22,18% a diferencia del modelo Regresión Logística que únicamente predice un 5% del total de menores con DCI.

#### 4.6. ANÁLISIS DE CORRESPONDENCIA DE LAS PRINCIPALES VARIABLES CATEGÓRICAS

Para este análisis se filtran las observaciones, únicamente con registros donde la variable *dcronica* sea igual a 1 y en el segundo caso donde esta variable sea igual a 0, es importante cambiar el tipo de datos, la mayoría de las variables al ser *dummies* tienen únicamente dos valores (0 y 1), por lo que las cambiamos a variables de tipo categóricas. Después de este paso obtenemos dos conjuntos de datos los cuales son la base del análisis de correspondencia, en el caso de las variables, se usa siglas para que la visualización e interpretación del gráfico sea más simple, estas se describen en la tabla 12:

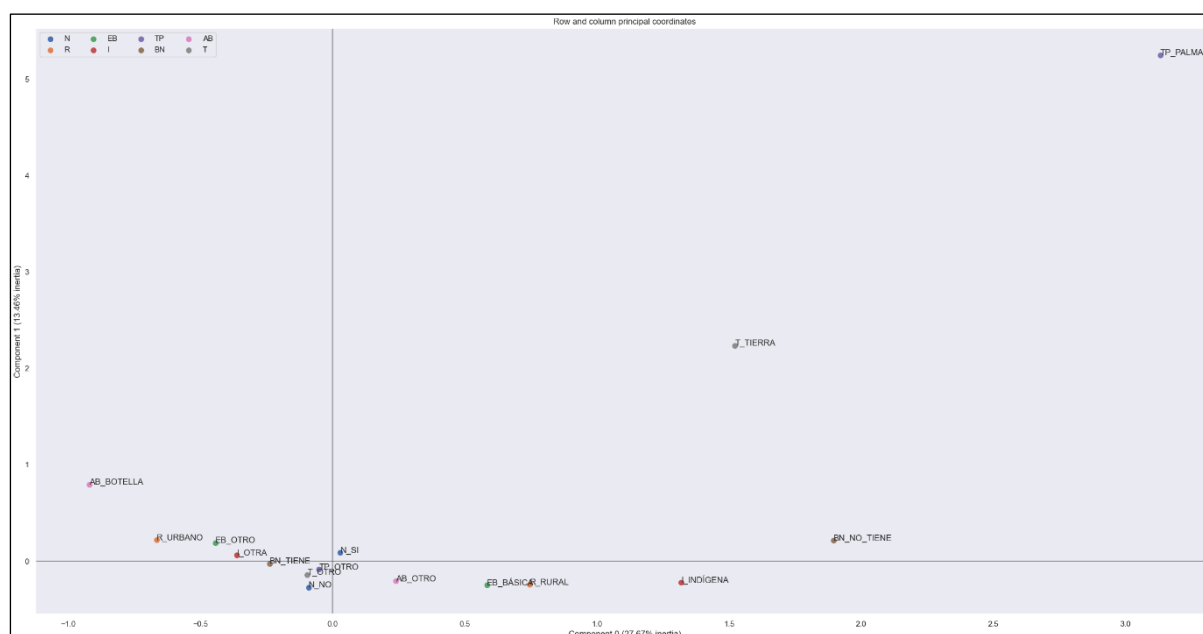
**Tabla 12:** Variables que intervienen en el análisis de correspondencia

VARIABLE	DESCRIPCIÓN	RESPUESTA
<i>N</i>	Segunda dosis de vacuna contra neumococo	SI/NO
<i>R</i>	Área donde se ubica la vivienda	URBANO/RURAL
<i>EB</i>	Nivel de instrucción de la madre: Educación Básica	OTRO/BÁSICA
<i>I</i>	Etnia del menor: Indígena	OTRA/INDÍGENA
<i>TP</i>	Material del techo de la vivienda del menor	OTRO/PALMA
<i>BN</i>	La vivienda del menor tiene o no baño	TIENE/NO_TIENE
<i>AB</i>	Fuente de agua para beber	OTRO/BOTELLA
<i>T</i>	Material del piso de la vivienda del menor	OTRO/TIERRA

