ESCUELA SUPERIOR POLITÉCNICA DE CHIMBORAZO

FACULTAD DE CIENCIAS

CARRERA ESTADÍSTICA

COMPARATIVA ENTRE MODELOS DE ÁRBOLES DE CLASIFICACIÓN Y REGRESIÓN PARA PREDECIR LA PARASITOSIS INTESTINAL EN NIÑOS DE 5 A 9 AÑOS EN EL HOSPITAL PEDÍATRICO ALFONSO VILLAGÓMEZ ROMÁN

Trabajo de Titulación

Tipo: Proyecto de Investigación

Requisito para obtener el grado académico de:

INGENIERO/A ESTADÍSTICO/A

AUTORES:

ISIN DAQUI WENDY ESTEFANIA

LÓPEZ SARMIENTO MAICOL AMABLE

DIRECTORA: ING. JOHANNA ENITH AGUILAR REYES, MGS.

Riobamba – Ecuador

2023

© 2023, Wendy Estefania Isin Daqui y Maicol Amable López Sarmiento

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

Nosotros, **WENDY ESTEFANIA ISIN DAQUI Y MAICOL AMABLE LÓPEZ SARMIENTO**, declaramos que el presente Trabajo de Titulación es de nuestra autoría y los resultados de este son auténticos. Los textos en el documento que provienen de otras fuentes están debidamente citados y referenciados.

Como autores asumimos la responsabilidad legal y académica de los contenidos de este Trabajo de Titulación; el patrimonio intelectual pertenece a la Escuela Superior Politécnica de Chimborazo.

Riobamba, 12 de abril de 2023

Wendy Estefania Isin Daqui
060404845-4

Maicol Amable López Sarmiento
140097283-0

ESCUELA SUPERIOR POLITÉCNICA DE CHIMBORAZO

FACULTAD DE CIENCIAS

CARRERA ESTADÍSTICA

El Tribunal del Trabajo de Titulación, certifica que: El Trabajo de Titulación; Tipo: Proyecto de Investigación, **COMPARATIVA ENTRE MODELOS DE ÁRBOLES DE CLASIFICACIÓN Y REGRESIÓN PARA PREDECIR LA PARASITOSIS INTESTINAL EN NIÑOS DE 5 A 9 AÑOS EN EL HOSPITAL PEDIÁTRICO ALFONSO VILLAGÓMEZ ROMÁN,** realizado por los señores: **WENDY ESTEFANIA ISIN DAQUI Y MAICOL AMABLE LÓPEZ SARMIENTO**, ha sido minuciosamente revisado por los Miembros del Tribunal del Trabajo de Titulación. El mismo que cumple con los requisitos científicos, técnicos, legales, en tal virtud el Tribunal autoriza su presentación.

	FIRMA	**FECHA**
Ing. Natalia Alexandra Pérez Londo, Mgs. **PRESIDENTE DEL TRIBUNAL**	_____	12-04-2023 _____
Ing. Johanna Enith Aguilar Reyes, Mgs. **DIRECTORA DEL TRABAJO DE TITULACIÓN**	_____	12-04-2023 _____
Ing. Paulina Fernanda Bolaños Logroño, Mgs. **ASESORA DEL TRABAJO DE TITULACIÓN**	_____	12-04-2023 _____

DEDICATORIA

Este trabajo va dedicado para mis padres Juan López y Shervin Sarmiento, que a pesar de la distancia siempre me mostraron su amor y cariño incondicional, para mi hermana Jahayra López esperando que algún día pueda cumplir sus metas propuestas, para la Dra. Kimberly Cambisaca quien ha sido mi compañera de travesía y mi impulso para jamás rendirme ante algún problema, a mi familia por ser el pilar más importante en mi vida y mi formación como persona.

Maicol

Este trabajo en primer lugar quiero dedicar a Dios por otorgarme vida y salud para culminar con un peldaño más en mi vida estudiantil. También a cada integrante de mi familia Isin Daqui por ser un soporte importante a lo largo de mi carrera universitaria, en especial a mis padres Orlando Isin y Gloria Daqui por depositar su confianza en mí y creer en todos los sueños y anhelos que me he propuesto además de su apoyo incondicional tanto económico y moral que me ha permitido hoy en día formarme como persona y profesional. Finalmente, a mis abuelitos María Sanaguano y Segundo Daqui (+) que con su amor y enseñanza me guiaron por el buen camino y con su ejemplo de superación y perseverancia me demostraron que soy fuerte y puedo vencer todos los obstáculos que se presenten.

Wendy

AGRADECIMIENTO

En primer lugar, agradezco a Dios por otorgarme esta magnífica oportunidad, agradezco también a la Dra. Kimberly Cambisaca, mi amante, amiga y compañera, por ser quien me ha apoyado en cada momento de esta travesía. A la Ing., Johanna Aguilar y a la Ing. Paulina Bolaños quienes me guiaron y tutelaron en el desarrollo de este trabajo de investigación. A mis padres quienes me han ayudado tanto económica como sentimentalmente durante toda mi vida. A la Biofísica Dalila Hidrovo, que gracias a ella decidí estudiar entra grandiosa carrera. A toda mi familia quienes me han apoyado y soportado en todas las facetas de mi vida. Mi corazón siempre estará agradecido con mis maestros y amigo con quienes compartí momentos de alegría, triunfos y derrotas, con todos aquellos que me apoyaron en este largo proceso, gracias.

Maicol

Quiero empezar agradeciendo a nuestras docentes Ing. Johanna Aguilar e Ing. Paulina Bolaños que con su experiencia y conocimiento guiaron este trabajo de investigación logrando alcanzar las expectativas y objetivos de este, a toda la planta docente y administrativa de la carrera de Estadística que con su tiempo, trabajo y dedicación permitieron que una generación más egrese como Ingenieros/as Estadísticos.

Al Hospital Pediátrico Alfonso Villagómez Román especialmente a la Ing. Valeria Vacacela por la oportunidad para poner en práctica los conocimientos adquiridos y a su vez contribuir para la realización del trabajo de titulación.

A mis compañeros y amigos con quienes compartí gratos momentos dentro y fuera de la institución y a todas las personas que en algún momento de mi vida con sus palabras y anécdotas contribuyeron para situarme en el lugar que hoy me encuentro.

Wendy

ÍNDICE DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

ÍNDICE DE GRÁFICOS

ÍNDICE DE ECUACIONES

ÍNDICE DE ANEXOS

ÍNDICE DE ABREVIATURAS

HPAVR:	Hospital Pediátrico Alfonso Villagómez Román
OMS:	Organización Mundial de la Salud
OPS:	Organización Panamericana de la Salud
CIBV:	Centro Infantil del Buen Vivir
IC:	Intervalo de confianza
CHAID:	Chi-squared Automatic Interaction Detection
ROC:	Receiver Operanting Characteristic
AUC:	Área bajo la curva ROC
AIC:	Criterio de Información Akaike
CART:	Árbol de clasificación y regresión
MCAR:	Missing completely at random
MAR:	Missing at random
NMAR:	Not missinng at random

RESUMEN

El presente trabajo de investigación compara dos técnicas de clasificación: Árboles de clasificación y Regresión Logística para predecir la parasitosis intestinal en niños de 5 a 9 años atendidos en el Hospital Pediátrico Alfonso Villagómez Román de la ciudad de Riobamba, con la finalidad de establecer el mejor modelo cuyos factores asociados sean significativos para la variable en estudio. La base de datos para el análisis fue proporcionada por el Área de Estadística de la casa de salud, a través de los repositorios existentes se tomó información del periodo 2019 -2021, para el procesamiento de información se depuró aquellos registros que podían presentar inconsistencias, la validación para la nueva matriz se realizó comparando las historias clínicas de los usuarios obteniendo un nuevo registro con 2675 pacientes y 14 variables de interés. Los modelos encontrados se obtuvieron usando el software SPSS y su modelador SPSS MODELER, empleado para evaluar la capacidad predictiva a través de las medidas de bondad de ajuste: Tasa de error y Curva Roc; con este análisis, el modelo por árboles de clasificación fue el mejor por tener una precisión general del 65,7% y una tasa de error del 24,49%, los factores asociados al diagnóstico son: Triage, Saturación de oxígeno, Temperatura axilar, Talla, Peso, Frecuencia cardiaca mínima y Tipo de seguro de salud. Cabe indicar que la diferencia en la eficiencia de predicción para cada técnica fue mínima por lo que sería factible probar otras técnicas, ya sea siguiendo la línea de árboles de decisión o modelos de regresión.

Palabras clave: <ESTADÍSTICA>, <PRONÓSTICOS>, <ÁRBOLES DE CLASIFICACIÓN>, <REGRESIÓN LOGÍSTICA>, <PARASITOSIS INTESTINAL>.

0731-DBRA-UPT-2023

SUMMARY

This research work compares two classification techniques: Classification Trees and Logistic Regression to predict intestinal parasitism in children from 5 to 9 years of age treated at the Alfonso Villagómez Román Pediatric Hospital in the city of Riobamba, in order to establish the best model whose associated factors are significant for the variable under study. The database for the analysis was provided by the Statistics Area of the health home, through the existing repositories, information was taken from the period 2019 -2021, for the information processing, those records that could present inconsistencies were purged, the validation for the new matrix was performed by comparing the clinical histories of the users, obtaining a new record with 2675 patients and 14 variables of interest. The models found were obtained using the SPSS software and its modeler SPSS MODELER, used to evaluate the predictive capacity through the goodness-of-fit measures: Error rate and Roc Curve; with this analysis, the classification tree model was the best for having a general precision of 65.7% and an error rate of 24.49%. The factors associated with the diagnosis are: Triage, Oxygen saturation, Axillary temperature, Height, Weight, Minimum heart rate and Type of health insurance. It should be noted that the difference in prediction efficiency for each technique was minimal, so it would be feasible to test other techniques, either by following the line of decision trees or regression models.

Keywords: <STATISTICS>, <FORECASTS>, <TREES CLASIFICATION>, <LOGISTIC REGRESSION>, <INTESTINAL PARASITES>.

Edgar Mesias Jaramillo Moyano
0603497397

INTRODUCCIÓN

La estadística es un área aplicable a distintos campos en los que el ser humano se desenvuelve, uno de ellos la medicina, en la cual, a través de las técnicas para la recolección, análisis y el procesamiento de la información generada por los establecimientos de salud tiene como objetivo otorgar posibles soluciones a los diversos problemas que enfrentan, contribuyendo a la correcta toma de decisiones, satisfaciendo las necesidades propias y de los usuarios.

Las enfermedades ocasionadas por parásitos intestinales está considerada como uno de los problemas más comunes en la salud humana, estos patógenos afectan a poblaciones de todas las edades, pero con mayor frecuencia son los niños quienes se encuentran vulnerables al contagio por distintos factores como el entorno social, alimentación, higiene y recientemente según estudios se ha incluido la migración como una causa relevante ya que la gran afluencia de personas de distintos lugares en el mundo provocó un incremento notorio en los diagnósticos, sobre todo ha alertado acerca de nuevas especies de parásitos no tan habituales en la región, siendo esto un problema para las diferentes casas de salud quienes se enfrentan a un panorama fuera de lo común y en ocasiones no cuenta con el personal ni el equipo adecuado para sobrellevar el tratamiento de la enfermedad.

Antecedentes investigativos

Las naciones unidas en su artículo "OMS alerta sobre infección de parásitos intestinales en países en desarrollo" menciona que, gran parte por no mencionar la mayoría de los niños residentes en países en vías de desarrollo se han visto afectados por gusanos intestinales contribuyendo a la malnutrición de estos y, disminuye las probabilidades de crecimiento, desarrollo y aprendizaje (OMS, 2008).

(Batista Rojas & Álvarez Hernández, 2013) en su artículo "Parasitismo intestinal en niñas y niños mayores de 5 años de Ciudad Bolívar" se aplicó un estudio descriptivo y transversal a 320 pacientes mayores de 5 años atendidos en este recinto durante julio 2011- marzo 2012 con la finalidad de caracterizarlos según algunas variables de interés: tipos de parásitos, particularidades del abastecimiento y tratamiento del agua de consumo, lugar de deposición, hábitos higiénico-sanitarios y síntomas más frecuentes se concluyó que, los pacientes presentaban poliparistismo cuyo síntoma relevante es el dolor abdominal y se atribuye a la ausencia de sanidad en el entorno puesto que los habitantes se abastecen de agua almacenada en pipas o tanques además de, la ausencia de baterías sanitarias y el hábito de andar.

(Montero Perez & Huilca Peralta, 2018) en su trabajo de investigación "Parasitosis intestinal, estado nutricional y diagnóstico bacteriológico en manos de niños de un jardín de la zona rural de Huancayo", consideró la investigación de tipo aplicada y diseño cuantitativo - no experimental descriptivo, se realizó un examen parasitológico, así como también un diagnóstico bacteriológico

donde se hizo la siembra en agar Macconkey, y por último se realizó un cuestionario para saber el estado nutricional de los niños concluyendo que el factor estado nutricional con relación a su edad, peso y talla fue baja.

En Ecuador los diversos síntomas y manifestaciones de la parasitosis constituyen las diez primeras causas de consulta pediátrica. Durante la infancia es frecuente la anemia en niños parasitados que, a largo plazo se convierte en alteraciones del desarrollo ponderal, psicomotriz e intelectual (Menacho Chávez, 2022).

La Organización Panamericana de la Salud, en el año 2016, realizó un estudio con el nombre de "Prevalencia de parasitismo intestinal en niños quechuas de zonas rurales montañosas del Ecuador", se analizó la relación entre algunas variables sanitarias (el uso de letrinas, la disponibilidad de métodos adecuados de almacenamiento y tratamiento de agua dentro de esto también se estudió la aplicación de proyectos comunitarios para proteger las fuentes de agua potable) además de las enfermedades y la prevalencia de parasitismo infantil. Se buscó una gran variedad de parásitos, la mayoría de ellos patógenos, que se transmiten por diversas vías, como el agua, los alimentos, el suelo y las heces fecales. Se realizó exámenes coproparasitario a 149 niños de comunidades rurales quechuas que viven en las montañas de la provincia de Chimborazo, dando como resultado el 85,7% de las muestras presentaban al menos uno de los 10 parásitos estudiados y 63,4% contenían dos o más tipos de parásitos. No se encontraron diferencias significativas entre el número de casos informados de enfermedades como: diarrea, fiebre, infecciones respiratorias, vómitos y otras (Revista Panamericana de la Salud Pública, 2008).

(Pazmiño Gómez, et al., 2018) en su artículo " Parasitosis intestinal y estado nutricional en niños de 1-3 años de un centro infantil del Cantón Milagro", corresponde a un estudio cuantitativo de carácter descriptivo, apoyado por una investigación de campo que permitió la recolección de datos antropométricos y las muestras fecales para la realización de exámenes coprológicos a 38 niños que asisten frecuentemente al CIBV, dando como resultados lo siguiente: 23 (60,5%) de los niños y niñas presentaron parasitosis intestinal, quienes se ven afectados con bajo peso debido a la presencia de parásitos, en la encuesta a los padres de familia como parte de los instrumentos de investigación, más del 65% de los hogares no tienen una adecuada norma de cuidado para prevenir la infección por parásitos intestinales, el consumo de agua sin hervir así como la ingesta de frutas y legumbres sin lavado previo. Concluyendo que la carencia de servicios básicos, hábitos de higiene alimentaria, el desconocimiento de los riesgos es determinanteX para la presencia de parásitos intestinales en los infantes y por ende afecta el estado nutricional.

Por otra parte, en el artículo titulado "Árboles de Clasificación vs. Regresión Logística en el desarrollo de habilidades genéricas en ingeniería" desde un enfoque experimental evalúa el desempeño de ambos modelos en el contexto de dos habilidades genéricas de ingeniería (razonamiento cuantitativo y comprensión lectora), incorporando dos escenarios predictores

separados (solo indicadores y solo construcciones derivadas del análisis de componentes principales “ACP”), dentro del cual los dos métodos presentan un desempeño similar respecto a indicadores mientras que en el escenario de construcciones la regresión logística destaca significativamente (Perez Rave & Gonzalez Echeverria, 2018).

En la tesis titulada “Comparación de modelos de clasificación: regresión logística y árboles de clasificación para evaluar el rendimiento académico” se comparan los modelos de regresión logística binaria y árboles de clasificación (CHAID) para evaluar el rendimiento académico. El comportamiento de estos modelos fue medido por cuatro indicadores: Sensibilidad, Curva ROC, Índice de GINI e Índice de Kappa, determinado que el árbol de clasificación es modelo más adecuado para clasificación y predicción (Lizares Castillo, 2017).

Todos los estudios descritos previamente son referencias importantes para el desarrollo del presente trabajo puesto que, cada investigación tiene un enfoque, metodología y resultados distintos, pero llegan a coincidir en que la infección por parásitos intestinales es una enfermedad común alrededor del mundo, siendo los niños el grupo más propenso a contraer estas bacterias debido a que por la falta de conocimiento y cuidado influyen en la correcta evolución de los menores; inclinándonos a indagar minuciosamente sobre la problemática y brindarle a la unidad de salud un antecedente correctamente argumentado que contribuya a la toma de decisiones para controlar o erradicar dicha patología.

Antecedentes históricos

El Hospital Pediátrico Alfonso Villagómez Román tiene su origen en el siglo XX entre los años 1928 -1929, es una obra histórica a la cual se le atribuye los reconocimientos respectivos a los médicos Miguel Ángel León Pontón y Alfonso Villagómez Román quienes en conjunto a la iglesia y la junta cívica que se encontraba celebrando el primer centenario de la Republica del Ecuador y con una destacada labor de las mujeres de la época se funda el Centro General de Cultura Social cuya finalidad estaba dedicada a atender a toda la población infantil en todos los ámbitos y con los recursos necesarios.

Con el auspicio del ministerio de Previsión Social y Trabajo con oficio N°361 del 21 de abril de 1929 expide el Acuerdo 326 en donde se aprueba los estatutos del Centro General de Cultura Social dando paso a la fundación en la ciudad de Riobamba “La Gota de Leche” con características de Dispensario Médico y Casa Cuna.

La inauguración se realizó el 17 de noviembre en presencia de los personajes más connotados de la Sultana de los Andes, posteriormente en 1938 se realiza la bendición e inauguración solemne del hospital a cargo del primer director el Señor Alfonso Villagómez, médico riobambeño nacido el 13 de diciembre de 1902 a quien se le atribuye la mayor obra social de la época y de la ciudad la creación del “Hospital de Niños” la primera casa de salud de la Republica del Ecuador (Vallejo Samaniego, 2010).

También fue el responsable de la creación de una clínica quirúrgica donde se llevó a cabo notables operaciones, su vocación como médico y colaborador social lo llevo a dedicar gran parte de su vida a hacer el bien y a curar los males humanos, una relevante muestra de esto es que prestaba gratuitamente sus servicios profesionales a la clase trabajadora.

Además, a inicios del año 1939 se presenció un brote de peste bubónica en Riobamba, a pesar del peligro que esta enfermedad mortal conlleva esta ilustre figura enfrento la misión y logro salvar la vida de la población, sin embargo, el no corrió con la misma suerte puesto que se contagió de esta enfermedad y falleció el 14 de febrero de 1939. En honor y reconocimiento a su indudable desempeño profesional y social la casa de salud lleva su nombre.

Para la fecha de inauguración la entidad ya contaba don dos salas de internación cuyos nombres estaban en homenaje a las primeras mujeres presidentas e impulsadoras de este trabajo, cada una contenía 10 camas y una sala de pensión con cuatro camas de dotación. En el transcurso de los años los bienes fueron incrementando hasta alcanzar 84 camas de dotación normal, la sala de consulta externa, una incipiente Botica área destinada a la despensa, cocina y comedor entre sus principales servicios.

Posterior a la muerte de quien fue su director y promotor Dr. Villagómez por consenso entre el directorio del centro y el municipio del cantón se realizó una expropiación y posterior venta del inmueble ubicado en la calle España, el cual fue utilizado para ampliar las dependencias de la entidad y así el 9 de julio de 1940 por todo lo alto se colocó la primera piedra en la construcción de lo que sería el nuevo pabellón del Hospital en la ciudad obra que estuvo a cargo del constructor reconocido Neptalí Tornen y según documento que reposan en el recinto se dice que el diseño de la estructura es similar a los construidos en Europa, probablemente de Suiza.

A lo largo de los años el funcionamiento de la entidad únicamente se atribuye a sus benefactores (Centro General de Cultura Social) aunque la administración fue llevada por varios dirigentes y en los últimos cuarenta años estuvo sujeto a la dirección del Señor Doctor Ramiro Guerrero Casares.

En la administración del presidente General Guillermo Rodríguez Lara, mediante Decreto Supremo 232 del 25 de abril de 1972 en su Art. IV. Dispone que el Hospital de Niños Alfonso Villagómez Román pase a manos del Ministerio de Salud.

El Hospital de niños como comúnmente es conocido nació por los deseos de servir a la población desvalida que sumada a las inmensas ganas de trabajar en ese voluntariado dieron paso a que el anhelo de la época sea realidad cabe mencionar que, no era frecuente que ciudad alguna del país cuente con un nosocomio exclusivo para atender enfermedades en los niños y al lograrse este proyecto se constituyó como un hito histórico en la medicina ecuatoriana y del voluntariado social (Bonilla Pulgar & Bonilla Nina, 2020).

Con lo antes mencionado cabe indicar que en este proyecto de investigación se establecen los principales factores que intervienen en la parasitosis intestinal a través de un marco teórico conceptual empleando la terminología adecuada misma que, se utilizó para la comparación entre las técnicas: árboles de clasificación y regresión aplicada para la categorización de individuos a través de indicadores asociados al modelo matemático y en conjunto con las pruebas para la validación del mismo: Tasa de error del área ROC y de la matriz de confusión se determinó como "el mejor" a aquel cuya efectividad de predicción para la parasitosis intestinal en niños de 5 a 9 años en el HPAVR fue mayor y así, contribuimos con la institución de salud entregando un precedente investigativo que posteriormente puede ser replicado para otras patologías y/o casas de salud a nivel local y nacional.

Finalmente, la investigación comprende varios capítulos, el capítulo I engloba el problema de investigación en conjunto con su respectiva justificación, objetivos y la hipótesis a comprobar al final de este trabajo investigativo. El capítulo II abarca el marco teórico referencial en donde con brevedad se describe teóricamente las técnicas empleadas: Árboles de clasificación y Regresión logística y, respecto a la parasitosis intestinal se menciona: conceptos, clasificación y factores asociados a este.

En el capítulo III con respecto al marco metodológico se indica el enfoque, nivel y diseño de investigación en donde se describe el proceso para la recolección y análisis de la información en base a la unidad de análisis y el tamaño muestral; la selección y descripción de variables que en conjunto con los algoritmos de clasificación empleados se determinó la factibilidad para la solución del problema.

En el capítulo IV se presenta el análisis de resultados obtenidos en la caracterización por árboles de clasificación y regresión logística para así finalmente presentar las conclusiones y recomendaciones respectivamente.

CAPÍTULO I

1. PROBLEMA DE INVESTIGACIÓN

1.1. Planteamiento del Problema

A lo largo del tiempo se ha establecido a la estadística y sus divisiones como una herramienta importante para la realización de investigaciones en diversas áreas, destacando a la salud y educación, aspectos sobre los cuales la aplicación de las innumerables técnicas para el análisis y modelado de datos son relevantes para la toma de decisiones puesto que según varias bibliografías el empleo de la instrumentación surge por la enorme incertidumbre que se presenta en diferentes fenómenos dentro de estos campos y por ende para su correcta comprensión es necesario diseñar procesos para la recolección, clasificación y tratamiento de los datos con la finalidad de extraer una gran cantidad de información relevante sobre los sujetos de estudio.

En el Ecuador según estudios, la población infantil es la más afectada por parásitos intestinales los cuales debido a distintos factores ya sea sociales u inmunológicos ingresan en el organismo de los niños influyendo en el correcto desarrollo atacándolos silenciosamente. Al tratarse de un fenómeno multifactorial originado por múltiples elementos que actúan a nivel relacional, para su prevención y atención nos vemos en la necesidad de establecer factores asociados a la patología y se consideran las siguientes técnicas: árboles de clasificación y regresión logística, mismas que establecerán lo antes mencionado y, además, cabe indicar que las pruebas están sujetas a errores sin embargo, individualmente sus ventajas podrían ser alentadoras es por ello la importancia de su comparación puesto que de ambas una se establecerá como mejor técnica cuando el modelo entregado tenga una alta capacidad predictiva.

1.2. Limitaciones y delimitaciones

Una limitación respecto a los registros de pacientes con parasitosis intestinal en el HPAVR es que a la casa de salud acuden niños con diagnósticos específicos y son atendidos por médicos especialistas, por ende, no se encuentra un número considerable de casos respecto a esta patología ya que la mayoría de estos son tratados en las unidades de primer o segundo nivel del MSP, además que, por cuestiones de confidencialidad el expediente médico es inaccesible sin la autorización del representante. Dentro de las delimitaciones, se tomará como muestra a los pacientes de 5 a 9 años registrados en el hospital desde el año 2019 hasta 2021.

1.3. Problema General de Investigación

¿Cuál de los modelos: ¿Árboles de clasificación o regresión, predice con mayor exactitud la parasitosis intestinal en niños de 5 a 9 años en el Hospital Pediátrico Alfonso Villagómez Román?

1.4. Problemas específicos de Investigación

- ¿Cuáles son los factores asociados a la parasitosis intestinal en los niños de 5 a 9 años?
- ¿Cómo ejecutar la clasificación de los pacientes mediante arboles de clasificación y regresión logística?
- ¿Cuáles son los resultados obtenidos de la comparación entre los modelos obtenidos por arboles de clasificación y regresión logística para predecir la parasitosis intestinal en niños de 5 a 9 años en el HPAVR?

1.5. Objetivos

1.5.1. Objetivo General

- Comparar las metodologías de árboles de clasificación y regresión para predecir la parasitosis intestinal en niños de 5 a 9 años en Hospital Pediátrico Alfonso Villagómez Román.

1.5.2. Objetivos Específicos

- Determinar a través de un marco teórico pertinente los posibles factores asociados a la parasitosis intestinal.
- Recolectar la información de los posibles factores, usando las historias clínicas de los niños de 5 a 9 años, atendidos en el Hospital Pediátrico Alfonso Villagómez Román.
- Predecir los factores asociados a la parasitosis intestinal a través de un modelo de árboles de clasificación y de regresión.
- Comparar la efectividad de los modelos usados con base a la precisión de la predicción.

1.6. Justificación

1.6.1. Justificación Teórica

En el grupo de enfermedades infecciosas, específicamente aquellas cuyos patógenos principales son parásitos se consideran un problema de salud importante para el ser humano; la mayoría de estos son agentes patógenos frecuentes en el mundo y, son una causa principal de morbilidad y mortalidad en regiones como África, Asia, América Central y América del Sur.

Esta investigación tiene como finalidad proporcionar información relevante sobre la clasificación de individuos de acuerdo a características estrechamente ligadas a un factor de interés, en estudios previos se menciona que a menudo se genera información de tamaño incalculable y cuyo tratamiento es necesario para generar una nueva, la misma que debe poder ser utilizada para investigaciones, estudios con predicciones contribuyendo a la toma de decisiones, además que con los software adecuados todo el contenido no sea solo un respaldo sino que con el proceso adecuado los datos proporcionen nuevo conocimiento. De acuerdo con lo descrito nuestro trabajo tiene como propósito describir la información sobre los pacientes con parásitos intestinales mediante pronósticos para contribuir al HPAVR una base investigativa fundamentada con aporte estadístico el cual permitirá a la casa de salud establecer parámetros para el diagnóstico de esta enfermedad en sus usuarios.

1.6.2. Justificación Metodológica

Con el presente trabajo de investigación a través de la revisión bibliográfica adecuada se determinará los posibles factores asociados a la parasitosis intestinal que permitirán la clasificación de los niños de 5 a 9 años registrados en el HPAVR generando un modelo asociado de acuerdo con las técnicas estadísticas: Árboles de clasificación y Regresión respectivamente.
Una vez obtenido dichos modelos se evaluará y se establecerá como el "mejor" aquel cuya capacidad de predicción sea mayor a través de la comparación de las medidas de bondad de ajuste: Tasa de error de la matriz de confusión y el área bajo la curva ROC y así, utilizarlo para la predicción de la parasitosis intestinal en los infantes.

1.6.3. Justificación Práctica

Una problemática de interés dentro del análisis estadístico es la clasificación de objetos o individuos ya sea en grupos o poblaciones, por ello se han desarrollado técnicas para alcanzar dicho objetivo, entre las conocidas está el análisis discriminante en conjunto con sus variaciones y es muy utilizada sin embargo requiere de los supuestos de normalidad y homocedasticidad mismo que frecuentemente no se cumplen, recurriendo a técnicas que no incluyan lo mencionado por ejemplo la regresión logística y en conjunto con otras técnicas como los árboles de clasificación que son poco estudiadas se pretende analizar el desempeño de dichas técnicas y determinar bajo qué condiciones cuál de los modelos posee una capacidad predictiva mayor para establecer la parasitosis intestinal en los niños (Serna Pineda, 2009).

1.7. Hipótesis

Los modelos de regresión tienen una capacidad predictiva mejor que los árboles de clasificación.

CAPITULO II

2. MARCO TEÓRICO

2.1. Bases teóricas

2.1.1. Modelo de regresión logística

Es una técnica estadística multivariante que ayuda a modelar una variable de respuesta categórica en función de variables explicativas o predictoras cuantitativaso categóricas. Forma parte de los GLM (modelos lineales generalizados) y se aplica en diferentes campos como las ciencias de la salud, ciencias sociales, economía, ecología, entre otros. El objetivo de los modelos logit es estimar probabilidades o predecir un suceso definido por la variable respuesta categórica en función de las variables predictoras (Congacha Ortega, 2020).

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

También, el análisis de regresión logística comprende la estimación de la probabilidad de que ocurra un evento (variable respuesta dicotómica) como función de los valores de p variables independientes. Consideremos Y una variable respuesta y p una colección de variables independientes expresadas por el vector $X' = (x_1, x_2, ..., x_p)$.

El modelo por RL con $p - variables$ predictoras se especifica de la siguiente forma:

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

Donde, se representa la probabilidad condicional de que el evento $Y = 1$ ocurra dada la ocurrencia de un conjunto de variables X (probabilidad de éxito) (Congacha Ortega, 2020).

Una transformación de (x) fundamental para el estudio de regresión logística es la transformación logit y se define en términos de (x), como:

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

Ecuación 1-2: Transformación logit

Donde β_0 es la constante y los β_i son los coeficientes de los predictores x_i del modelo. La importancia de esta transformación es que (x) posee muchas de las propiedades deseables de un modelo de regresión lineal. La función logit es lineal en sus parámetros, puede ser continuo y variar de $-\infty\ a + \infty$, dependiendo del rango de x (Congacha Ortega, 2020).

El modelo logístico puede expresarse en términos de odds (disparidad o ventaja) de ocurrencia de eventos. Esta razón se define como el cociente entre la probabilidad de éxito y la probabilidad de fracaso (Congacha Ortega, 2020). Esto es:

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

Ecuación 2-2: Odds

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

2.1.1.1. Estimación de los parámetros del modelo

Debido a que la distribución de Y dado un conjunto de variables $X = (x_1, x_2, \ldots, x_p)$ no es normal y no existe homocedasticidad en los errores, la estimación del vector $\hat{\beta} = (\widehat{\beta_1}, \widehat{\beta_2}, \ldots, \widehat{\beta_p})$ por el método de mínimos cuadrados no tiene propiedades óptimas en su lugar emplearemos el método de máxima verosimilitud para obtener los valores de los parámetros desconocidos que maximizan la probabilidad de obtener el conjunto observado de datos

La función de verosimilitud adopta la forma:

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

Ecuación 3-2: Función de verosimilitud

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

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

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

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

$$\sum_{i=1}^{n}[y_i - \pi(x_i)] = 0 \ \forall\, j = 1, \ldots, p$$

Para encontrar la solución de este conjunto de ecuaciones se utiliza el método iterativo de Newton Raphson que se lo puede realizar mediante algún paquete estadístico.

2.1.1.2. Contrastes o pruebas de significancia del modelo

La diagnosis de la regresión logística considera algunos contrastes o pruebas estadísticas de significancia como:

Contraste de Wald: Evalúa estadísticamente los coeficientes de regresión logística. Se quiere contrastar si un parámetro $\beta_i = 0$, con $i = 1, 2, \ldots, k$ frente a si son significativamente distintos de 0, es decir:

$$H_0: \beta_i = 0$$
$$H_1: \beta_i \neq 0$$

mediante el estadístico de Wald:

$$W_i = \frac{\widehat{\beta_\iota}}{SE(\widehat{\beta_\iota})} \sim N(0,1)$$

Ecuación 4-2: Estadístico de Wald

$\widehat{\beta_\iota}$ y $SE(\widehat{\beta_\iota})$ son estimaciones del modelo para $\widehat{\beta_\iota}$ y el error estándar de β_i. Los coeficientes son significativos si tienen un valor $p < 0.05$. También se puede determinar IC para β_i puesto que el estadístico de prueba se distribuye según una normal estándar, entonces los extremos inferior y superior son respectivamente.

$$\widehat{\beta_\iota} - z_{\alpha/2}SE(\hat{\beta}_i) \text{ y } \widehat{\beta_\iota} + \widehat{\beta_\iota} - z_{\alpha/2}SE(\hat{\beta}_i).$$

Existe una estrecha relación entre contrastes de hipótesis e intervalos de confianza. Si el intervalo de confianza incluye el 0, significa que al nivel α elegido no podría rechazar la hipótesis nula de que $\beta_i = 0$ (Congacha Ortega, 2020).

2.1.1.3. Pseudo estadísticas R^2

R² de Cox y Snell: Coeficiente de determinación generalizado utilizado para estimar la proporción de la variabilidad de la variable dependiente explicada por sus predictores. Se compara del logaritmo de la verosimilitud para el modelo respecto al logaritmo de la verosimilitud para un modelo de línea base cuyos valores oscilan entre 0 y 1.

R² de Nagelkerke: Versión corregida de la pseudo estadística anterior esta tiene un valor máximo inferior de 1, incluido el modelo perfecto. Este corrige la escala del estadístico para cubrir el rango completo de 0 a 1.

2.1.1.4. Evaluación de la bondad de ajuste del modelo

AIC: El criterio de información de Akaike es una herramienta estadística útil para elegir el número de retrasos de $p\ y\ q$ basándose en la suma de los cuadrados de los errores que busca minimizarlos a partir de varias combinaciones de los parámetros.

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

Ecuación 5-2: Criterio de Akaike

Donde:
ln: Logaritmo neperiano
$\widehat{\sigma^2}$: Suma residual de cuadrados dividida entre el número de observaciones
n: Número de observaciones
r: Número total de parámetros (incluyendo el término constante)

Test de Hosmer y Lemeshow: Esta prueba evalúa la bondad de ajuste del modelo utilizando una estrategia de agrupamiento para obtener la estadística de bondad de ajuste y, se obtiene mediante el cálculo de la estadística Chi-cuadrado de Pearson de una tabla de frecuencias observadas y esperadas estimadas (Congacha Ortega, 2020). Se prueba las siguientes hipótesis:

$$H_0: No\ existen\ diferencias\ entre\ los\ valores\ observaados\ y\ predichos$$

$$H_1: Existen\ diferencias\ entre\ los\ valores\ observaados\ y\ predichos$$

Si rechazamos H_0, el modelo ajustado no es el adecuado. Se dividen todos los casos en deciles en base a las probabilidades predichas donde el primer decil contabiliza los casos cuyas probabilidades son altas, siendo el estadístico de prueba

$$\hat{C} = \sum_{k=1}^{g} \frac{(0_k - n_k \bar{\pi}_k)^2}{n_k \bar{\pi}_k (1 - \bar{\pi}_k)}$$

Ecuación 6-2: Test de Hosmer y Lemeshow

Donde:

O_k: Número de respuestas entre las covariables

n_k: Número de covariables en el k-ésimo decil

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

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

$$\hat{C} > \chi^2_{1-\alpha}(g - 2)$$

Concluyendo que el modelo no es adecuado (Congacha Ortega, 2020).

2.1.1.5. Ventajas

- Fácil aplicabilidad y comprensión
- El sobreajuste es casi inexistente
- Fácil entrenamiento para grandes grupos de datos gracias a su versión estocástica
- Resultados interpretables

2.1.1.6. Desventajas

- En ocasiones es demasiado simple al momento de captar relaciones complejas entre variables
- Alteración en presencia de datos atípicos.

2.1.2. Árboles de Clasificación y Regresión

Un árbol de clasificación o regresión es una representación gráfica y analítica ante eventos o sucesos que pueden surgir a partir de una elección asumida en un determinado momento y contribuye a tomar una decisión "acertada" con base probabilística permitiendo analizar los resultados y visualizar cómo el modelo se desglosa (Lizares Castillo, 2017).

Algunos métodos basados en la técnica por arboles son:

- Árbol de clasificación: Aplicable a variables categóricas de tipo nominal u ordinal
- Árbol de regresión: Aplicable a variables continuas

2.1.2.1. Algoritmos

CART (Árboles de clasificación y regresión)

(Serna Pineda, 2009) en su trabajo de investigación titulado "Comparación de árboles de regresión y clasificación y regresión logística" menciona que, Leo Breiman en 1984 desarrollo el algoritmo CART, una técnica no paramétrica cuyo resultado en general es un árbol de decisión, en donde las ramas representan conjuntos de decisiones y estas generan reglas sucesivas para continuar la clasificación o partición formando grupos homogéneos respecto a la variable a discriminar. Las particiones se realizan en forma recursiva hasta alcanzar un criterio de "*paranada",* este método utiliza datos históricos para construir el árbol de decisión, y este se emplea para clasificar a los nuevos datos.

(Timofeev, 2004) menciona que, el análisis de árboles CART generalmente consiste en tres pasos:

- Construcción del árbol máximo
- Poda del árbol
- Selección del árbol optimo mediante un procedimiento de validación cruzada ("cross-validation").

CHAID

Chi.square Automatic Interaction Detección (Detección automática de interacciones mediante chi-cuadrado) es un algoritmo estadístico rápido y multidireccional que rápida y eficientemente explora los datos construyendo segmentos y perfiles respecto a la variable dependiente.

CHAID Exhaustivo

Corrección de la técnica CHAID el cual examina todas las posibles divisiones de la variable respuesta.

QUEST

Desarrolla un algoritmo estadístico que selecciona las variables sin sesgo y a partir de estas construye arboles binarios precisos, rápidos y eficientes.

Los árboles de decisión son una herramienta estadística que segmenta, estratifica, predice, reduce información con un filtro de variables identificando interacciones que fusionen categorías para la discreción de variables continuas (Lizares Castillo, 2017).

2.1.2.2. Construcción del árbol máximo

El árbol máximo se construye utilizando un procedimiento de partición binario, comenzando en la raíz, el mismo es un modelo que describe el conjunto de entrenamiento (datos originales) y generalmente es sobre ajustado, es decir contiene una gran cantidad de niveles y nodos que no producen una mejor clasificación y puede ser demasiado complejo.

Cada grupo está caracterizado por la distribución (respuesta categórica), o por la media (respuesta numérica) de la variable dependiente, el tamaño del grupo y los valores de las variables explicativas que lo definen. Gráficamente, el árbol está representado con el nodo raíz (los datos sin ninguna división) al iniciar y las ramas y hojas debajo (cada hoja es el final de un grupo).

2.1.2.3. Calidad del nodo: Función de Impureza

Esta función es una medida que permite determinar la calidad de un nodo, será denotada por $i(t)$. Se diferencian varias medidas de impureza (criterios de particionamiento) que permiten analizar varios tipos de respuesta, las medidas más comunes propuestas por su creador son tres:

Índice de información

Conocido también como índice de entropía el cual se define por:

$$i(t) = \sum_{j} p(j|t) ln\, p(j|t)$$

Ecuación 7-2: Índice de información

El objetivo es encontrar la partición que maximice $\Delta i(t)$ en la ecuación

$$\Delta i(t) = -\sum_{j}^{k} p(j|t) lnp(j|t)$$

Donde $j = 1, ..., k$ es el número de clases de la variable respuesta categórica y $p(j|t)$ la probabilidad de clasificación correcta para la clase j en el nodo t.

Índice de Gini

Se encuentra definido por: $i(t) = \sum_{i \neq j} p(j|t) p(i|t)$

Ecuación 8-2: Índice de Gini

Encontrar la partición que maximice $\Delta \mathrm{i}(t)$ en

$$\Delta i(t) \; = \; -\sum_{j=1}^{k} [p_j(t)]^2$$

Este índice es el más utilizado, en cada división tiende a separar la categoría más grande en un grupo aparte, mientras que el índice de información tiende a formar grupos con más de una categoría en las primeras decisiones.

Índice "Towing"

A diferencia del anterior, este índice busca las dos clases que juntas formen más del 50% de los datos, esto define dos "super categorías" en cada división para las cuales la impureza es definida por el índice de Gini. Aunque el Índice de Towing produce árboles más balanceados, el algoritmo es más lento que la regla de Gini (Serna Pineda, 2009). Para calcular este índice se selecciona la partición s, que maximice

$$\frac{p_L p_R}{4} [\sum_{j} |p(j|t_L) - p(j|t_R)]^2$$

Ecuación 9-2: Índice de Towing

Donde tL y tR representan los nodos hijos izquierdo y derecho respectivamente, pL y pR representan la proporción de observaciones en t que pasaron a tL y a tR en cada caso.