**Fuente:** Tabla obtenida en Python con Visual Studio Code, a partir de la base abierta de ENSANUT.

#### 4.6.1. ANÁLISIS DE CORRESPONDENCIA DE OBSERVACIONES DE NIÑOS MENORES DE 5 AÑOS CON DESNUTRICIÓN CRÓNICA INFANTIL

En este análisis disponemos de un conjunto de datos con un total de 5.849 observaciones, que es el total de ecuatorianos menores de 5 años con DCI, en la figura 32, se muestra el análisis de correspondencia, de donde obtenemos algunos indicios de la similitud entre las categorías de ciertas variables:



**Figura 33:** Análisis de correspondencia de las principales variables categóricas con la data de menores ecuatorianos de 5 años con DCI.

**Fuente:** Gráfico obtenido usando Python en Visual Studio Code, a partir de la base abierta de ENSANUT.

- #### 4.6.2. ANÁLISIS DE CORRESPONDENCIA DE OBSERVACIONES DE NIÑOS MENORES DE 5 AÑOS SIN DESNUTRICIÓN CRÓNICA INFANTIL

[illegible]

**Fuente:** Gráfico obtenido usando Python en Visual Studio Code, a partir de la base abierta de ENSANUT.



- Un grupo donde las variables con gran similitud son: “Educación del menor es diferente a educación básica”, “Etnia del menor diferente a indígena” y “La vivienda del menor tiene baño”
- Otro grupo de variables con similitud es: “El techo de la vivienda es diferente a paja/palma” y “El piso de la vivienda es diferente a tierra”.
- La variable que indica “Nivel de instrucción de la madre: Educación Básica” tiene similitud con la variable “Área donde se encuentra la vivienda: rural”
- La variable que indica “La vivienda del menor no tiene baño” tiene similitud con “Etnia del menor: indígena”.
- La variable que no tienen similitud con ninguna es “El material del piso de la vivienda: tierra”

## 5. CAPITULO 5: CONCLUSIONES Y RECOMENDACIONES

### 5.1. CONCLUSIONES

- Durante el análisis exploratorio de datos se observaron variables con *outliers*, los cuales no fueron eliminados debido a que, en este caso pueden ser el punto de interés y varían de acuerdo a múltiples factores como el social, económico, geográfico, cultural, etc., tal es el caso de: tiempo que demoró la madre en hacerse el primer control posparto, tiempo que demoró la madre en hacerse el primer control médico durante el embarazo, semanas de embarazo al hacerse el primer control médico, número de controles antes del parto, años cumplidos de la madre y total de hijos en casa.
- Los dos modelos desarrollados, tienen un porcentaje de exactitud (*accuracy*) bueno, el modelo de Regresión Logística tiene un porcentaje de 71,16% y el modelo *Random Forest* 68,2%. De ambos modelos se obtuvo la matriz de confusión para medir su rendimiento, el modelo de regresión logística es mejor para predecir casos de menores sin DCI, mientras que el modelo *Random Forest* es mejor prediciendo casos de menores con DCI.
- En ambos casos, se obtuvo una curva ROC mayor a 0,6 por lo que se determina que los dos modelos tienen una probabilidad alta de diferenciar entre valores positivos y negativos.
- Las variables más importantes para identificar la DCI son: segunda dosis de la vacuna contra el Neumococo: no, área donde habita el menor: rural, nivel de instrucción de la madre: Educación Básica, etnia del menor: indígena, material del techo de la vivienda del menor: palma/ paja/ hoja, baño de la vivienda del menor: no tiene, fuente del agua que bebe el menor: embotellada /envasada, material del piso de la vivienda del menor: tierra.
- Después de la evaluación de los modelos, el modelo óptimo para predecir la DCI en ecuatorianos menores de 5 años, es *Random Forest* ya que es el modelo que más casos de verdaderos positivos predice correctamente.
- Existen varios métodos automáticos y manuales para seleccionar las mejores variables y crear modelos con un buen rendimiento, Python ofrece algunas librerías que seleccionan de manera automática el mejor grupo de variables para la creación del modelo óptimo.
- La aplicación del análisis de correspondencia es ventajosa ya que visualmente nos permite descubrir y mostrar las similitudes entre categorías, la distancia entre dos puntos muestra la fuerza de la similitud entre dos características.