2.1.2.4. Poda del árbol

El árbol obtenido es generalmente sobreajustado por tanto es podado, cortando sucesivamente ramas o nodos terminales hasta encontrar el tamaño "adecuado" del árbol. Existen algunas ideas básicas para resolver el problema de seleccionar el mejor árbol, de forma computacional el procedimiento a realizar es complejo pero viable, solo se necesita considerar un árbol de cada tamaño, es decir, los árboles de la secuencia anidada.

(De'ath & Fabricius, 2000) indica que otra forma de podado es buscar una serie de árboles anidados de tamaños decrecientes, cada uno de los cuales es mejor de todos los árboles de su tamaño. Estos árboles pequeños son comparados para determinar el óptimo. Esta comparación está basada en una función de costo complejidad, $R_\propto(T)$.Para cada árbol T, la función costo-complejidad se define como: $$R_\propto(T) = R(T) + \propto |\tilde{T}|$$

Ecuación 10-2: Costo-complejidad

Donde $R(T)$ es el promedio de la suma de cuadrados entre los nodos, pueden ser la tasa de mala clasificación total o la suma de cuadrados de residuales total dependiendo del tipo de árbol,

$|\tilde{T}|$ es la complejidad del árbol, definida como el número total de nodos del subárbol y α es el parámetro de complejidad.

El parámetro α es un número real mayor o igual a cero, cuando α = 0 se tiene el árbol más grande y a medida que α se incrementa, se reduce el tamaño del árbol.

La función $R_{\propto}(T)$ siempre será minimizado por el árbol más grande, por tanto, se necesitan mejores estimaciones del error, para esto proponen obtener estimadores *"honestos"* del error por "*validación cruzada".*

2.1.2.5. Selección del árbol optimo

Del grupo de árboles anidados se selecciona el árbol óptimo para el proceso se requiere estimar con precisión el error de predicción, para lo cual se realiza un procedimiento de validación cruzada, cuyo objetivo es encontrar la proporción óptima entre la tasa de mala clasificación siendo está el cociente entre las observaciones mal clasificadas y el total de observación, y la complejidad del árbol.

La implementación del procedimiento por validación cruzada puede realizarse de dos formas: Si los datos son suficientes se parte de la muestra, extrayendo la mitad o menos de los datos para construir la secuencia de árboles con los datos que permanecen para predecir en cada árbol la respuesta de los datos sacados al inicio del proceso obteniendo así el error de las predicciones y elegir el que contiene menor error.

En caso de no contar con los datos suficientes se procede por el método de validación cruzada con partición en V (v-fold cross validation), la cual parte de extraer de la muestra de aprendizaje una muestra prueba, con la primera se calculan los estimadores, y el subconjunto sacado se utiliza para comprobar la efectividad de los estimadores obtenidos y utilizarlos como "datos nuevos". El error de predicción es acumulado para estimar el error medio absoluto del conjunto de prueba.

2.1.3. Curva ROC

Evaluar las técnicas de clasificación es importante para la validación del modelo sobre el conjunto de entrenamiento, además permite la comparación entre las técnicas aplicadas y selecciona aquel con tenga mayor precisión (Congacha Ortega, 2020).

La curva ROC (Receiver Operating Characteristic curve) mide la bondad de ajuste de forma gráfica comparando la tasa de negativos verdaderos frente a la tasa de positivos verdaderos según el umbral de discriminación; donde, los modelos de predicción que se encuentran por encima de

la línea discriminante son mejor cuanto más separados están de la línea. Si los modelos coinciden con la línea pueden clasificarse como aleatorios y los que están por debajo de esta se consideran peores o que tienen algún error en la variable explicativa (Lizares Castillo, 2017).

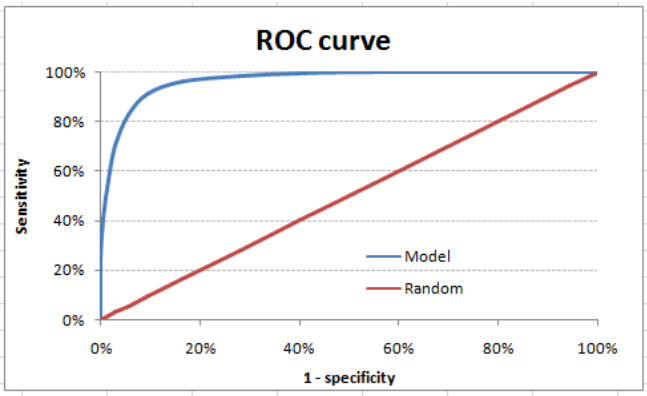


Figura 1-2: Parcela Curva ROC

Fuente: Srivastava, 2019

La elección de la mejor curva ROC se realiza comparando el espacio bajo la curva denominada AUC está comprendida entre 0.5 y 1, donde el máximo valor indica predicción perfecta y 0.5 una predicción aleatoria. Si los valores obtenidos son menores a 0.5 puede haber problemas de concepto por lo que siempre se elige aquella curva que tenga mayor AUC con respecto a otras (Srivastava, 2019).

2.1.4. Datos Perdidos o faltantes

La presencia de valores perdidos (información ausente o faltante) es un problema común en cualquier investigación, y no puede ser ignorado en el análisis de datos, pues puede ser de grave repercusión en la pérdida de potencia del análisis, hasta en la aparición de sesgos inaceptables. La eliminación de entes con este problema limita la representatividad o validez externa de los resultados del estudio, a pesar de que es algo prácticamente inevitable en las investigaciones (Abellana Sangra & Farran Codina, 2015).

Según (Viada, et al., 2016) los datos perdidos se clasifican en tres tipos y son:

- **Datos perdidos completamente al azar (MCAR):** Cuando la probabilidad de qué un sujeto presente un valor ausente en una variable no depende ni de la propiavariable ni de ninguna otra variable recogida.
- **Datos perdidos al azar (MAR):** Cuando la probabilidad de noobservar un dato depende de otras variables, pero no de los valores de la variable con valores perdidos.

- **Datos perdidos no debidos al azar (NMAR):** Cuando la probabilidad de que un sujeto presente un valor faltante depende de dicha variable con valores perdidos.

2.1.4.1. Tratamiento para datos faltantes

Little y Rubin en 1989 presentaron una técnica para probar si los datos ausentes constituyen o no un conjunto de números aleatorios mediante el denominado test conjunto de aleatoriedad de Little, basado en la distribución $\chi 2$ (chi-cuadrado). Una vez que se prueba que existe aleatoriedad en los datos ausentes, se puede iniciar cualquier análisis estadístico.

(Viada, et al., 2016) menciona que tiempo atrás las únicas herramientas para tratar el problema de datos perdidos eran métodos como la eliminación de los casos que contiene valores faltantes o la imputación de ellos por valores plausibles como la media de la variable o la estimación obtenida utilizando regresión sobre las demás variables.

2.1.4.2. Eliminación de los casos

(Abellana Sangra & Farran Codina, 2015) indica dos formas de eliminar datos perdidos: eliminación de los casos (listwise) o eliminación por pares (pairwise). En el primer caso el sujeto con datos perdidos se elimina del análisis; en los datos MCAR la eliminación no presenta sesgo pero el tamaño de la muestra se reduce por ende afectaría en el contraste de hipótesis (disminuyendo) en el error estándar de estimación (aumentando) y además se descartaría la demás información del sujeto. Por otro lado, en la eliminación por pares llamado también análisis de los casos disponibles, se elimina el sujeto del análisis cuando los datos son perdidos en la variable que se precisa para el análisis, pero se incluye el sujeto en los análisis en los que se disponga información. En este caso el tamaño de la muestra a analizar es consistente en todas las estimaciones realizadas.

2.1.4.3. Imputación simple o múltiple

(Abellana Sangra & Farran Codina, 2015) menciona que, la imputación es un proceso de reemplazar los datos perdidos por estimaciones e indica varios métodos para proceder: imputación mediante la media, imputación mediante regresión, imputación mediante el algoritmo de esperanza-maximización e imputación múltiple.

El primer método consiste en reemplazar los datos perdidos por la media de los datos no perdidos, el inconveniente es que puede atenuar cualquier correlación entre las variables que se han imputado valores. En el método por regresión los datos perdidos son reemplazados por el valor predicho de la regresión que se deriva de los datos. En contraste con la imputación de la media, este valor está condicionado a la información que se dispone de los sujetos. En otro caso si se

procede por el algoritmo de expectación-maximización (EM algorithm), se asume una distribución de los datos perdidos parcialmente y la inferencia se base en la verosimilitud bajo esta distribución. Al ser un proceso iterativo, se repiten los siguientes pasos hasta la convergencia; en el paso E se calcula la expectativa condicional de los datos perdidos, condicionando a los valores observados y las estimaciones actuales de los parámetros y estas estimaciones se imputan a los datos perdidos; en el paso M, se calculan las estimaciones máximo-verosímiles de los parámetros, y este método no considera la incertidumbre de los datos perdidos.

De otra manera, en la imputación múltiple en lugar de imputar un valor único para cada dato perdido, estos se sustituyen por m datos simulados que indican la incertidumbre del valor a imputar; cada computo genera un conjunto de datos diferentes los cuales se analizan por separado, obteniendo m estimaciones con sus respectivos errores estándar; donde, la estimación global corresponde al promedio de todas las estimaciones y el error estándar de la estimación se realiza calculando la varianza intra-imputaciones, promedio de los errores estándar m, así como la varianza entre las imputaciones, varianza muestral de las m estimaciones, se suman las dos varianzas y la raíz cuadrada determina el error estándar de la estimación. En este método se ingresa la incertidumbre de los datos perdidos en el error estándar estimado y la varianza entre las m estimaciones refleja incertidumbre estadística debido a los datos perdidos. Finalmente, si los datos son del tipo MCAR no hay sesgo en los datos y si son escasos el método listwise es una buena opción; si son MAR es recomendable la imputación múltiple, aunque la imputación máxima verosimilitud y por regresión también son recomendables; y, si lo datos son NMAR entonces estos métodos a menudo están sesgados y existen métodos específicos para este tipo de datos.

2.2. Bases conceptuales

2.2.1. Parásito

Un parásito es aquel organismo que vive sobre un huésped o en el interior de este y consume los nutrientes de este, existen tres clases distintivas de parásitos que pueden llegar a provocar enfermedades en los humanos y son: protozoos, helmintos y ectoparásitos (Global Health, Division of Parasitic Diseases and Malaria, 2022).

Se considera parasito a todo ser vivo, animal o vegetal que pasa una parte o la totalidad de su existencia en el interior o en el exterior de otros seres vivos, animales o vegetales de diferente especie, a expensas del cual se nutren, ocasionándole daño aparente o inaparente. Es importante mencionar que un parasito puede vivir por un periodo como comensal, pero siempre tendrá la potencialidad genética de producir daño; en cambio, un comensal nunca provoca daño (Apt Baruch, 2013).

2.2.2. Parasitosis intestinal

Los parásitos intestinales, son organismos parasitarios que se alojan o viven en los intestinos incluso llegando a reproducirse dentro de los mismos, los cuales causan graves enfermedades en los seres humanos como por ejemplo infecciones, diarrea, heces acuosas. Los parásitos pueden ser adquiridos en guarderías, escuelas o incluso donde la higiene es muy mala o se encuentra ausente, estos casos generalmente se dan en las afueras de una localidad donde la población de estos sectores no está presenta una cultura de aseo muy sofisticada, siendo los niños los más vulnerables o propensos en contagiarse con estos parásitos.

La OMS/OPS calcula que 20-30% de todos los latinoamericanos están infectados por helmintos intestinales, parásitos intestinales, mientras que las cifras en los barrios pobres alcanzan con frecuencia el 50% y hasta el 95% en algunos grupos indígenas. La mayor frecuencia de estas enfermedades enteroparasitarias se observa en los sectores rurales, por las condiciones de vida para el individuo (Paho, s.f.) .

2.2.2.1. Características generales

El tracto digestivo en el ser humano puede almacenar una gran diversidad de parásitos como los antes mencionados, los mismos que pueden ser comensales o patógenos, este último lo pueden desarrollar los parásitos y no tiene relación el tamaño de este puesto que si comparamos la medida de una ameba en micrones esta puede causar la muerte mientras que una lombriz solitaria con varios metros de longitud apenas causa una sintomatología.

La mayor vía de contagio para contraer una infección por parásitos es la digestiva y algunas veces cutánea, dentro de los mecanismos de transmisión algunos tienen relación con los ciclos evolutivos y genéricamente respectivamente entre estos podemos mencionar cuatro modalidades:

- Infección por fecalismo
- Infección por carnivorismo
- Infección por el ciclo ano-mano-boca
- Infección por la piel

Dentro de lo que concierne al concepto de saneamiento básico este abarca las calidades de la disposición de las excretas, agua de bebida y riego, eliminación de basura, pululación de las moscas y mataderos consideradas las más importantes. Los factores socioeconómicos y la cultura higiénica son de importancia obvia y decisiva en la difusión de la parasitosis intestinal.

Ante la susceptibilidad del huésped hacia las infecciones parasitarias estas dependen de la inmunidad natural ya sea por factores genéticos y de nutrición.

En lo que respecta a la prevalencia de la parasitosis con sus variantes a lo largo del tiempo se ha mantenido estable pues las actuales cifras son iguales a las de hace 20 años y se dice que hay endemicidad estable resultado de reinfecciones repetidas y, se puede atribuir lo dicho a factores ambientales responsables de la difusión y desarrollo de las formas infectantes (Lojano Collaguazo & Lojano Punin, 2018).

2.2.2.2. Tipo de parásito intestinal

Como se indicó anteriormente entre los parásitos intestinales existen dos familias que afectan a los seres humanos y se clasifican en Protozoos y Helmintos según la (Global Health, Division of Parasitic Diseases and Malaria, 2022)

Protozoos: Son organismos unicelulares microscópicos que pueden ser de vida libre o de naturaleza parasitaria, capaces de multiplicarse en el humano contribuyendo a su supervivencia y por ende permite el desarrollo de infecciones graves a partir de tan solo un organismo. La transmisión de protozoos que viven en el intestino humano a otro generalmente ocurre por vía fecal-oral, es decir, en alimentos o agua contaminados o el contacto de persona a persona. Estos parásitos viven en la sangre o en tejidos humanos y se trasmiten a otros seres humanos mediante un artrópodo vector (por ejemplo, la picadura de un mosquito o jején), puede clasificarse en cuatro grupos según su modo de movimiento:

- Sarcodinos, o amebas, como: Entamoeba
- Mastigóforos, o flagelados, como: Giardia, Leishmania
- Ciclióforos, o ciliados, como: Balantidium
- Esporozoos, organismos cuya etapa adulta no es móvil como: Plasmodium, Cryptosporidium.

Helmintos: Se deriva de la palabra griega "gusano" y son organismos grandes multicelulares que por lo general se observan a simple vista cuando son adultos. Al igual que el anterior, los helmintos pueden ser de vida libre o de naturaleza parasitaria. En la etapa adulta, los helmintos no pueden multiplicarse en los seres humanos y se clasifican en tres grupos importantes:

- Gusanos planos (platelmintos): incluyen los trematodos (duelas) y cestodos (tenias).
- Gusanos de cabeza espinosa (acantocéfalos): en la fase adulta estos gusanos se alojan en el tracto gastrointestinal y se cree que estos parásitos son una forma intermedia entre los cestodos y los nemátodos.

- Gusanos cilíndricos (nemátodos): en sus formas adultas los gusanos pueden alojarse en el tracto gastrointestinal, la sangre, el sistema linfático o tejidos subcutáneos por otra parte, en estados inmaduros (larvas) pueden provocar enfermedades por infección de diversos tejidos corporales.

Algunos consideran dentro de este grupo a los gusanos segmentados (anélidos) que desde el punto de vista médico son importantes las sanguijuelas a pesar de que estos usualmente no son considerados parásitos.

2.2.2.3. Tipo de hospedero

Se denomina huésped, hospedador u hospedante a aquel organismo que recibe o alberga a parásitos en su interior o lo porta sobre sí (Lema Punín & Inga Miguitama, 2018).

Según (Quimica.es, s.f.) y (Apt Baruch, 2013) de acuerdo con la utilidad para el parásito los hospederos o anfitriones son de tipo:

- Definitivo: se designa un ser vivo quien será imprescindible para el parásito ya que desarrollará su fase adulta en el anfitrión.
- Intermediario: es aquél en donde el parásito no alcanza la madurez sexual y alberga formas intermedias como: larvas o se multiplica asexualmente.
- Habitual: aquel hospedero donde el parásito desarrolla normalmente su ciclo de vida.
- Accidental: aquel organismo que circunstancialmente alberga un parasito.

2.2.2.4. Diagnóstico

(Lema Punín & Inga Miguitama, 2018) y (Lojano Collaguazo & Lojano Punin, 2018) mencionan que:

- El diagnóstico de infecciones por parásitos intestinales tiene como base los signos y síntomas que el paciente presenta.
- Se detecta mediante exámenes coproparasitarios por microscopia directa una técnica metodológica apropiada que permite la identificación de un gran número de parásitos como trofozoítos o quistes de protozoos y huevos o larvas de helmitos presentes en muestras fecales, orgánicas tomadas por un aspirado duodenal y biliar o biopsias. También, se emplea métodos serológicos para detección de anticuerpos, técnicas de, detección de coproantígenos usando anticuerpos o análisis isoenzimático y de biología molecular por ejemplo la reacción en cadena de la polimerasa (PCR) detectando genomas parasitarios.

2.2.3. Factor de riesgo

Según la OMS un factor de riesgo es cualquier rasgo, característica o exposición de un individuo que aumente su probabilidad de sufrir una enfermedad o lesión, es decir, cada una de las características o factores de naturaleza hormonal, genética, personal o ambiental que modifican las posibilidades de contraer una enfermedad.

Los factores de riesgo son aquellas características y atributos (variables) que están asociados diversamente con la enfermedad o el evento estudiado. En epidemiología, es toda circunstancia o situación que aumenta las probabilidades de una persona de contraer una enfermedad o cualquier otro problema de salud. Implican que las personas afectadas por dicho factor de riesgo presentan un riesgo sanitario mayor al de las personas sin este factor. Hay que diferenciar factores de riesgo de los factores de pronóstico, que son aquellos que predicen el curso de una enfermedad una vez que ya está presente. Existe también marcadores de riesgo que son características de la persona que no pueden modificarse (edad, sexo, estado socioeconómico). Existen factores de riesgo (edad, hipertensión arterial, raza, condiciones de trabajo, entre otras) que cuando aparece la enfermedad son a su vez factores pronóstico (mayor probabilidad de que se desarrolle un evento). Los factores de riesgo no son necesariamente las causas, sólo sucede que están asociadas con el evento de manera que incrementan el mismo. Como constituyen una probabilidad medible, tienen valor predictivo y pueden usarse con ventajas tanto en prevención individual como en la comunidad (Lojano Collaguazo & Lojano Punin, 2018).

2.2.3.1. Factores de riesgo de la parasitosis intestinal

Inadecuada higiene personal*:* La ausencia de higiene personal es condicionante al contraer enfermedades al organismo humano, los niños se encuentran vulnerables ante efectos negativos por lo que se debe resaltar la importancia de un adecuado aseo diario puesto que, al estar en un proceso de crecimiento, las actividades y el ambiente donde las realizan involucran tierra, sudor y factores que condicionan una acumulación de gérmenes. En lo que al hogar respecta el cuidado personal de los infantes es responsabilidad de quienes estén a cargo de su cuidado pues seguir las normas de higiene mantienen la salud del cuerpo y permite disfrutar la vida sanamente siempre enseñando a preservarse de agentes que pueden alterar la salud a través del mantenimiento de la integridad física, intelectual y psíquica previniendo enfermedades infectocontagiosas o su propagación.

Inadecuada higiene de los alimentos: En ocasiones por desconocimiento personas que residen en condiciones insalubres no mantienen una adecuada cultura de higiene en los alimentos por ende hay que hacer hincapié sobre el correcto lavado de frutas y vegetales además de la correcta

preparación de estos y su cocción, puesto que un tratamiento adecuado y con medidas higiénicas correctas permiten que el producto este en perfectas condiciones de seguridad. Las frutas y verduras están consideradas como alimentos propios para una alimentación saludable, no obstante, el consumo generalizado puede ocasionar un importante medio de contagios de origen infeccioso.

Inadecuado consumo de agua: Una gran parte de las infecciones por parásitos se contraen por la ingesta de agua que no tiene un tratamiento o no es el correcto para que el líquido sea apto al consumo humano. Una manera segura para beber agua y un método empleado por las familias es hervirla ya que se elimina la mayoría de las bacterias o microorganismos, pero no elimina suciedad e impurezas presentes lo que requiere de otros procesos como la destilación que separa los contaminantes del líquido (Reynolds, 2021).

Inadecuado manejo de basura: La acumulación de desechos y residuos es un problema diario, en lo que al hogar se refiere la basura doméstica principalmente contiene plásticos, cartones, papel, residuos de comida y entre otros que a su vez son acumulados en espacios al aire libre originando problemas en la higiene dando pie a la proliferación de virus y bacterias causantes de varias enfermedades como plagas, ratas, cucarachas e insectos perjudiciales a la salud además que en conjunto con la lluvia la acumulación de estos contaminan las aguas cuando es llevada hasta ríos, lagos y depósitos subterráneos de agua. Por otra parte, algunas veces la basura es eliminada a través de la incineración cuya técnica origina el desprendimiento de grandes cantidades de gases tóxicos que contaminan la atmósfera y que al ser depositados a cielo abierto los microorganismos producidos son transportados por el viento infectando aire, suelo, agua y los alimentos ya que gran parte de los residuos sólidos no son desagradables y se acumulan causando la perdida en la calidad y productividad de los suelos y el agua. La manipulación de los desechos sólidos se sintetiza en el ciclo donde a partir de la generación y acumulación temporal se continua con su recolección, transporte y trasferencia para finalizar con la acumulación total de estos y es en este punto donde el verdadero problema comienza ya que los basureros o centros de recolección se convierten en focos permanentes de contaminación.

Insuficiente educación: La carencia de conocimientos acerca de la transmisión de los parásitos y en general todo lo referente a la prevención de enfermedades es un hecho común en gran parte de los grupos poblacionales en América Latina, además que con los elevados porcentajes de analfabetismo en zonas rurales refleja la absoluta carencia de un mínimo nivel cultural o educativo.

2.2.3.2. Factores epidemiológicos asociados a la parasitosis intestinal

La complejidad de los factores que condicionan la parasitosis intestinal y la dificultad en su control determina que las infecciones por parásitos se encuentren ampliamente difundidas y su prevalencia sea similar en la actualidad.

(Lema Punín & Inga Miguitama, 2018) menciona los siguientes elementos:

Contaminación Fecal: Es un factor predominante en el parasitismo puesto que la contaminación de la tierra o del agua se da en sitios con déficit de saneamiento. En el suelo los elementos patógenos pueden ingresar al organismo a través de una defecación directa por ausencia de excretas, empleo de heces como abono, utilizar aguas cloacales como sistema de riego, contaminación con basura patológica o deposición de animales.

Condiciones ambientales: La temperatura adecuada, humedad, clima cálido, vegetación, precipitaciones favorecen el desarrollo y sobrevivencia de parásitos. Las viviendas cuya construcción es con adobe de barro da pie al ingreso de artrópodos y la presencia de aguas estancadas permite el desarrollo del vector, el cual ayuda a la diseminación de la enfermedad parasitaria.

Hacinamiento: El gregarismo de ciertas comunidades por ejemplo campos de concentración, refugios, escuelas contribuyen a la aparición de enfermedades por parásitos debido a la ausencia o perdida de costumbres higiénicas y alimenticias, existencia de limitación al uso de baños etc. En las guarderías los parásitos más frecuentes son la giardiosis y oxiuriosis.

Costumbres alimenticias*:* El consumo de carnes crudas o mal cocidas produce infecciones por Tenias, Toxoplasma y Trinchinella; por otra parte, la cestodiasis, trematodiasis son producto de la ingesta de carne cruda de pescado, crustáceos (cangrejos, langostas) y otros mariscos.

2.2.4. Medidas de Prevención

- Lavarse las manos antes y después de ingerir alimentos, después de salir del baño o las veces que se considere necesario.
- Consumir agua potable, en caso de la ausencia del servicio es recomendable verter 2 gotas de cloro por cada litro de agua y hervirla en un período de 3 a 5 minutos
- Lavar y desinfectar frutas y verduras antes de consumirlas
- Mantener las uñas cortas y evitar onicofagia
- Evitar el consumo de comida en la calle o en sitios con condiciones higiénicas deficientes
- Implementar una buena estructura para la disposición de excrementos

CAPITULO III

3. MARCO METODOLÓGICO

3.1. Tipo de investigación

Por el método de investigación el presente proyecto de investigación es de tipo cuantitativo, debido a que se analizó el número de pacientes diagnosticados con parasitosis intestinal en el HPAVR, según el objeto de estudio es aplicado porque se utilizó diferentes análisis para comparar cuál de los dos métodos es mejor para realizar pronósticos, según el nivel de profundización en el objeto de estudio es exploratorio e inferencial ya que este estudio no se lo ha realizado antes, según la manipulación de las variables es no experimental pues los datos se obtuvieron mediante las historias clínicas de los pacientes pediátricos, según la inferencia es inductiva puesto que se analiza las metodologías de árboles de clasificación y regresión para especificar que método es mejor y según el periodo temporal es longitudinal donde se analizó pacientes pediátricos internados en diferentes años calendario.

3.2. Diseño de investigación

Se utiliza el método de investigación cuantitativo y según la manipulación de variables es un diseño no experimental, en vista que en el trascurso del desarrollo del proyecto de investigación se trabajó con diferentes algoritmos y técnicas para la clasificación de individuos, específicamente los árboles de clasificación y regresión logística donde a través de estos se obtiene un modelo matemático que mediante las medidas de bondad de ajuste se determina el "mejor" a aquel cuya capacidad predictiva es mayor utilizando la tasa de error entre ellos mediante la matriz de confusión y el área bajo la curva ROC a través del AUC.

3.2.1. Localización del Estudio

La investigación planteada se llevó a cabo en el Hospital Pediátrico Alfonso Villagómez Román en la ciudad de Riobamba en las calles España entre José Orozco y Av. José Veloz.

3.2.2. Población de estudio

La población de estudio pertenece a los pacientes atendidos en el periodo 2019-2021 en el Hospital Pediátrico Alfonso Villagómez Román.

3.2.3. Tamaño de la muestra

La muestra pertenece a todos los pacientes de 5 a 9 años diagnosticados con parasitosis intestinal en el Hospital Pediátrico Alfonso Villagómez Román.

3.2.4. Método de muestreo

No se aplica un método de muestreo ya que los datos fueron proporcionados por la casa de salud a través de sus repositorios, además se revisó las historias clínicas de los pacientes.

3.2.5. Técnicas de recolección de datos

No se aplicó una técnica de recolección de datos ya que la información con los pacientes registrados en la entidad de salud fue entregada por el área de estadística de la institución.

3.2.6. Modelo Estadístico

La comparativa entre modelos de regresión y árboles de clasificación estará sujeta a la revisión previa del historial clínico de los pacientes que asisten a la casa de salud, de quienes obtendremos los datos para las variables independientes, las que serán utilizadas para modelar la parasitosis intestinal a través de las técnicas planteadas.

3.3. Identificación de variables

Variable dependiente

- Diagnóstico de alta (Parasitosis intestinal y Otro caso)

Variables independientes

- Cantón
- Grupo Cultural
- Edad
- Género
- Frecuencia cardiaca mínima
- Frecuencia respiratoria mínima
- Triage
- Temperatura axilar (°C)
- Peso (Kg)
- Talla (cm)
- Saturación de oxigeno
- Tipo de seguro

3.4. Operacionalización de variables

Tabla 1-3: Descripción de variables

Variable	Descripción	Tipo	Escala de medición
Diagnóstico de alta	Estada de salud con el que egresa el paciente del hospital (Parasitosis intestinal u Otro caso)	Cualitativa Dicotómica	Nominal
Cantón	Unidad de división administrativa y territorial de algunos países; puede constituir el primer nivel de división, como ocurre en algunos estados federales, o estar por debajo de entidades mayores, como provincias, departamentos, etc.	Cualitativa Politómica	Nominal
Grupo cultural	Registra la autoidentificación étnica (Afroecuatoriano, blanco, indígena, mestizo, montubio, mulato, negro, otro)	Cualitativa Politómica	Nominal
Edad	Edad en años cumplidos (5-9 años)	Cuantitativa Discreta	Razón
Género	Registra la información correspondiente al sexo-biológico (Femenino, Masculino)	Cualitativa Dicotómica	Nominal
Frecuencia cardiaca mínima	Número de contracciones del corazón o de pulsaciones por unidad de tiempo	Cuantitativa Discreta	Razón
Frecuencia respiratoria mínima	Número de respiraciones que realiza un ser vivo en un periodo especifico	Cuantitativa Discreta	Razón
Triaje	Clasifica al paciente de acuerdo con las necesidades terapéuticas y los recursos disponibles	Cualitativa Politómica	Ordinal

Temperatura axilar °C	Registra la temperatura tomada bajo la axila del paciente	Cuantitativa Continua	Razón
Peso (Kg)	Registra el peso o la masa del paciente	Cuantitativa Continua	Razón
Talla (m)	Registra la estatura del paciente	Cuantitativa Continua	Razón
Saturación de oxigeno	Registra la cantidad de oxígeno disponible en la sangre	Cuantitativa Continua	Razón
Tipo de seguro de salud	Registra el seguro de salud del paciente (IESS, ISSPOL, ISSFA, otro ninguno)	Cualitativa Politómica	Nominal

Fuente: Elaboración propia

Realizado por: Isin Wendy, López Maicol, 2022

CAPITULO IV

4. RESULTADOS Y DISCUSIÓN

El análisis estadístico toma en consideración un colectivo de 2675 individuos o pacientes pediátricos de 5 a 9 años que fueron atendidos en el área de emergencia del Hospital Pediátrico Alfonso Villagómez Román (HPAVR) de la ciudad de Riobamba en el periodo 2019 – 2021, la matriz de información contiene 14 variables, 7 de tipo cuantitativo y 7 de tipo cualitativo.

4.1. Análisis Exploratorio de Datos

Para el presente trabajo de investigación en primer lugar se realizó un análisis descriptivo de las variables mediante representaciones gráficas de las frecuencias de cada una de ellas, así como también un análisis estadístico básico para las variables cuantitativas seleccionadas. Los resultados encontrados se detallan a continuación:

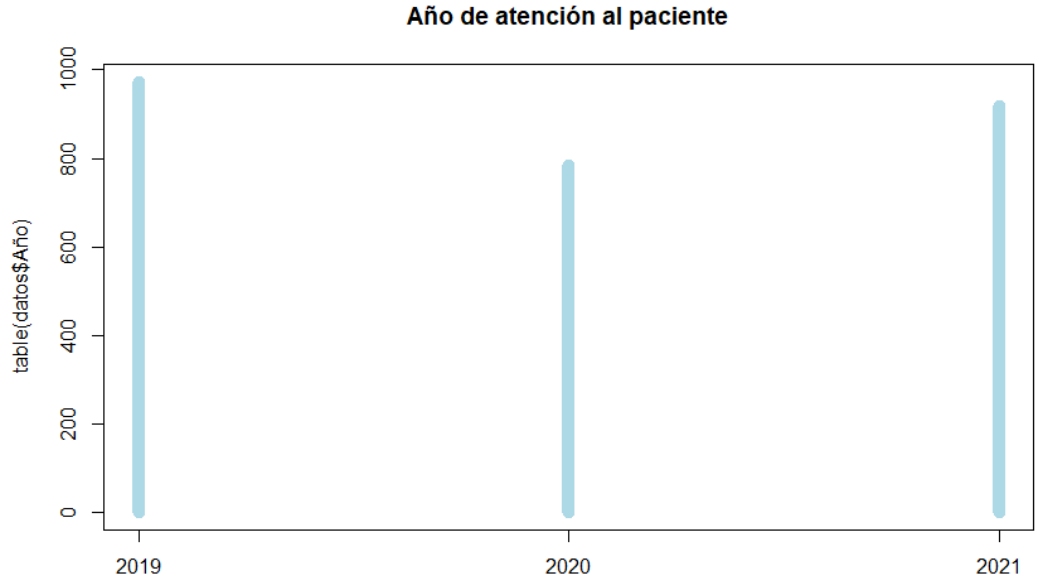


Gráfico 1-4: Distribución de la variable "Año"

Realizado por: Isin W., López M. 2022

En la gráfica se representa la afluencia de pacientes que ingresaron al hospital en los 3 años de estudio donde, observamos que el año con mayor ingreso fue en el 2019, esta variación podemos decir que se origina a raíz de la pandemia por COVID-19 que afecto a la población y provoco una crisis sanitaria en el sistema de salud. En 2020 y 2021 se observa un descenso y un leve incremento en los ingresos hospitalarios respectivamente.

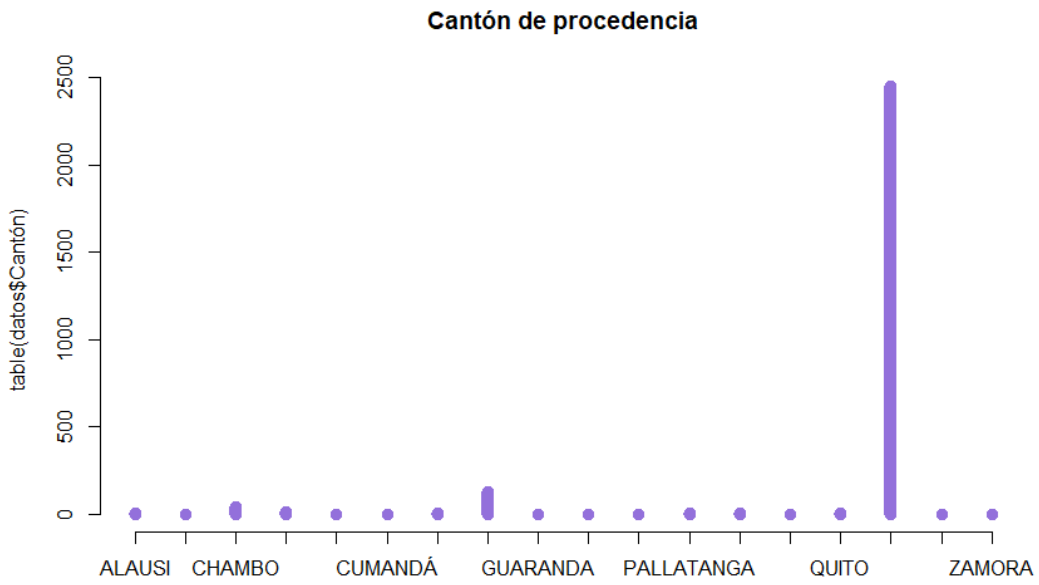


Gráfico 2-4: Distribución de la variable "Cantón"

Realizado por: Isin W., López M. 2022

En lo que respecta al cantón de residencia según el gráfico indica que la mayoría de los pacientes que acudieron a la casa de salud son locales es decir oriundos del cantón Riobamba, sin embargo, también se registra pacientes de otros cantones del país en menor proporción, esto último puede estar ligado a que al ser un hospital de tercer nivel y específicamente para pacientes pediátricos quienes llegan al lugar posiblemente son transferidos de otra entidad de salud en su cantón natal.

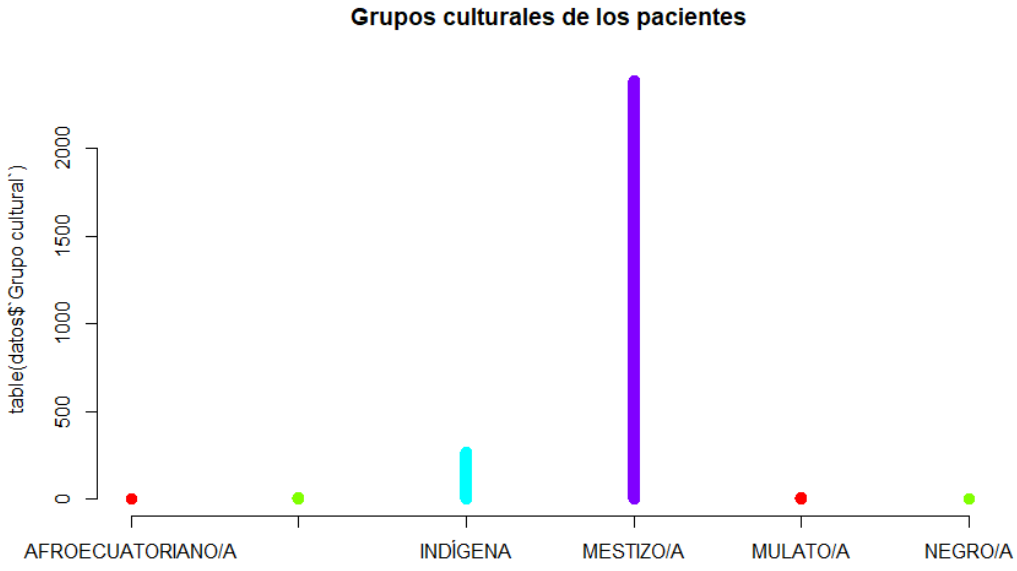


Gráfico 3-4: Distribución de la variable "Grupo cultural"

Realizado por: Isin W., López M. 2022

Con respecto al grupo cultural con el cual se identifican los pacientes según el gráfico hay un gran grupo de individuos que se identifican como Mestizo/a, seguido de aquellos que se identifican como Indígenas y en menor proporción se encuentran los grupos blanco, mulato y negro.

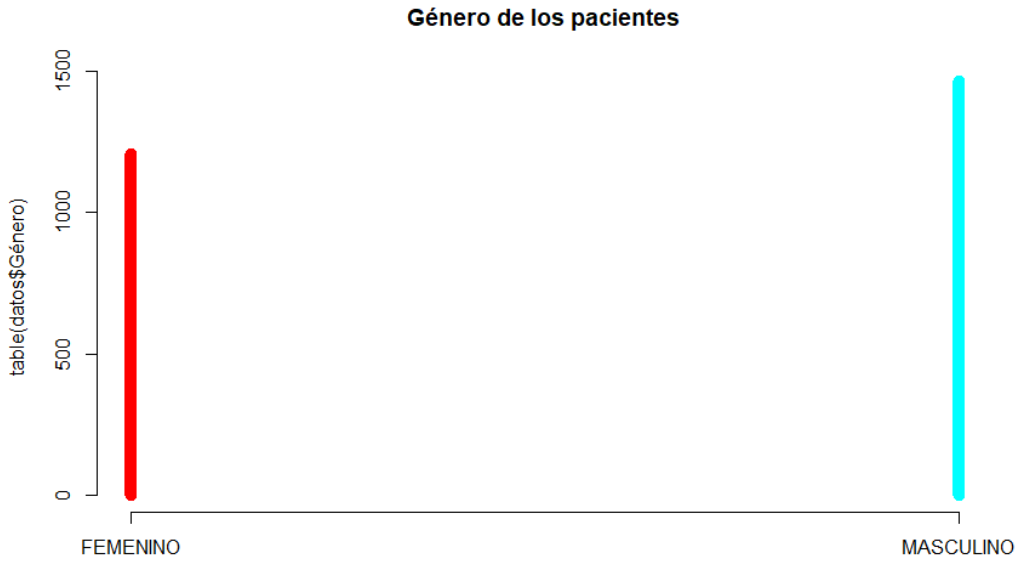


Gráfico 4-4-: Distribución de la variable "Género"

Realizado por: Isin W., López M. 2022

Según los registros de la casa de salud en los tres años de estudio en el hospital se registró un gran conglomerado de pacientes masculinos y en leve proporción pacientes femeninas.

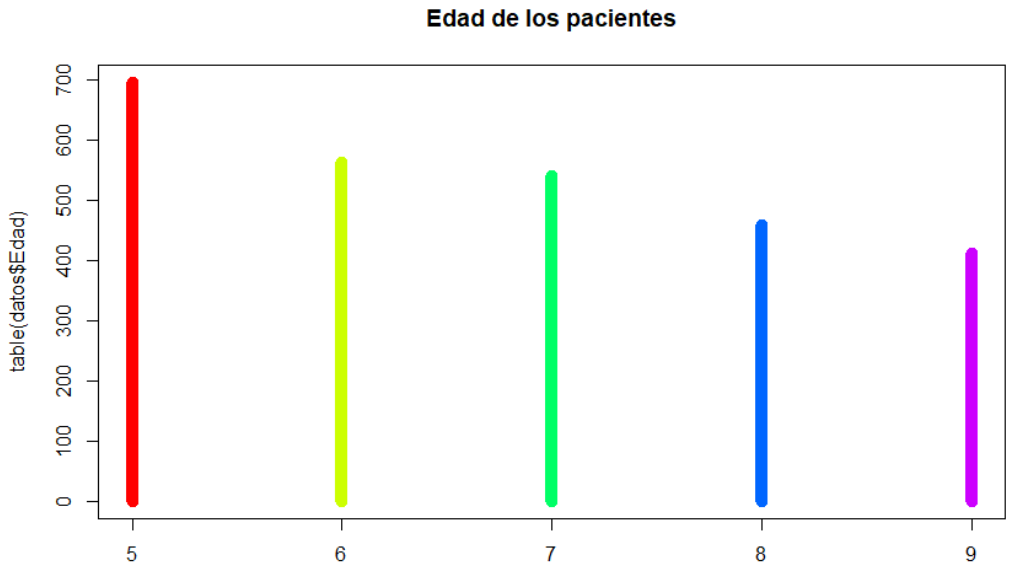


Gráfico 5-4: Distribución de la variable "Edad"

Realizado por: Isin W., López M. 2022

Con respecto a la edad cabe recalcar que para el estudio se limitó a pacientes pediátricos entre 5-9 años por lo tanto podemos observar que la edad con mayor frecuencia es 5 años, además observamos que el resto de las edades presenta un descenso en sus proporciones.