## 5.2. RECOMENDACIONES

- Al seleccionar las variables independientes para la creación de los modelos es importante eliminar las correlaciones fuertes (multicolinealidad) entre las variables independientes para que al final los modelos sean fáciles de interpretar y las variables independientes representen una única información.
- Uno de los pasos importantes para el modelado, es el ajuste de hiperparámetros ya que cada conjunto de datos requiere diferentes hiperparámetros, para esto es necesario realizar varias pruebas experimentando con diferentes valores hasta conseguir un modelo con los mejores resultados.
- El análisis exploratorio de datos permite constituir el conjunto de datos ideal, es decir preparar los datos para la implementación de métodos estadísticos, por lo que se recomienda ser cuidadoso en todo el proceso previo al modelado de datos.
- Se recomienda a las organizaciones nacionales hacer uso de los resultados del presente estudio, para que enfoquen su ayuda a disminuir de manera efectiva la desnutrición crónica infantil, teniendo una visión global del problema podrán crear proyectos, viables y principalmente oportunos para detectar casos de DCI en los primeros meses de vida del recién nacido, evitando un retraso en su crecimiento, causa que es irreversible.
- Existen otras fuentes que se pueden contemplar, como ENDEMAIN que proporciona datos con cobertura nacional clasificados por regiones y provincias, sobre características de los miembros del hogar, se recomienda obtener variables significativas de esta fuente, que influyan en la desnutrición crónica infantil, y agregarlas a los modelos ya existentes, con el fin de analizar su comportamiento y obtener modelos más precisos y con mejor rendimiento.