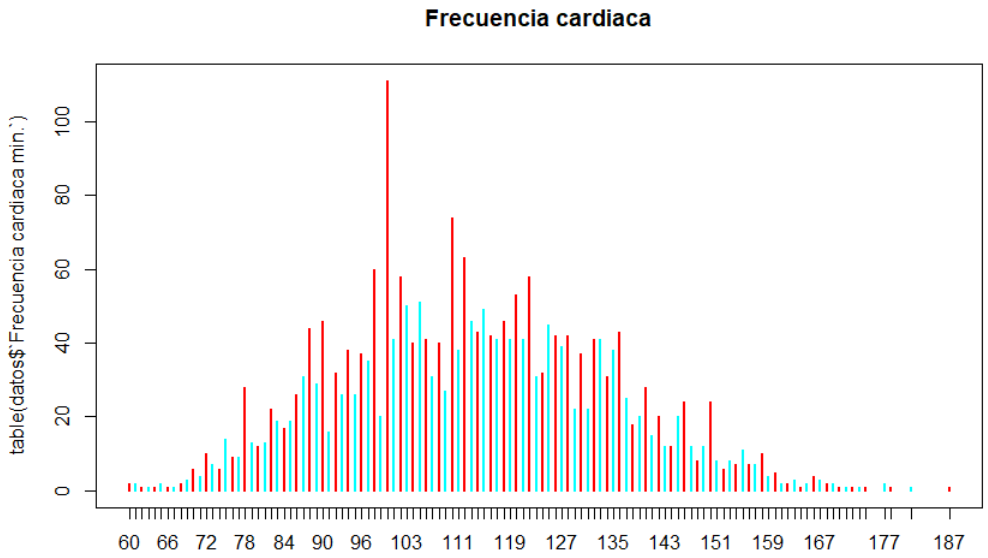


Gráfico 6-4: Distribución de la variable "Frecuencia cardiaca"

Realizado por: Isin W., López M. 2022

La frecuencia cardiaca registra el número de pulsaciones del paciente por unidad de tiempo, en el gráfico podemos observar que los valores se encuentran entre 60 y 187 pulsaciones por unidad de tiempo siendo 100 el valor que mayor frecuencia se presenta en los pacientes. Además, cabe indicar que algunos valores estarán por encima o debajo de los rangos normales de frecuencia cardiaca en los niños de 5-9 años y esta variación puede deberse a otros factores o condiciones en las que ingreso el paciente al hospital.

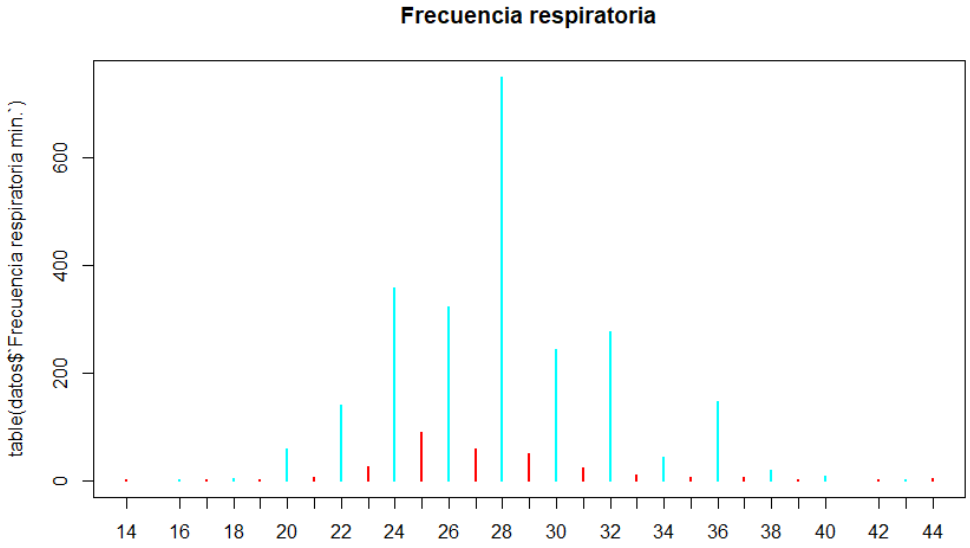


Gráfico 7-4: Distribución de la variable "Frecuencia respiratoria"

Realizado por: Isin W., López M. 2022

La frecuencia de respiración mínima contabiliza el número de respiraciones del paciente en un periodo específico, respecto a los registros del hospital y como se observa en el gráfico la variable se encuentra en un intervalo de 14 a 44 respiraciones, y hay una gran proporción de niños entre 5-9 años cuya frecuencia respiratoria es de 28 mientras que los demás se encuentran por encima o debajo de este valor y, como se indicó anteriormente los valores registrados varían con respecto a los valores normales para estas edades debido a factores que inciden para que esta se eleve o disminuya respectivamente.

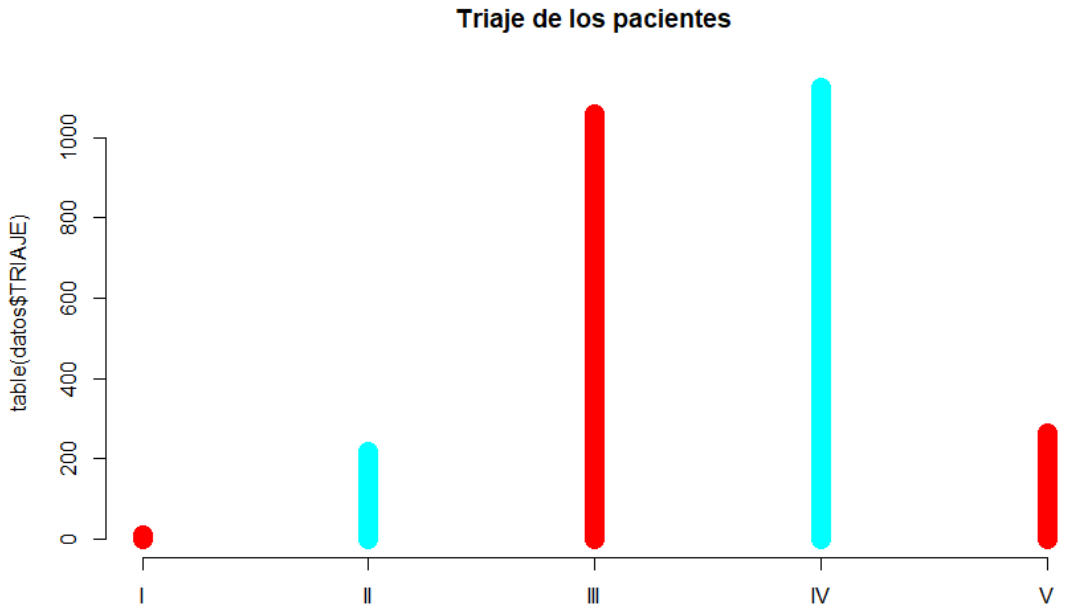


Gráfico 8-4: Distribución de la variable "Triage"

Realizado por: Isin W., López M. 2022

El Triage hace referencia al nivel de gravedad con la que un paciente ingresa al servicio de emergencia del hospital pediátrico, siendo el Nivel I el más crítico y el Nivel V el menos urgente a ser atendido, para el presente estudio se obtuvo que los pacientes ingresaban en su mayoría con un nivel de emergencia III y IV, a pesar que se presentaron pacientes con un Nivel I de emergencia, fueron pocos, lo cual es gratificante ya que la vida de la mayoría de pacientes no se encontraba en un riesgo crítico.

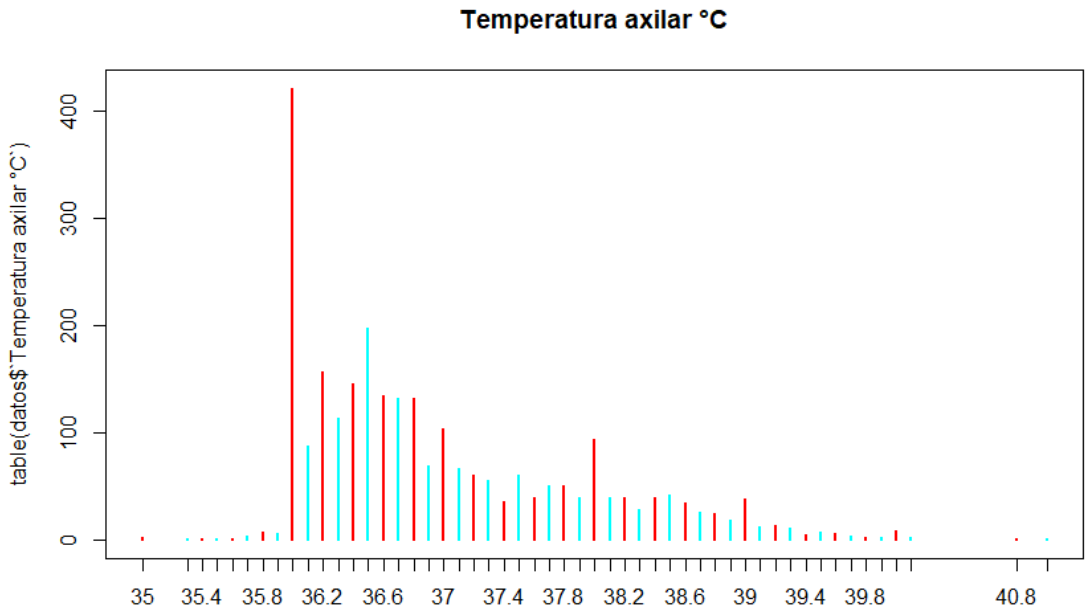


Gráfico 9-4: Distribución de la variable "Temperatura axilar"

Realizado por: Isin W., López M. 2022

En el análisis de la temperatura axilar se puede observar que el mayor número de pacientes presentan una temperatura entre los 36 °C, valor considerado dentro de los parámetros normales, pero existen casos que nos llegaron a preocupar debido a que tenían temperaturas de 35 °C y 41 °C estando estas consideradas dentro de los parámetros de hipotermia e hiperpirexia (fiebre) respectivamente, dándonos a entender que los pacientes acudieron a la casa de salud en condiciones críticas de salud.

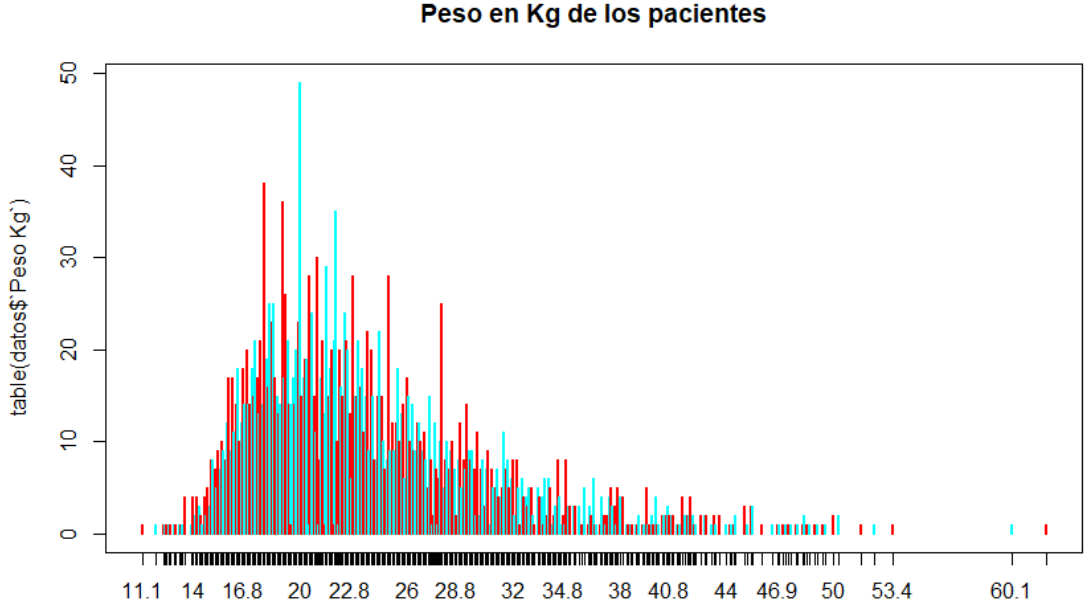


Gráfico 10-4: Distribución de la variable "Peso"

Realizado por: Isin W., López M. 2022

Con respecto al peso, podemos decir que la media es equivalente a 23.91 Kg dándonos a entender que la gran mayoría de pacientes registrados mantienen un peso adecuado para la edad que presentan, lo que si llama la atención es que existieron pacientes con tendencia a la desnutrición, ya que sus pesos fueron un tanto bajos con valores mínimos de 11.1 Kg.

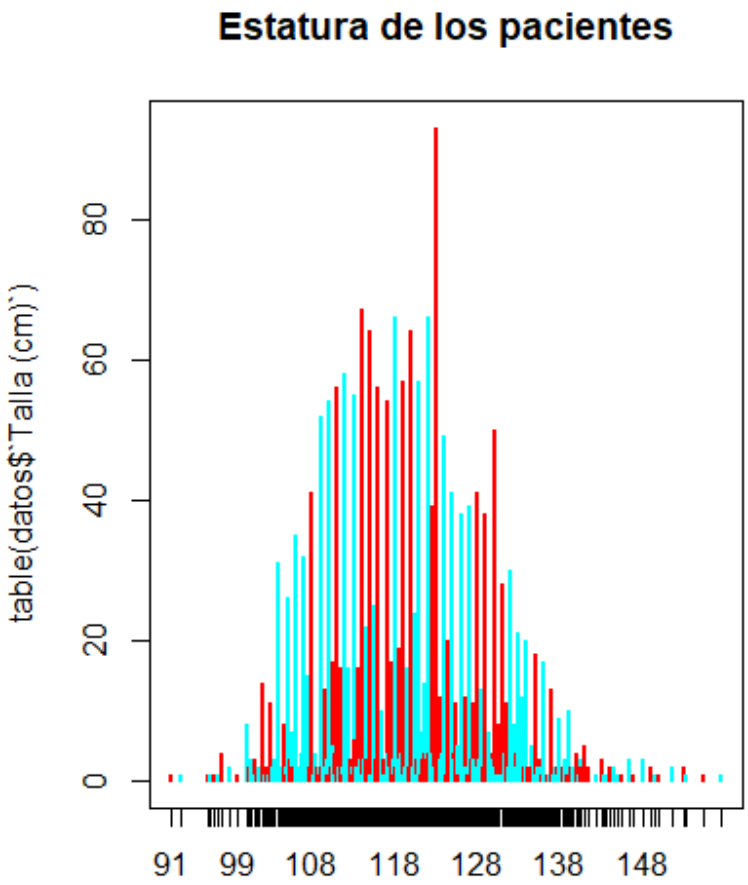


Gráfico 11-4: Distribución de la variable "Talla"

Realizado por: Isin W., López M. 2022

Con respecto al análisis de la estatura o talla, podemos decir que la media de los pacientes estudiados ronda en una media de 119.8 cm de estatura, esto puede ser debido a que la estatura promedio de los ecuatorianos no es muy alta, pero se pudo registrar que la talla más baja es equivalente a 91.0 cm dándonos a entender que se trata de un paciente de 5 años, siendo la edad más corta establecida para este estudio.

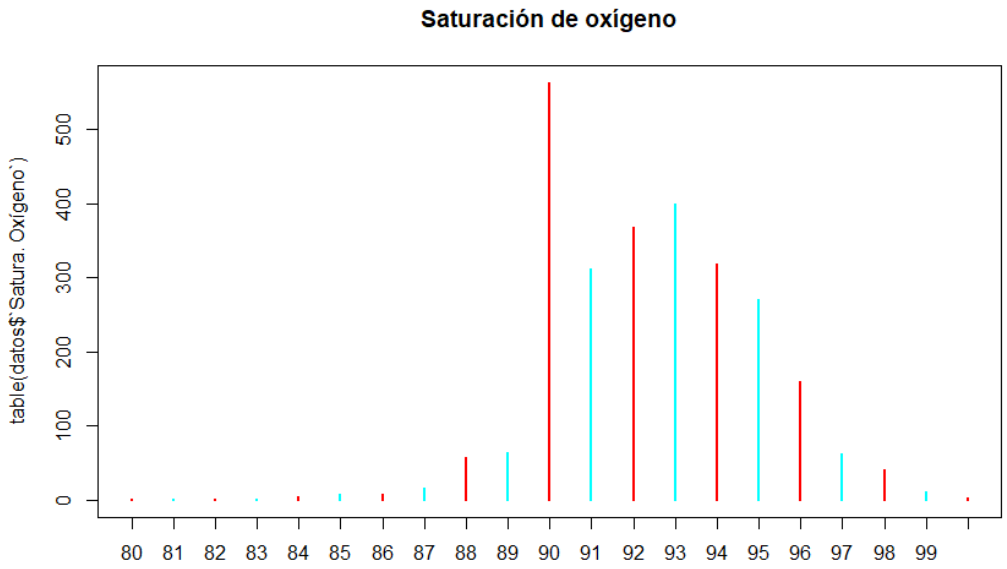


Gráfico 12-4: Distribución de la variable "Saturación de oxígeno"

Realizado por: Isin W., López M. 2022

La saturación de oxígeno hace referencia a las cantidades de oxígeno que circulan por el torrente sanguíneo, siendo así, un valor de 80 se encuentra fuera de los parámetros normales indicando que a los pacientes se les debió otorgar servicios de salud de manera urgente y prioritaria para salvaguardad sus vidas.

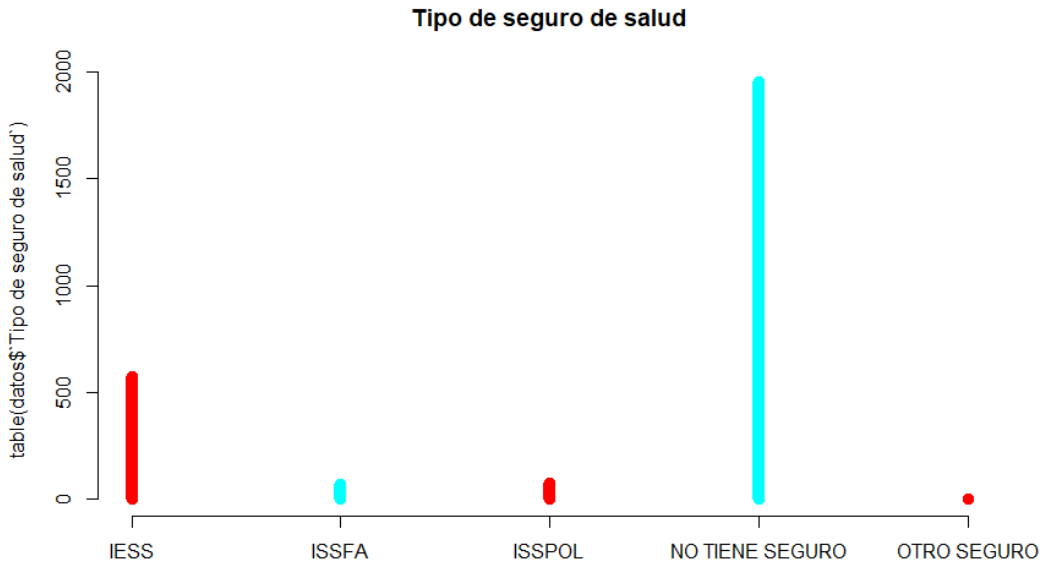


Gráfico 13-4: Distribución de la variable "Peso"

Realizado por: Isin W., López M. 2022

Gracias a este análisis pudimos darnos cuenta que la gran mayoría de pacientes que acuden a esta casa de salud no poseen un seguro social, por lo cual los gastos de salud se deben cubrir por parte del estado, esto puede deberse a la situación tan crítica en la que se encuentra nuestro país

actualmente en el ámbito económico, ya que los ecuatorianos, y en este estudio, los riobambeños para obtener un seguro de salud deben hacer aportaciones mensuales, las cuales solo una minoría de pacientes contaba con algún seguro público o privado.