## 6. BIBLIOGRAFIA

- Acción contra el hambre. (2020). *NUTRICIÓN Y SALUD*. Obtenido de Desnutrición: prevención, diagnóstico y tratamiento: <https://www.accioncontraelhambre.org/es/que-hacemos/nutricion-salud>
- Agencia de la ONU para los Refugiados. (5 de mayo de 2020). *Desnutrición infantil en el mundo*. Obtenido de UNHCR ACNUR: <https://eacnur.org/es/actualidad/noticias/emergencias/desnutricion-infantil-en-el-mundo#:~:text=La%20regi%C3%B3n%20m%C3%A1s%20afectada%20por,que%20sufren%20may>
- Cardellino, F. (22 de marzo de 2021). *Tutorial para un clasificador basado en bosques aleatorios: cómo utilizar algoritmos basados en árboles para el aprendizaje automático*. Obtenido de freeCodeCamp: <https://www.freecodecamp.org/espanol/news/random-forest-classifier-tutorial-how-to-use-tree-based-algorithms-for-machine-learning/>
- Comunidad de Python. (28 de mayo de 2023). *Python*. Obtenido de <https://pypi.org/project/featurewiz/#introduction>
- Dagnino, J. (2014). *Datos Faltantes*. Obtenido de Revista chile de Anestesia: <https://revistachilenadeanestesia.cl/datos-faltantes-missing-values/>
- Desarrolladores de scikit learn. (2023). *scikit learn*. Obtenido de scikit learn: [https://scikit-learn.org/stable/modules/feature\\_selection.html](https://scikit-learn.org/stable/modules/feature_selection.html)
- Detección de outliers en Python*. (2 de junio de 2020). Obtenido de Aprende Machine Learning: <https://www.aprendemachinelearning.com/deteccion-de-outliers-en-python-anomalia/>
- E., S. (2023). *Understand Random Forest Algorithms with examples*. Obtenido de Analytics Vidhya: <https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/>
- IBM. (13 de septiembre de 2022). *Métodos de selección de variables en el análisis de regresión lineal*. Obtenido de IBM: <https://www.ibm.com/docs/es/spss-statistics/saas?topic=regression-linear-variable-selection-methods>
- Instituto Nacional de Estadística y Censos. (2018). *Guía de uso de base de datos de ENSANUT*. Obtenido de INEC: [chrome-extension://efaidnbmnnnibpcajpcgclefindmkaj/https://www.ecuadorencifras.gob.ec/documentos/web-inec/Estadisticas\\_Sociales/ENSANUT/ENSANUT\\_2018/Guia%20de%20BDD%20ENSANUT%202018.pdf](chrome-extension://efaidnbmnnnibpcajpcgclefindmkaj/https://www.ecuadorencifras.gob.ec/documentos/web-inec/Estadisticas_Sociales/ENSANUT/ENSANUT_2018/Guia%20de%20BDD%20ENSANUT%202018.pdf)
- Instituto Nacional de Estadísticas y Censo. (2018). *Diseño muestral de la Encuesta Nacional de Salud y Nutrición 2018*. Obtenido de INEC: [chrome-extension://efaidnbmnnnibpcajpcgclefindmkaj/https://www.ecuadorencifras.gob.ec/documentos/web-inec/Estadisticas\\_Sociales/ENSANUT/ENSANUT\\_2018/Metodologia%20del%20diseño%20muestral%20ENSANUT%202018.pdf](chrome-extension://efaidnbmnnnibpcajpcgclefindmkaj/https://www.ecuadorencifras.gob.ec/documentos/web-inec/Estadisticas_Sociales/ENSANUT/ENSANUT_2018/Metodologia%20del%20diseño%20muestral%20ENSANUT%202018.pdf)