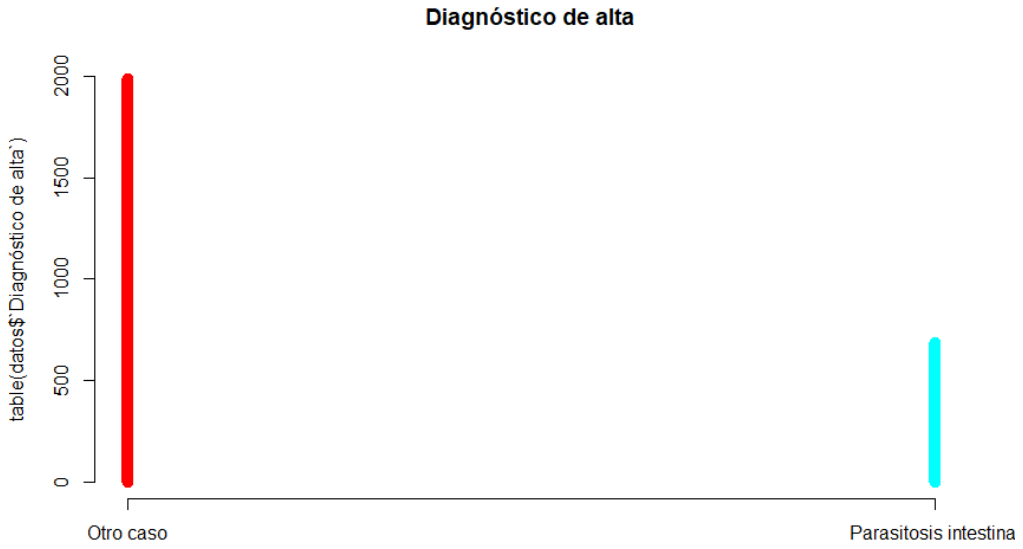


Gráfico 14-4: Distribución de la variable "Diagnóstico de alta"

Realizado por: Isin W., López M. 2022

Para el presente estudio podemos decir que, en la ciudad de Riobamba y sus alrededores, presentan casos de parasitosis intestinal, y al estar esta casa de salud al servicio para transferencia de pacientes en estado crítico, asumimos que los casos atendidos de parasitosis fueron de carácter severo, lo cual puede estar atentando seriamente la salud de los niños e inclusive atentando contra la vida de estos.

4.2. Técnicas de Modelado

4.2.1. Modelo de clasificación: Regresión Logística Binaria

La clasificación mediante regresión logística tiene como objetivo estimar, clasificar e identificar las variables influyentes en la parasitosis intestinal; para la construcción del modelo mediante esta técnica se utilizó todo el conjunto de estudio.

La selección de las variables para el modelo se lo realizó con el método de escalonado hacia adelante que inicia solo con el término constante y va agregando al mismo modelo aquellas variables independientes cuyo nivel de asociación sea significativo estadísticamente respecto a la variable dependiente, obteniendo el modelo que a continuación se presenta:

		B	Error estándar	Wald	gl	Sig.	Exp(B)
Paso 1[a]	Triage	-,254	,056	20,439	1	,000	,776
	Constante	-,179	,199	,806	1	,369	,836
Paso 2[b]	Triage	-,298	,057	27,207	1	,000	,742
	Satur.oxig	,091	,018	24,580	1	,000	1,096
	Constante	-8,471	1,686	25,243	1	,000	,000
Paso 3[c]	Triage	-,354	,063	31,265	1	,000	,702
	Temp.axilar	-,112	,055	4,157	1	,041	,894
	Satur.oxig	,082	,019	18,775	1	,000	1,085
	Constante	-3,257	3,052	1,139	1	,286	,038
Paso 4[d]	Frec.car.min	,008	,003	10,237	1	,001	1,008
	Triage	-,347	,064	29,814	1	,000	,707
	Temp.axilar	-,198	,062	10,401	1	,001	,820
	Satur.oxig	,090	,019	22,049	1	,000	1,094
	Constante	-1,758	3,090	,324	1	,569	,172
Paso 5[e]	Frec.car.min	,010	,003	13,565	1	,000	1,010
	Frec.resp.min	-,029	,012	5,588	1	,018	,971
	Triage	-,362	,064	32,096	1	,000	,696
	Temp.axilar	-,188	,062	9,300	1	,002	,828
	Satur.oxig	,085	,019	19,426	1	,000	1,089
	Constante	-,993	3,117	,101	1	,750	,370

Figura 1-4: Variables del Modelo de Regresión por el método de pasos hacia adelante

Fuente: IBM SPSS STATISTICS

En el cuadro presentando anteriormente se observa que las variables en estudio: Triage, Temperatura axilar, Saturación de oxígeno, Frecuencia respiratoria mínima y Frecuencia cardíaca mínima son significativas $p - valor < 0.05$ por lo tanto están incluidas en el modelo y concluimos que influyen sobre el diagnóstico de los pacientes atendidos en el HPAVR con un nivel de confianza del 95%. El modelo obtenido mediante regresión logística es el siguiente:

$$p_i = \frac{1}{1 + e^z}$$

Donde

Z = -0,993 + 0,85(Saturación oxígeno) - 0,188(Temperatura axilar) - 0,362(Triage) - 0,29(Frecuencia respiratoria mínima) + 0,01(Frecuencia cardíaca mínima)

Del modelo presentado se observa que el signo de los coeficientes en algunas variables es positivo y significa que la variable aumenta la probabilidad del suceso estudiado, o lo que es lo mismo decir que aumenta la probabilidad de que un paciente sea diagnosticado con parasitosis intestinal.

4.2.1.1. Prueba ómnibus para la significancia del modelo

$$\boldsymbol{H_0} = \textit{El modelo no es significativo}$$
$$\boldsymbol{H_1} = \textit{El modelo es significativo}$$

		Chi-cuadrado	gl	Sig.
Paso 1	Escalón	20,609	1	,000
	Bloque	20,609	1	,000
	Modelo	20,609	1	,000
Paso 2	Escalón	24,921	1	,000
	Bloque	45,530	2	,000
	Modelo	45,530	2	,000
Paso 3	Escalón	4,196	1	,041
	Bloque	49,726	3	,000
	Modelo	49,726	3	,000
Paso 4	Escalón	10,283	1	,001
	Bloque	60,009	4	,000
	Modelo	60,009	4	,000
Paso 5	Escalón	5,646	1	,017
	Bloque	65,656	5	,000
	Modelo	65,656	5	,000

Figura 2-4: Prueba ómnibus de coeficientes del modelo

Fuente: IBM SPSS STATISTICS

El contraste sobre el nivel de significancia global del modelo se contrastó con el estadístico Razón de Verosimilitud (Prueba Ómnibus). El cuadro anterior nos muestra un valor de Chi-cuadrado de 65,656 con un p-valor (0.00) menor al nivel establecido 0.05, por lo que concluimos que hay una relación significativa entre las variables independientes y su predictora entonces el modelo es significativo.

4.2.1.2. Tabla de clasificación de Regresión Logística

El modelo estimado a través de regresión logística binaria presento un error de predicción de 25,607%, la matriz de confusión asociada al modelo es la siguiente:

Tabla 1-4: Matriz de Confusión regresión

		Diagnóstico	
		Otro caso	Parasitismo
Diagnóstico	Otro caso	1991	0
	Parasitismo	684	0

Fuente: IBM SPSS MODELER

Realizado por: Isin Wendy, López Maicol, 2022

4.2.2. Modelo de clasificación: Árboles de Clasificación

La selección de las variables para el modelo por árboles de clasificación se utilizó el método de crecimiento CHAID, con un número de casos mínimo aceptable en un nodo padre de 100 y nodo hijo de 50 con profundidad 4 obtuvimos el siguiente resultado.

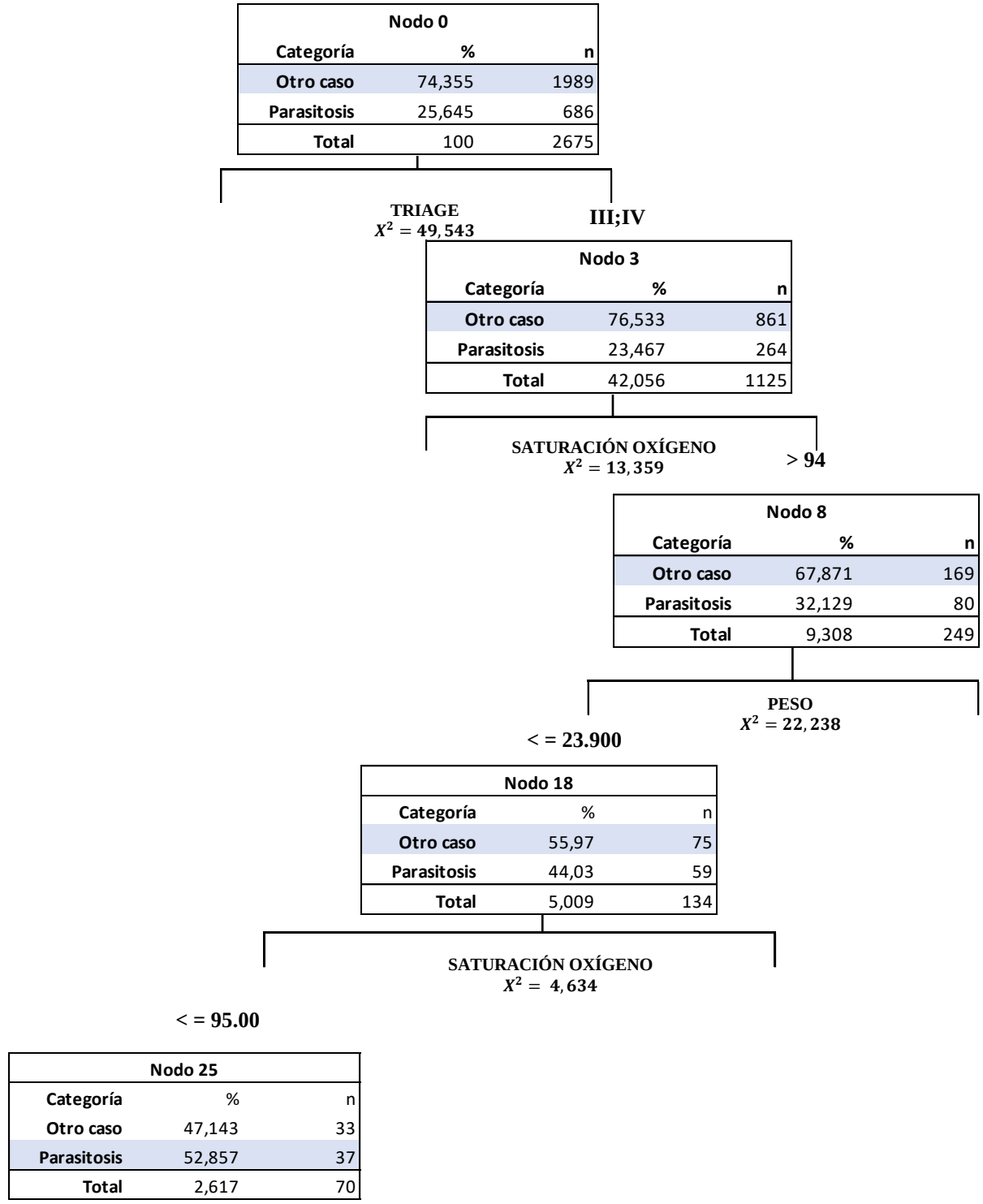


Gráfico 15-4: Árbol de clasificación

Realizado por: Isin W., López M. 2022

El nodo 0 describe la variable dependiente: diagnóstico de alta, especificando el porcentaje de los pacientes registrados con parasitosis intestinal u otra enfermedad.

a) La variable dependiente se ramifica en 4 nodos derivados de la variable "Triage" considerando esta como variable principal predictora.
b) Fijamos atención en el nodo 2 puesto el valor de Chi-cuadrado es mayor a los otros, como nuestro objetivo de interés es conocer los factores que modelan la parasitosis intestinal nos concentramos en los nodos cuya ramificación se extienda hasta alcanzar el objeto de estudio.
c) Los nodos 2, 3, 4 ramifican a los nodos 5,6,7,8,9,10 según la temperatura axilar, saturación de oxígeno y frecuencia cardiaca mínima respectivamente.
d) Los nodos ramificados anteriormente se vuelven a abrir y se incluyen las variables Talla, Peso y Tipo de seguro de salud, llegando al penúltimo nivel del árbol las clasificaciones se han realizado adecuadamente para el diagnóstico 0 = "otro caso", pues si observamos la selección de los grupos con respecto a los porcentajes a una profundidad de 3 la mayoría de los pacientes registro otra enfermedad diferente a la parasitosis intestinal
e) A una profundidad de 4 el nodo terminal 25 partiendo del nodo 3 con saturación de oxígeno mayor a 94 donde el 67.87% no son pacientes con parasitosis, este nodo se ramifica de acuerdo con el Peso el nodo 18 agrupando el 55.97% de pacientes con otro diagnostico si el peso es inferior a 23.9. Continuando la ruta para pacientes con saturación de oxígeno menor o igual a 95 se visualiza el objetivo y en este nodo el 52,86% está clasificado con diagnóstico de parasitosis intestinal.
f) Finalmente, en síntesis, los nodos que definen el diagnóstico de los pacientes con parasitosis intestinal son: Frecuencia cardíaca mínima, Triage, Temperatura axilar, Peso, Talla, Saturación de oxígeno, Tipo seguro de salud.

4.2.2.1. Tabla de clasificación: Algoritmo CHAID

El modelo estimado a través de regresión logística binaria presento un error de predicción de 24,495%, la matriz de confusión asociada al modelo es la siguiente:

Tabla 2-4: Matriz de confusión árbol clasificación

		Diagnóstico	
		Otro caso	Parasitismo
Diagnóstico	Otro caso	1993	0
	Parasitismo	682	0

Fuente: IBM SPSS MODELER
Realizado por: Isin Wendy, López Maicol, 2022

4.3. Evaluación de las técnicas de clasificación

4.3.1. Comparación de los modelos: Regresión Logística y Árboles de Clasificación

Considerando la tasa de error mediante la matriz de confusión de cada modelo encontrado, (Tabla 4-1, Tabla 4-2) se analizó la capacidad predictiva de estos mediante el AUC de la curva ROC y sus graficas respectivas, a continuación, se presenta los resultados:

Tabla 3-4: Áreas bajo la curva (AUC)

	Área bajo la curva (AUC)
Regresión Logística	62,8%
Árbol de clasificación	65,7%

Fuente: IBM SPSS MODELER

Realizado por: Isin Wendy, López Maicol, 2022

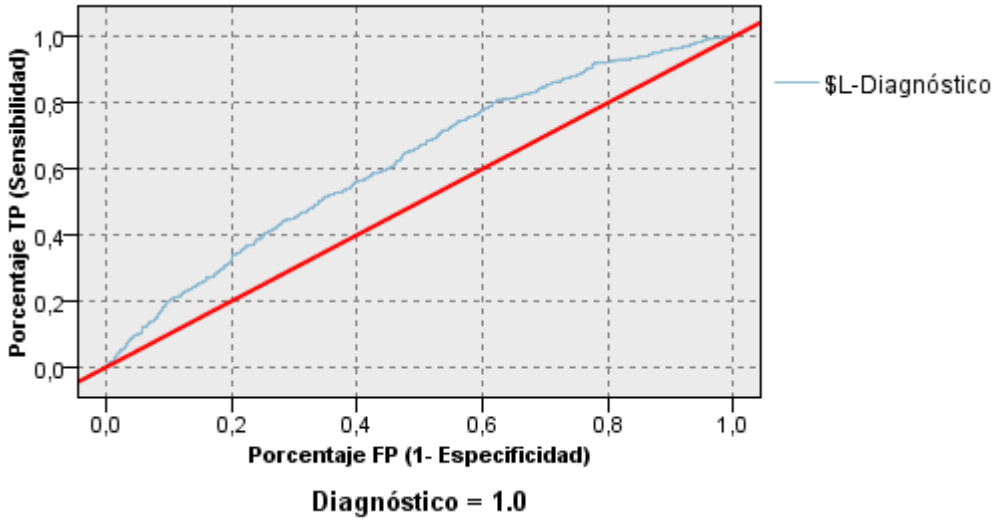


Figura 3-4: Curva ROC: Regresión Logística

Fuente: IBM SPSS MODELER

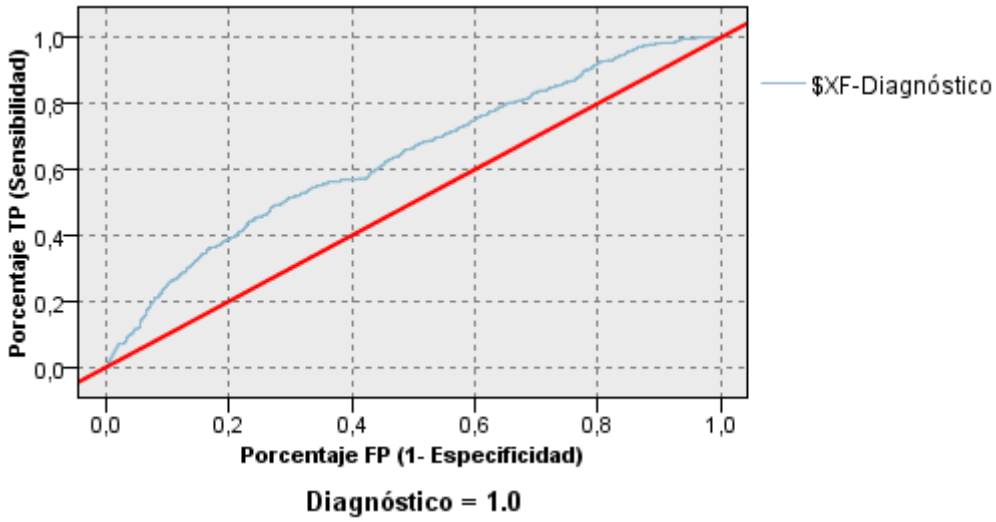


Figura 4-4: Curva ROC: Árbol de Clasificación

Fuente: IBM SPSS MODELER

Una vez analizando las representaciones gráficas de la curva ROC y el área bajo la curva (AUC) de los modelos encontrados, se observa que el modelo encontrado mediante Árbol de clasificación presenta un mayor AUC = 65,7% por lo tanto tiene mayor capacidad predictiva a comparación de la regresión logística cuyo AUC = 62,8%, y esto tiene sentido ya que al comparar con la matriz de confusión según la tasa del error, los árboles de clasificación, clasificaría el diagnóstico de los pacientes con un error mínimo.