- Instituto Nacional de Estadísticas y Censos. (2018). *Evolución histórica de la Encuesta Nacional de Salud y Nutrición 2018 - 2019*. Obtenido de INEC: [chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.ecuadorencifras.gob.ec/documentos/web-inec/Estadisticas\\_Sociales/ENSANUT/ENSANUT\\_2018/Evolucion%20Historica%20de%20ENSANUT%202018.pdf](chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.ecuadorencifras.gob.ec/documentos/web-inec/Estadisticas_Sociales/ENSANUT/ENSANUT_2018/Evolucion%20Historica%20de%20ENSANUT%202018.pdf)
- Instituto Nacional de Estadísticas y Censos. (2018). *Salud, Salud Reproductiva y Nutrición*. Obtenido de INEC: <https://anda.inec.gob.ec/anda/index.php/catalog/891>
- Lantz, B. (2019). *Estadística y Machine Learning con R*. Obtenido de Métodos de selección de variables en el modelo general de aprendizaje: <https://bookdown.org/content/2274/metodos-de-clasificacion.html>
- Lantz, B. (2019). *Métodos de clasificación*. Obtenido de Estadística y Machine Learning con R: <https://bookdown.org/content/2274/metodos-de-clasificacion.html>
- Malo, N., Mejía, M., & Vinuesa, B. (2015). *Situación de la desnutrición crónica en niños y niñas de servicios de Desarrollo Infantil Integral*. Obtenido de Ministerio de Unclusión Económica y Social: [chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/http://sitp.pichincha.gob.ec/repositorio/disenos\\_paginas/archivos/Desnutricion%20investigacion%20MIES.pdf](chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/http://sitp.pichincha.gob.ec/repositorio/disenos_paginas/archivos/Desnutricion%20investigacion%20MIES.pdf)
- Márquez, C. (5 de agosto de 2021). *Alarma por alto índice de desnutrición infantil en Ecuador*. Obtenido de EL COMERCIO: <https://www.elcomercio.com/actualidad/ecuador/desnutricion-infantil-ecuador-enfermedades-alimentacion.html>
- Ministerio de Asuntos Económicos y Tranformación Digital. (22 de septiembre de 2021). *Ministerio de Asuntos Económicos y Tranformación Digital*. Obtenido de [datos.gob.ec](https://datos.gob.es/es/documentacion/guia-practica-de-introduccion-al-analisis-exploratorio-de-datos): <https://datos.gob.es/es/documentacion/guia-practica-de-introduccion-al-analisis-exploratorio-de-datos>
- Ministerio de Salud Pública. (2019). *Encuesta Nacional de Salud y Nutrición*. Obtenido de MSP: <https://www.salud.gob.ec/encuesta-nacional-de-salud-y-nutricion-se-presenta-este-miercoles/>
- Organización Mundial de la Salud. (9 de junio de 2021). *Malnutrición*. Obtenido de OMS: <https://www.who.int/es/news-room/fact-sheets/detail/malnutrition>
- Pramoditha, R. (30 de abril de 2022). *Why do we set a random state in machine learning models?* Obtenido de <https://towardsdatascience.com/why-do-we-set-a-random-state-in-machine-learning-models-bb2dc68d8431>
- Ramírez, M. (26 de agosto de 2014). *Desnutrición infantil: causas, tipos y repercusiones*. Obtenido de Alimenta: <https://www.dietistasnutricionistas.es/desnutricion-infantil/>
- Random Forest Algorithm*. (s.f.). Obtenido de javaTpoint: <https://www.javatpoint.com/machine-learning-random-forest-algorithm>

- Random Forest, el poder del Ensemble.* (17 de junio de 2019). Obtenido de Aprende Machine Learning: <https://www.aprendemachinelearning.com/random-forest-el-poder-del-ensemble/>
- Roy, B. (10 de julio de 2020). *Técnicas de Imputación de datos faltantes (Un análisis profundo).* Obtenido de DATASOURCE.AI: <https://www.datasource.ai/es/data-science-articles/todo-sobre-el-manejo-de-datos-faltantes>
- Sancho, F. (7 de enero de 2021). *Medir la eficacia de un aprendizaje.* Obtenido de <http://www.cs.us.es/~fsancho/?e=231>
- Statistical Discovery.* (2022). Obtenido de Análisis exploratorio de datos: [https://www.jmp.com/es\\_co/statistics-knowledge-portal/exploratorydata-analysis.html](https://www.jmp.com/es_co/statistics-knowledge-portal/exploratorydata-analysis.html)
- TIBCO. (2023). *¿Qué es el aprendizaje supervisado?* Obtenido de TIBCO: <https://www.tibco.com/es/reference-center/what-is-supervised-learning>
- TIBCO. (2023). *¿Qué es la regresión logística?* Obtenido de TIBCO: <https://www.tibco.com/es/reference-center/what-is-logistic-regression>
- TIBCO CLOUD. (2023). *ANALISIS DE CORRESPONDENCIA.* Obtenido de ¿Qué es el análisis de correspondencias?: <https://www.tibco.com/es/reference-center/what-is-correspondence-analysis>
- UNICEF. (2019). *La desnutrición crónica infantil afecta el desarrollo económico y social del Ecuador.* Obtenido de UNICEF: [chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.unicef.org/ecuador/sites/unicef.org.ecuador/files/2021-04/DCI\\_Desarrollo-economico-social\\_UNICEF%20.pdf](chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.unicef.org/ecuador/sites/unicef.org.ecuador/files/2021-04/DCI_Desarrollo-economico-social_UNICEF%20.pdf)
- UNICEF. (octubre de 2019). *Niños, alimentos y Nutrición.* Obtenido de UNICEF: <chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.unicef.org/media/61091/file/Estado-mundial-infancia-2019-resumen-ejecutivo.pdf>
- UNICEF. (22 de junio de 2022). *Cada minuto, un niño sufre malnutrición grave en 15 países afectados por la crisis mundial del hambre.* Obtenido de UNICEF: <https://www.unicef.org/es/comunicados-prensa/cada-minuto-nino-sufre-malnutricion-grave-15-paises-afectados-crisis>