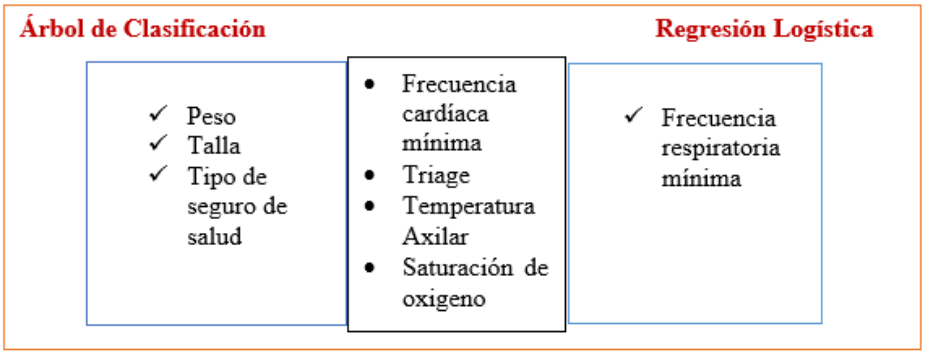


Gráfico 16-4: Comparativa entre los factores asociados a la parasitosis intestinal mediante RL y AC

Realizado por: Isin W., López M. 2022

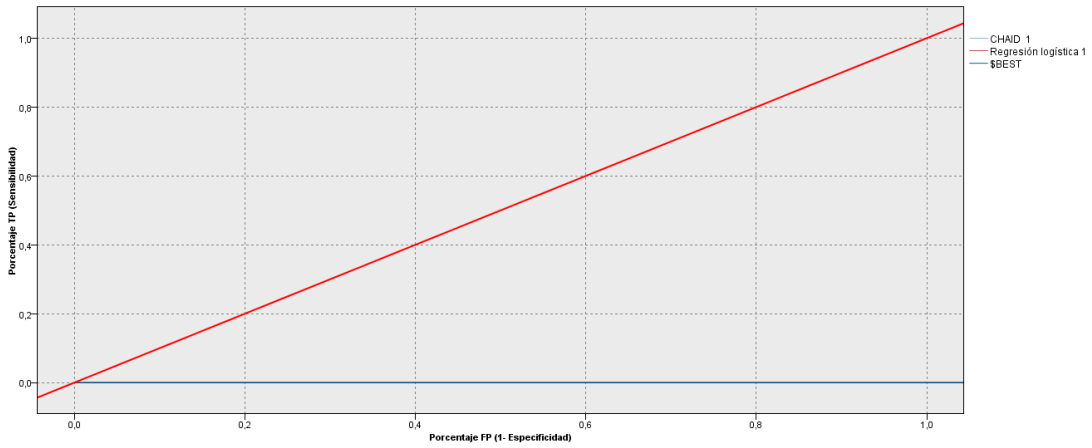


Figura 5-4: Curva ROC: Regresión Logística vs. Árbol de Clasificación

Fuente: IBM SPSS MODELER

CONCLUSIONES

- Gracias al presente estudio, conforme a los datos obtenidos, se demostró que los árboles de clasificación, con un área bajo la curva ROC equivalente al 65.7%, son más eficientes que los modelos de regresión logística, los cuales presentan un área bajo la curva ROC del 62.8%.
- Los posibles factores asociados a la parasitosis intestinal fueron determinados mediante un análisis teórico dándonos como resultado los siguientes: Temperatura axilar, peso, talla, edad, saturación de oxígeno, presión arterial, grupo cultural, frecuencia respiratoria, frecuencia cardíaca, y las variables anexadas al estudio fueron determinas como posibles factores influyentes en el sector de estudio.
- El análisis de las historias clínicas brindó las variables para este estudio, y a su vez permitió consolidar la base de datos que fue utilizada, aplicando las respectivas restricciones como el grupo etario, análisis discriminante de datos para poder obviar a los pacientes con signos vitales erróneamente establecidos, etc. mejorando la calidad del trabajo presentado.
- Para realizar predicciones sobre parasitosis intestinal en el Hospital Pediátrico Alfonso Villagómez Román podemos utilizar los árboles de clasificación, si bien no tienen una capacidad predictora muy elevada, funcionan de mejor manera que la regresión logística.
- Los árboles de clasificación presentan un 2.9% más eficiencia que el modelo de regresión logística obtenido. Una causa de este suceso es que los árboles de clasificación presentan porcentajes de discriminación dejando en claro cuáles son los límites de aceptación en los valores de las variables estudiadas.

RECOMENDACIONES

- Este trabajo de investigación es de utilidad para la institución donde se realizó el estudio, por lo que se considera pertinente su socialización con el personal de salud del servicio de emergencias del Hospital Pediátrico Alfonso Villagómez Román.
- Utilizar de manera rutinaria los árboles de clasificación podría resultar útil para prevenir al personal de salud sobre los posibles casos de parasitosis intestinal, lo cual podría contribuir a un manejo adecuado de esta patología.
- Los árboles de clasificación son una herramienta útil pero no indispensable en el sector de la salud, ya que en otros campos como por ejemplo el administrativo; la técnica es sumamente importante puesto que se podría saber un número tentativo de pacientes que acudirán a la casa de salud con una determinada enfermedad y de esta manera poder adquirir los medicamentos necesarios para un óptimo funcionamiento.

BIBLIOGRAFÍA

ABELLANA SANGRA, R. & FARRAN CODINA, A., 2015. *Identifiación, impacto y tratamiento de datos perdidos y atípicos en epidemiología nutricional.* [En línea] Available at: https://www.renc.es/imagenes/auxiliar/files/NUTR.%20COMUN.%20SUPL.%201-2015_Tratamiento%20atipicos.pdf
[Último acceso: 1 01 2023].

APT BARUCH, W. L., 2013. *Parasitología Humana.* 1ra ed. s.l.:M.G.H.

BATISTA ROJAS, O. & ÁLVAREZ HERNÁNDEZ, Z., 2013. Parasitismo intestinal en niñas mayores de 5 años de Ciudad Bolívar. *Scielo,* 17(4).

BONILLA PULGAR , G. E. & BONILLA NINA, G. E., 2020. *Revista de Historia, Ptrimonio, Arqueologia y Antropología Americana.* [En línea] Available at: http://www.rehpa.net/ojs/index.php/rehpa/article/view/27/57
[Último acceso: 20 10 2022].

CONGACHA ORTEGA, G. N., 2020. *DSpace ESPOCH.* [En línea] Available at: http://dspace.espoch.edu.ec/handle/123456789/14551
[Último acceso: 24 7 2022].

DE'ATH, G. & FABRICIUS, K., 2000. Classification and Regression Trees: A Powerful Yet Simple Technique for Ecological Data Analysis. *Ecology.*

GLOBAL HEALTH, Division of Parasitic Diseases and Malaria, 2022. *CDC.* [En línea] Available at: https://www.cdc.gov/parasites/es/about.html
[Último acceso: 11 11 2022].

GUJARATI, D. N. & PORTER, D. C., 2010. *Econometría.* Quinta ed. México DF(DF): Mc Graw-Hill.

LEMA PUNÍN, D. C. & INGA MIGUITAMA, M. A., 2018. *FRECUENCIA DE PARASITOSIS INTESTINAL POR MICROSCOPÍA DIRECTA EN LOS ESTUDIANTES DE LAS ESCUELAS RURALES DE LA PARROQUIA SAN BARTOLOMÉ-2017.* [En línea] Available at: http://dspace.ucuenca.edu.ec/handle/123456789/30073
[Último acceso: 12 11 2022].

LIZARES CASTILLO, M., 2017. Comparación de modelos de clasificación: regresión logística y árboles de clasificación para evaluar el rendimiento académico. *CYBERTESIS.*

LOJANO COLLAGUAZO, R. I. & LOJANO PUNIN, M. A., 2018. *PREVALENCIA DE ENTEROPARASITOSIS Y FACTORES DE RIESGO EN ESCOLARES DE LA UNIDAD EDUCATIVA CHIQUINTAD, 2017.* [En línea] Available at: http://dspace.ucuenca.edu.ec/handle/123456789/30073
[Último acceso: 14 Noviembre 2022].

MENACHO CHÁVEZ, C. M., 2022. *Repositorio Universidad Técnica del Norte.* [En línea] Available at: http://repositorio.utn.edu.ec/bitstream/123456789/12737/2/06%20ENF%201308%20TRABAJO%20DE%20GRADO.pdf
[Último acceso: 2 2 2023].

MONTERO PEREZ, A. . P. & HUILCA PERALTA, G. D. P., 2018. *Repositorio Institucional Continental.* [En línea]
Available at: https://hdl.handle.net/20.500.12394/8796
[Último acceso: 2 2 2023].

OMS, 2008. *Noticias ONU.* [En línea]
Available at: https://news.un.org/es/story/2008/08/1140951
[Último acceso: 18 10 2022].

PAHO, s.f. *Paho.org.* [En línea]
Available at: https://www3.paho.org/spanish/ad/dpc/cd/psit-program-page.htm#:~:text=La%20OPS%2FOMS%20calcula%20que,95%25%20en%20algunas%20tribus%20ind%C3%ADgenas.
[Último acceso: 12 11 2022].

PAZMIÑO GÓMEZ, B. J. Y OTROS, 2018. Parasitosis intestinal y estado nutricional en niños de 1-3 años de un centro infantil del cantón Milagro. *Ciencia Unemi,* 11(26), pp. 143-149.

PEÑA, D., 2002. Datos Atípicos. En: *Análisis de Datos Multivariantes.* s.l.:s.n.

PEREZ RAVE, J. & GONZALEZ ECHEVERRIA, F., 2018. Árboles de clasificación vs regresión logística en el desarrollo de competencias genéricas en ingeniería.. *Scielo,* 22(4), pp. 1519-1541.

QUIMICA.ES, s.f. *Huesped (biología).* [En línea]
Available at: https://www.quimica.es/enciclopedia/Hu%C3%A9sped_%28biolog%C3%ADa%29.html
[Último acceso: 12 11 2022].

REVISTA PANAMERICANA DE LA SALUD PÚBLICA, 2008. Prevalencia de parasitismo intestinal en niños quechuas de zonas rurales montañosas de Ecuador. *Scielo.*

REYNOLDS, L., 2021. *eHOW En Español. ¿Hervir el agua la hace destilada?.* [En línea]
Available at: https://www.ehowenespanol.com/hervir-agua-destilada-como_173637/
[Último acceso: 16 11 2022].

SERNA PINEDA, S. C., 2009. *Comparación de árboles de regresión y clasificación y regresión logística.* [En línea]
Available at: https://repositorio.unal.edu.co/handle/unal/2421
[Último acceso: 2 2 2023].

SRIVASTAVA, T., 2019. 11 Important Model Evaluation Metrics for Machine Learning Everyone should know. *Analytics Vidhya.*

TIMOFEEV, R., 2004. *Classification and Regression Trees (CART) Theory and Applications.* [En línea]
Available at: https://www.academia.edu/13700196/Classification_and_Regression_Trees_CART_Theory_and_Applications
[Último acceso: 2023 2 2].

VALLEJO SAMANIEGO, D. G., 2010. *Wordpress.* [En línea]
Available at: https://digvas.wordpress.com/personajes/alfonso-villagomez-roman/#:~:text=Este%20m%C3%A9dico%20riobambe%C3%B1o%20naci%C3%B3%20el,y%

20a%20aliviar%20los%20dolores%20humanos.
[Último acceso: 20 10 2022].

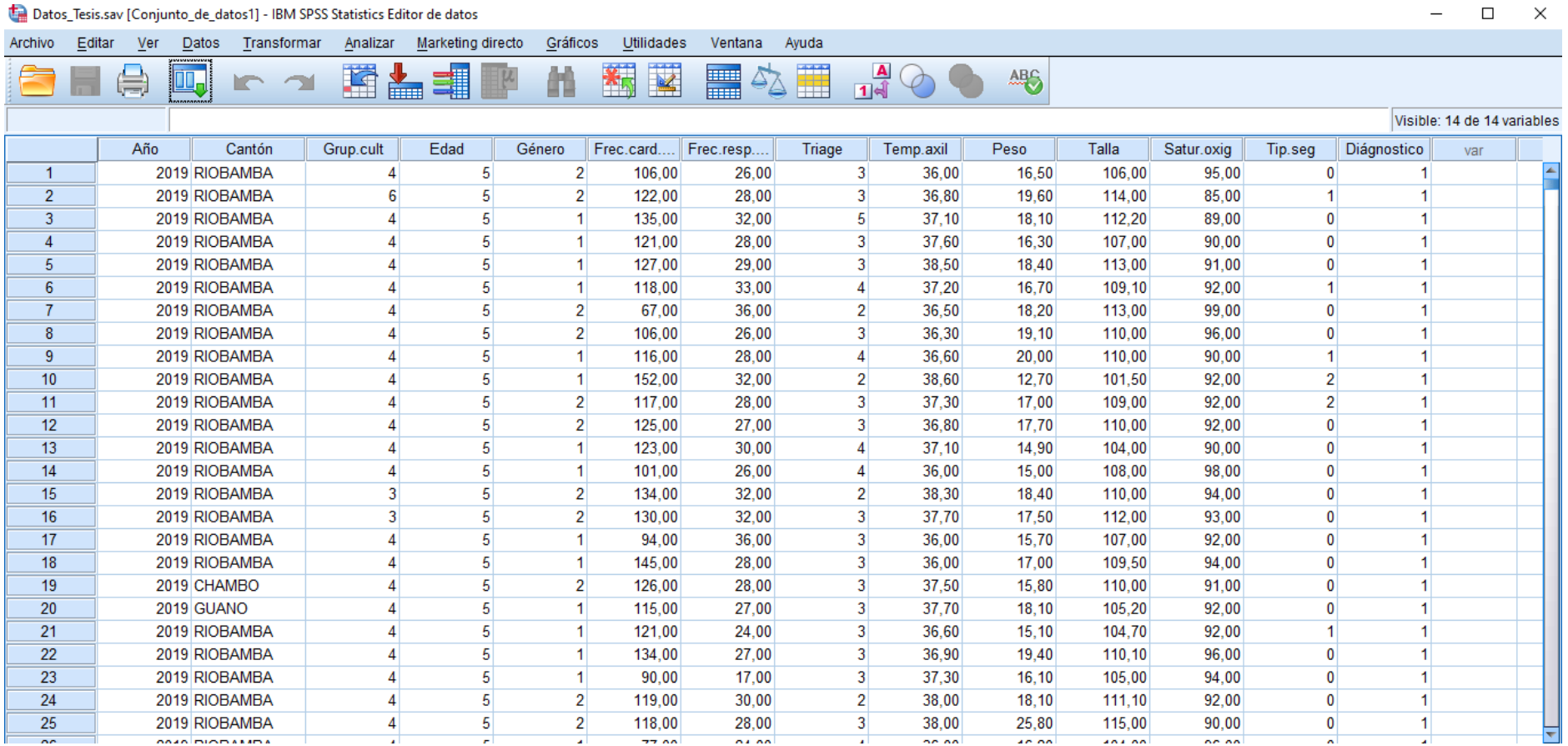
VIADA, C. Y OTROS, 2016. REVISIÓN SISTEMÁTICA DE LOS MÉTODOS DE IMPUTACIÓN DE DATOS FALTANTES. En: s.l.:s.n., pp. 113-130.

ANEXOS

ANEXO A: Parte base de datos pacientes atendidos por emergencia en el HPAVR (2019-2021)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Año	Cantón	Grupo cultural	Edad en años cumplidos	Género	Frecuencia cardiaca min.	Frecuencia respiratoria min.	TRIAJE	Temperatura axilar °C	Peso Kg	Talla (m)	Satura. Oxígeno	Tipo de seguro de salud	Diagnóstico de alta
2	2019	GUANO	MESTIZO/A	8 Años 7 Meses	FEMENINO	92	28	II	38,9	24,9	128,7	92	IESS	Diarrea y gastroenteritis de presunto origen infeccioso
3	2019	RIOBAMBA	MESTIZO/A	5 Años 1 Meses	MASCULINO	106	26	III	36	16,5	106	95	NO TIENE SEGURO	Diarrea y gastroenteritis de presunto origen infeccioso
4	2019	RIOBAMBA	MESTIZO/A	6 Años 11 Meses	MASCULINO	163	28	III	38,6	26,2	121,9	92	IESS	Diarrea y gastroenteritis de presunto origen infeccioso
5	2019	RIOBAMBA	MESTIZO/A	8 Años 6 Meses	FEMENINO	177	32	II	37,4	24,7	127,8	90	NO TIENE SEGURO	Diarrea y gastroenteritis de presunto origen infeccioso
6	2019	RIOBAMBA	MESTIZO/A	8 Años 11 Meses	MASCULINO	110	32	II	36,2	28	131	93	NO TIENE SEGURO	Diarrea y gastroenteritis de presunto origen infeccioso
7	2019	RIOBAMBA	INDÍGENA	7 Años 7 Meses	MASCULINO	102	30	IV	36	20,6	118	92	IESS	Parasitosis intestinal, sin otra especificacion
8	2019	RIOBAMBA	INDÍGENA	9 Años 3 Meses	FEMENINO	118	23	II	38,7	32,9	122,5	92	IESS	Diarrea y gastroenteritis de presunto origen infeccioso
9	2019	RIOBAMBA	MESTIZO/A	9 Años 1 Meses	MASCULINO	94	28	III	37,2	31,3	125,6	91	NO TIENE SEGURO	Diarrea y gastroenteritis de presunto origen infeccioso
10	2019	RIOBAMBA	INDÍGENA	6 Años 9 Meses	FEMENINO	108	32	II	37	17,3	111	91	NO TIENE SEGURO	Diarrea y gastroenteritis de presunto origen infeccioso
11	2019	RIOBAMBA	MULATO/A	5 Años 5 Meses	MASCULINO	122	28	III	36,8	19,6	114	85	IESS	Diarrea y gastroenteritis de presunto origen infeccioso
12	2019	RIOBAMBA	MESTIZO/A	8 Años 3 Meses	FEMENINO	81	24	III	37,8	27,6	120,1	91	NO TIENE SEGURO	Diarrea y gastroenteritis de presunto origen infeccioso
13	2019	RIOBAMBA	MESTIZO/A	8 Años 7 Meses	MASCULINO	117	28	I	39	28	132	89	IESS	Diarrea y gastroenteritis de presunto origen infeccioso
14	2019	RIOBAMBA	MESTIZO/A	9 Años 2 Meses	MASCULINO	78	36	III	36,2	31,5	131,1	99	NO TIENE SEGURO	Diarrea y gastroenteritis de presunto origen infeccioso
15	2019	RIOBAMBA	INDÍGENA	9 Años 11 Meses	FEMENINO	112	28	II	38,1	43,3	144	93	NO TIENE SEGURO	Diarrea y gastroenteritis de presunto origen infeccioso
16	2019	RIOBAMBA	MESTIZO/A	9 Años 10 Meses	MASCULINO	128	26	III	39	31,1	131,5	91	ISSPOL	Diarrea y gastroenteritis de presunto origen infeccioso
17	2019	RIOBAMBA	MESTIZO/A	8 Años 1 Meses	MASCULINO	112	28	IV	37,1	29,2	125,9	91	NO TIENE SEGURO	Diarrea y gastroenteritis de presunto origen infeccioso
18	2019	RIOBAMBA	MESTIZO/A	9 Años 9 Meses	MASCULINO	96	24	III	36	42,8	139,5	93	IESS	Diarrea y gastroenteritis de presunto origen infeccioso
19	2019	RIOBAMBA	INDÍGENA	6 Años 0 Meses	MASCULINO	82	24	IV	36	19,7	111	93	IESS	Otras helmintiasis intestinales especificadas
20	2019	GUANO	MESTIZO/A	6 Años 7 Meses	FEMENINO	130	28	III	36	15,8	114	94	OTRO SEGURO	Diarrea y gastroenteritis de presunto origen infeccioso
21	2019	RIOBAMBA	MESTIZO/A	5 Años 6 Meses	FEMENINO	135	32	V	37,1	18,1	112,2	89	NO TIENE SEGURO	Diarrea y gastroenteritis de presunto origen infeccioso
22	2019	RIOBAMBA	MESTIZO/A	8 Años 1 Meses	FEMENINO	125	28	II	37,1	19,5	19,5	91	NO TIENE SEGURO	Diarrea y gastroenteritis de presunto origen infeccioso
23	2019	RIOBAMBA	MESTIZO/A	5 Años 2 Meses	FEMENINO	121	28	III	37,6	16,3	107	90	NO TIENE SEGURO	Diarrea y gastroenteritis de presunto origen infeccioso
24	2019	RIOBAMBA	MESTIZO/A	8 Años 3 Meses	MASCULINO	136	35	III	36,3	8	73	90	NO TIENE SEGURO	Diarrea y gastroenteritis de presunto origen infeccioso
25	2019	GUANO	MESTIZO/A	7 Años 0 Meses	MASCULINO	92	24	II	36,5	18,4	111,5	92	IESS	Diarrea y gastroenteritis de presunto origen infeccioso
26	2019	RIOBAMBA	MESTIZO/A	7 Años 0 Meses	FEMENINO	100	24	III	37,8	16,8	109	93	NO TIENE SEGURO	Diarrea y gastroenteritis de presunto origen infeccioso
27	2019	RIOBAMBA	MESTIZO/A	6 Años 1 Meses	FEMENINO	96	28	IV	36,7	21,7	121	94	IESS	Parasitosis intestinal, sin otra especificacion