## ANEXO 2. CÓDIGO DE LAS FUNCIONES CREADAS PARA FACILITAR EL ANÁLISIS DE LOS DATOS

- *FUNCION PARA REEMPLAZAR LOS VALORES NULOS DE UN DATAFRAME POR LA MEDIA DE LOS VALORES VALIDOS: NUMERICO*

```
def reemplazar_media(df):  
    for column in df.columns:  
        media = df[column].mean()  
        df[column] = df[column].fillna(round(media,2))  
    return df
```

- *FUNCION PARA REEMPLAZAR LOS VALORES NULOS DE UN DATAFRAME POR EL VALOR PREDOMINANTE: CATEGORICA*

```
def nulos(df):  
    for column in df.columns:  
        df_valores = df[column].value_counts()  
        fd=df_valores.index  
        df[column] = df[column].fillna(fd[0])  
    return df
```

- *FUNCION PARA CONVERTIR A BOOLEANO*

```
def convertir_booleano(df):  
    for column in df.columns:  
        df[column] = df[column].replace(['si'],1)  
        df[column] = df[column].replace(['no'],0)  
        df[column] = df[column].astype(bool)  
    return df
```

- *FUNCION PARA ELIMINAR OUTLIERS*

```
def eliminar_outliers(df):  
    for column in df.columns:  
        Q1 = df[column].quantile (q = .25)  
        Q3 = df[column].quantile (q = .75)  
        IQR = df[column].apply (stats.iqr)  
        df[column] = df[column] [~ ((df[column] <(Q1-1.5 * IQR)) | (df[column]> (Q3 + 1.5 * IQR)))]  
    return df
```

- *FUNCION PARA OBTENER LAS VARIABLES DUMMIES*

```
def dummies(df,columna,nombre1,nombre2,df_global):
    cp=pd.get_dummies(df)
    cp=cp.rename(columns={'no':nombre1,'si':nombre2})
    df_global=pd.concat([df_global,cp],axis=1)
    df_global.drop([columna],axis = 'columns', inplace= True)
    return df_global
```

- *FUNCION PARA GRAFICAR DISTRIBUCIÓN DE VARIABLES SOBRE TARGET*

```
def grafico_uno(df,columna,titulo1,titulo2,label1):
    cros=pd.crosstab(index=df['dcronica'], columns=df[columna], normalize="index")
    cros1=cros*100
    cros1=round(cros1, 1)
    cros1= cros1.transpose()
    cros_dci = cros1.sort_values(True,False,ascending=False)
    ax = cros_dci.plot(kind='bar', figsize=(20,10), stacked=False)
    for c in ax.containers:
        ax.bar_label(c, label_type='center',fontsize=20)
    plt.xticks(rotation=360, fontsize=20)
    plt.xlabel(label1)
    plt.ylabel(titulo1,fontsize=20)
    plt.title(titulo2,fontsize=20)
    return plt
```

- *FUNCION PARA GRAFICAR BIVARIADO*

```
def grafico_bivariado(df,columna,titulo,titulo1):
    total = float(len(df))
    ax=sns.countplot(data=df, x=columna,hue="dcronica",palette='Set2')
    for label in ax.containers:
        ax.bar_label(label)
    ax.set(title = titulo, xlabel = titulo1, ylabel = 'NIÑOS ECUATORIANOS MENORES DE 5 AÑOS')
    for p in ax.patches:
        percentage = '{:.1f}%'.format(100 * p.get_height()/total)
        x = p.get_x() + p.get_width()
        y = p.get_height()
```



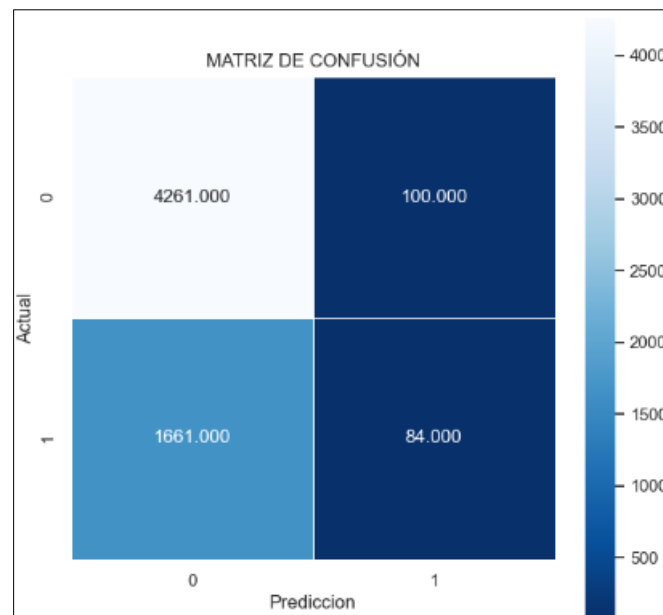
```

        ax.annotate(percentage, (x, y), ha='center')
    return plt

• FUNCION PARA EXPORTAR DATASET
def exportar_dataset(df,columna,link):
    df = (df.groupby('dcronica')[columna]
          .value_counts(normalize=True)
          .reset_index(name='porcentaje'))
    df.to_excel(link)
    return df

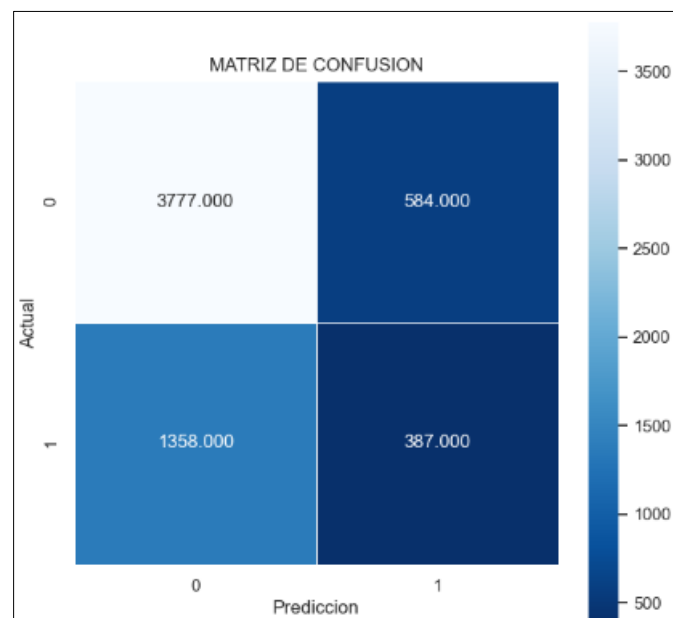
```

### ANEXO3. MATRIZ DE CONFUSIÓN DEL MODELO REGRESIÓN LOGÍSTICA Y RANDOM FOREST.



**Figura 35:** Matriz de confusión del modelo Regresión Logística

**Fuente:** Gráfico obtenido usando Python en Visual Studio Code a partir del modelo Regresión Logística



**Figura 36:** Matriz de confusión del modelo Random Forest

**Fuente:** Gráfico obtenido usando Python en Visual Studio Code a partir del modelo Random Forest