ANEXO B: Parte base de datos categorizada en SPSS



	Año	Cantón	Grup.cult	Edad	Género	Frec.card....	Frec.resp....	Triage	Temp.axil	Peso	Talla	Satur.oxig	Tip.seg	Diágnostico	var
1	2019	RIOBAMBA	4	5	2	106,00	26,00	3	36,00	16,50	106,00	95,00	0	1	
2	2019	RIOBAMBA	6	5	2	122,00	28,00	3	36,80	19,60	114,00	85,00	1	1	
3	2019	RIOBAMBA	4	5	1	135,00	32,00	5	37,10	18,10	112,20	89,00	0	1	
4	2019	RIOBAMBA	4	5	1	121,00	28,00	3	37,60	16,30	107,00	90,00	0	1	
5	2019	RIOBAMBA	4	5	1	127,00	29,00	3	38,50	18,40	113,00	91,00	0	1	
6	2019	RIOBAMBA	4	5	1	118,00	33,00	4	37,20	16,70	109,10	92,00	1	1	
7	2019	RIOBAMBA	4	5	2	67,00	36,00	2	36,50	18,20	113,00	99,00	0	1	
8	2019	RIOBAMBA	4	5	2	106,00	26,00	3	36,30	19,10	110,00	96,00	0	1	
9	2019	RIOBAMBA	4	5	1	116,00	28,00	4	36,60	20,00	110,00	90,00	1	1	
10	2019	RIOBAMBA	4	5	1	152,00	32,00	2	38,60	12,70	101,50	92,00	2	1	
11	2019	RIOBAMBA	4	5	2	117,00	28,00	3	37,30	17,00	109,00	92,00	2	1	
12	2019	RIOBAMBA	4	5	2	125,00	27,00	3	36,80	17,70	110,00	92,00	0	1	
13	2019	RIOBAMBA	4	5	1	123,00	30,00	4	37,10	14,90	104,00	90,00	0	1	
14	2019	RIOBAMBA	4	5	1	101,00	26,00	4	36,00	15,00	108,00	98,00	0	1	
15	2019	RIOBAMBA	3	5	2	134,00	32,00	2	38,30	18,40	110,00	94,00	0	1	
16	2019	RIOBAMBA	3	5	2	130,00	32,00	3	37,70	17,50	112,00	93,00	0	1	
17	2019	RIOBAMBA	4	5	1	94,00	36,00	3	36,00	15,70	107,00	92,00	0	1	
18	2019	RIOBAMBA	4	5	1	145,00	28,00	3	36,00	17,00	109,50	94,00	0	1	
19	2019	CHAMBO	4	5	2	126,00	28,00	3	37,50	15,80	110,00	91,00	0	1	
20	2019	GUANO	4	5	1	115,00	27,00	3	37,70	18,10	105,20	92,00	0	1	
21	2019	RIOBAMBA	4	5	1	121,00	24,00	3	36,60	15,10	104,70	92,00	1	1	
22	2019	RIOBAMBA	4	5	1	134,00	27,00	3	36,90	19,40	110,10	96,00	0	1	
23	2019	RIOBAMBA	4	5	1	90,00	17,00	3	37,30	16,10	105,00	94,00	0	1	
24	2019	RIOBAMBA	4	5	2	119,00	30,00	2	38,00	18,10	111,10	92,00	0	1	
25	2019	RIOBAMBA	4	5	2	118,00	28,00	3	38,00	25,80	115,00	90,00	0	1	

ANEXO C: Codificación variable categórica “Cantón”-Regresión Logística en SPSS

Codificaciones de variables categóricas[a]

			Codificación de parámetro													
		Frecuencia	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Cantón	ALAUSI	4	1,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000
	BUENA F	1	,000	1,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000
	CHAMBO	41	,000	,000	1,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000
	COLTA	11	,000	,000	,000	1,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000
	CUENCA	1	,000	,000	,000	,000	1,000	,000	,000	,000	,000	,000	,000	,000	,000	,000
	CUMANDÁ	1	,000	,000	,000	,000	,000	1,000	,000	,000	,000	,000	,000	,000	,000	,000
	GUAMOTE	10	,000	,000	,000	,000	,000	,000	1,000	,000	,000	,000	,000	,000	,000	,000
	GUANO	131	,000	,000	,000	,000	,000	,000	,000	1,000	,000	,000	,000	,000	,000	,000
	GUARANDA	3	,000	,000	,000	,000	,000	,000	,000	,000	1,000	,000	,000	,000	,000	,000
	GUAYAQUI	2	,000	,000	,000	,000	,000	,000	,000	,000	,000	1,000	,000	,000	,000	,000
	ORELLANA	1	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	1,000	,000	,000	,000
	PALLATAN	4	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	1,000	,000	,000
	PENIPE	6	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	1,000	,000
	PUTUMAYO	1	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	1,000
	QUITO	4	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000
	RIOBAMBA	2452	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000
	SAN MIGU	1	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000
	ZAMORA	1	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000	,000

a. Esta codificación genera coeficientes de indicador.

ANEXO D: Prueba Hosmer y Lemeshow-Regresión Logística en SPSS

Prueba de Hosmer y Lemeshow

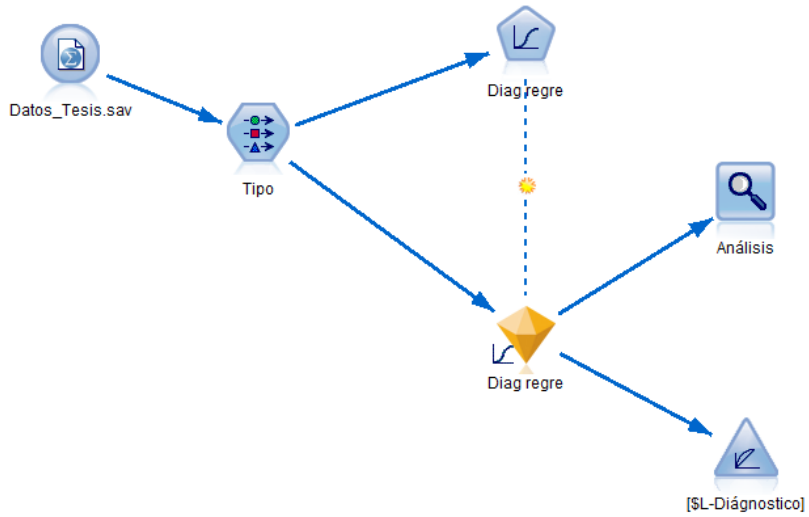
Escalón	Chi-cuadrado	gl	Sig.
1	29,183	2	,000
2	20,423	8	,009
3	17,193	8	,028
4	3,905	8	,866
5	6,100	8	,636

ANEXO E: Pseudo Estadística R^2- Regresión Logística en SPSS

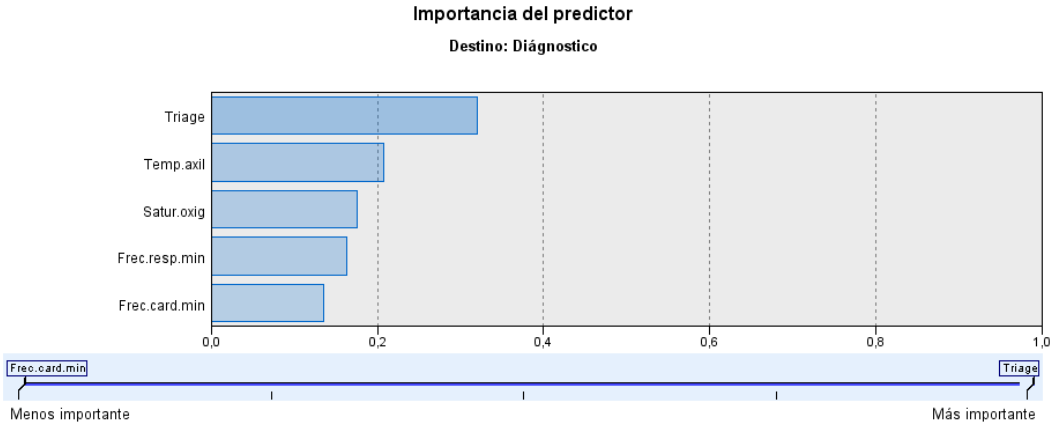
Resumen del modelo

Escalón	Logaritmo de la verosimilitud -2	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	3025,196[a]	,008	,011
2	3000,275[a]	,017	,025
3	2996,079[a]	,018	,027
4	2985,796[a]	,022	,033
5	2980,150[a]	,024	,036

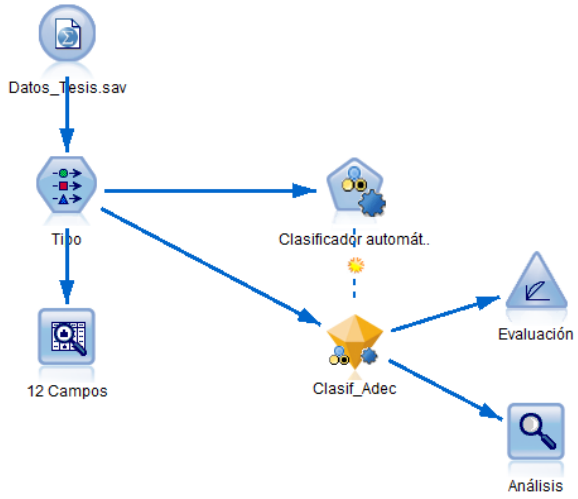
ANEXO D: Ruta de modelado para RL en IBM SPSS MODELER



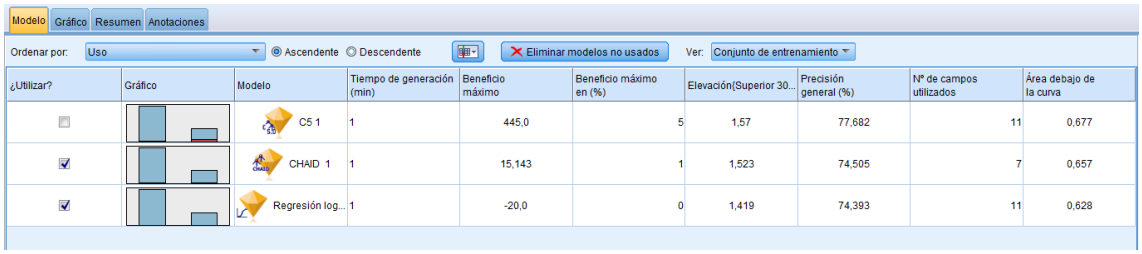
ANEXO G: Importancia de los predictores en el modelo por Regresión Logística



ANEXO E: Ruta utilizando el nodo "clasificador automático" de IBM SPSS MODELER

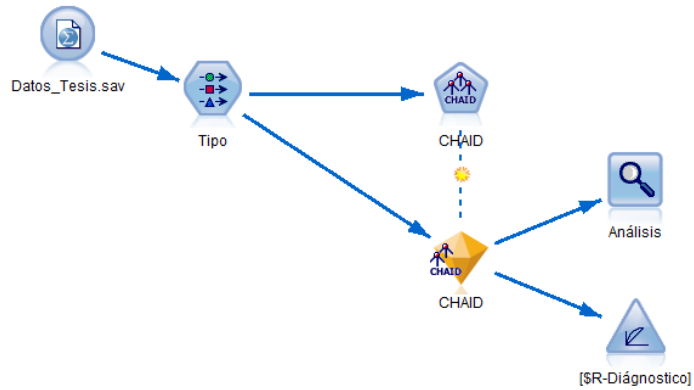


ANEXO F: Modelos propuestos por el clasificador automático de IBM SPSS MODELER

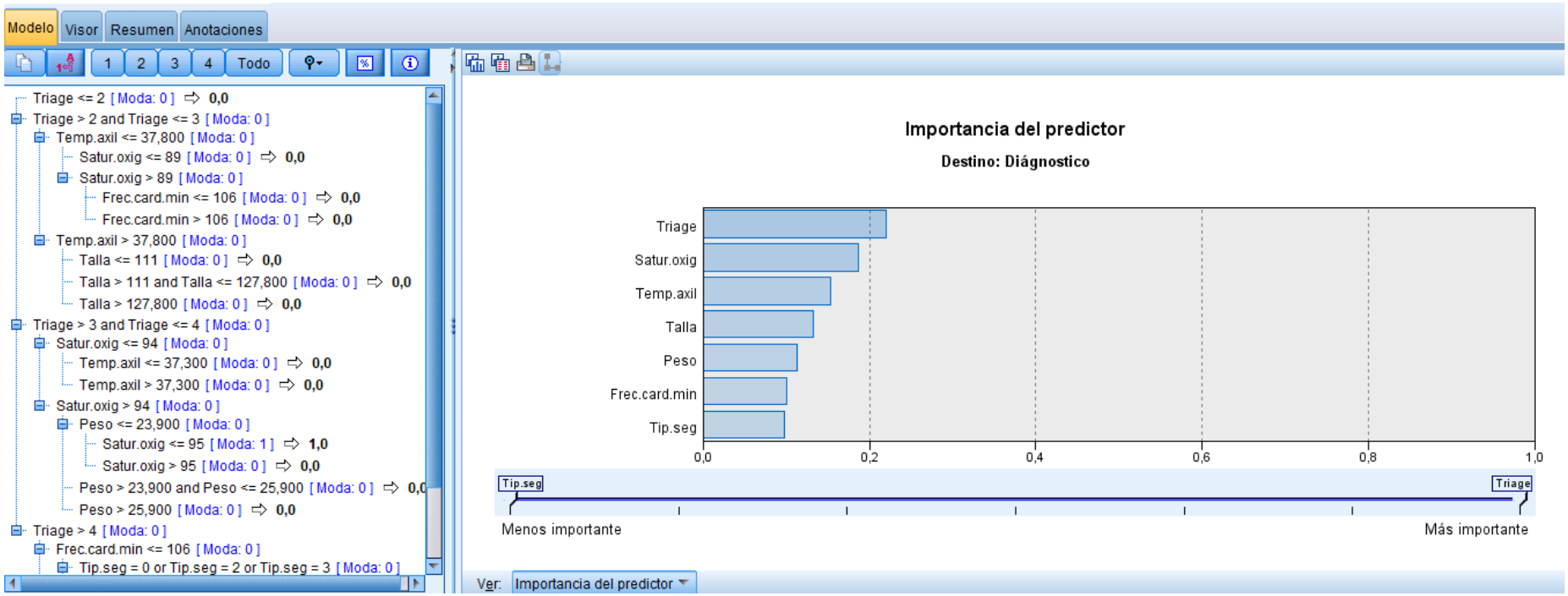


¿Utilizar?	Gráfico	Modelo	Tiempo de generación (min)	Beneficio máximo	Beneficio máximo en (%)	Elevación(Superior 30...	Precisión general (%)	Nº de campos utilizados	Área debajo de la curva
☐		C5 1	1	445,0	5	1,57	77,682	11	0,677
☑		CHAID 1	1	15,143	1	1,523	74,505	7	0,657
☑		Regresión log...	1	-20,0	0	1,419	74,393	11	0,628

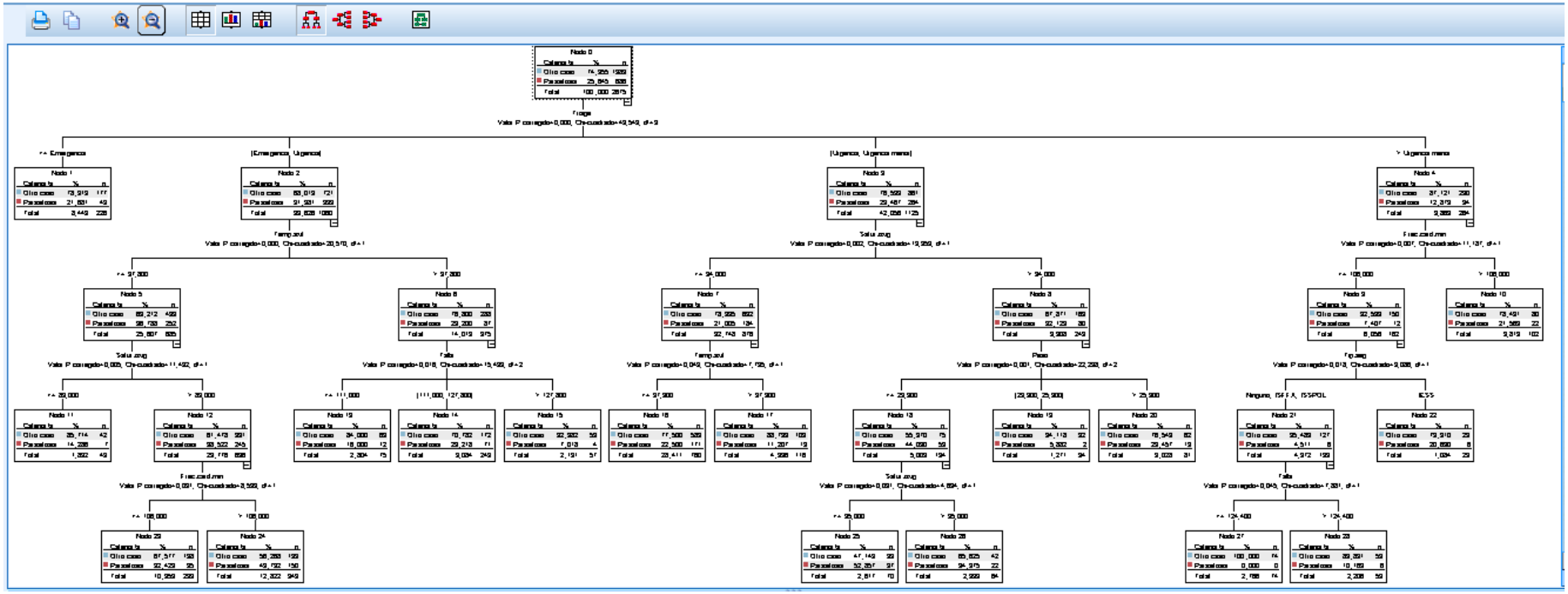
ANEXO G: Ruta con el nodo CHAID para el árbol de clasificación en IBM SPSS MODELER



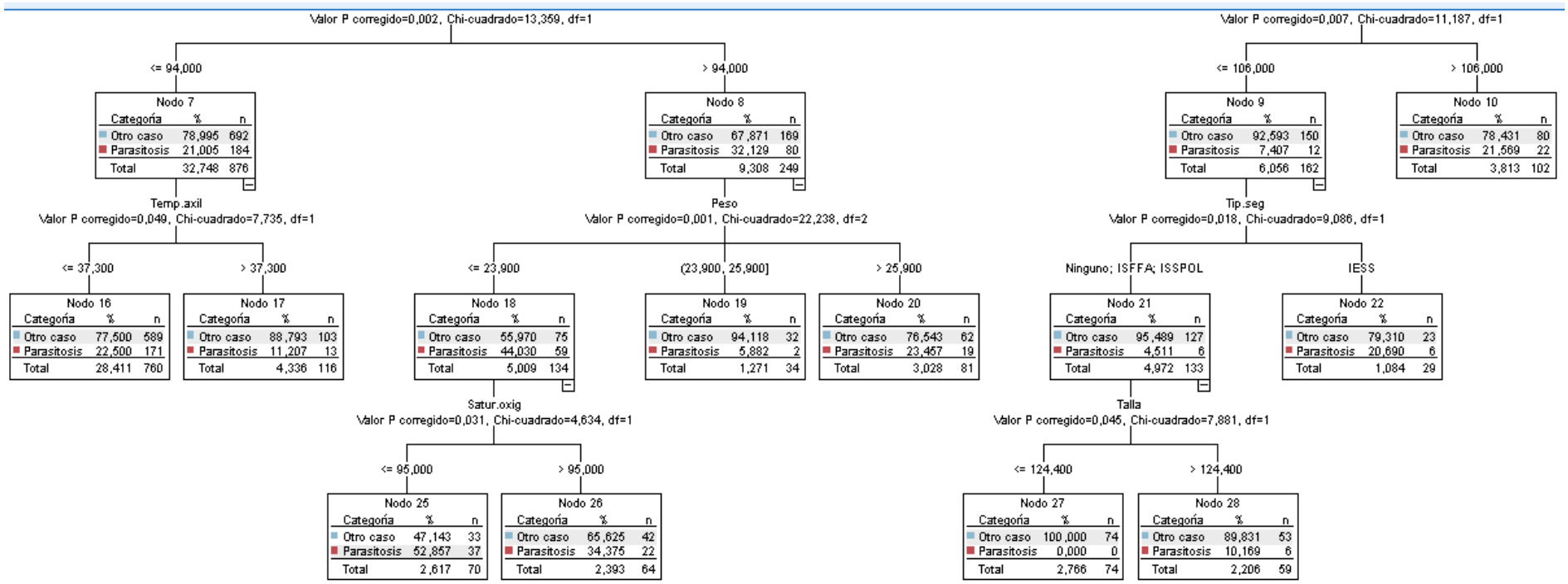
ANEXO H: Descripción del árbol de clasificación en IBM SPSS MODELER



ANEXO I: Árbol de clasificación con el algoritmo CHAID-IBM SPSS MODELER



ANEXO J: Parte del árbol de clasificación que predice la parasitosis intestinal



ESCUELA SUPERIOR POLITÉCNICA DE CHIMBORAZO

DIRECCIÓN DE BIBLIOTECAS Y RECURSOS DEL APRENDIZAJE

UNIDAD DE PROCESOS TÉCNICOS

REVISIÓN DE NORMAS TÉCNICAS, RESUMEN Y BIBLIOGRAFÍA

Fecha de entrega: 04/ 05 / 2023

INFORMACIÓN DE LOS AUTORES
Nombres – Apellidos: Wendy Estefania Isin Daqui Maicol Amable López Sarmiento
INFORMACIÓN INSTITUCIONAL
Facultad: Ciencias
Carrera: Estadística
Título a optar: Ingeniera/o Estadística/o
f. Analista de Biblioteca responsable: Ing. Fernanda Arévalo M.