
# ESCUELA SUPERIOR POLITÉCNICA DE CHIMBORAZO

## FACULTAD DE CIENCIAS

### CARRERA ESTADÍSTICA

# MODELO LOGÍSTICO Y REDES NEURONALES PARA PRONÓSTICO DE ANEMIA EN MENORES DE 5 AÑOS EN EL HOSPITAL PEDIÁTRICO ALFONSO VILLAGÓMEZ ROMÁN PERIODO 2020-2021

**Trabajo de Titulación**

Tipo: Proyecto de Investigación

**Presentado para optar al grado académico de:**

## INGENIERA/O ESTADÍSTICA/O


## AUTORES:

EVELYN MISHEL SÁNCHEZ BARRIGA

SEBASTIAN ISRAEL TENESACA BUENAÑO

**DIRECTORA:** ING. JOHANNA ENITH AGUILAR REYES, MGS.


**Riobamba – Ecuador**

**2023**

Nosotros, Sánchez Barriga Evelyn Mishel y Tenesaca Buenaño Sebastian Israel 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, 28 de noviembre de 2023

**Evelyn Mishel Sánchez Barriga**
**060532218-9**

**Sebastian Israel Tenesaca Buenaño**
**175131549-8**

**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, **MODELO LOGÍSTICO Y REDES NEURONALES PARA PRONÓSTICO DE ANEMIA EN MENORES DE 5 AÑOS EN EL HOSPITAL PEDIÁTRICO ALFONSO VILLAGÓMEZ ROMÁN PERIODO 2020-2021,** realizado por los señores: **SÁNCHEZ BARRIGA EVELYN MISHEL y TENESACA BUENAÑO SEBASTIAN ISRAEL**, 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** |
|---|---|---|
| Biof. Tania Paulina Morocho Barrionuevo, Mgs.<br>**PRESIDENTE DEL TRIBUNAL** |  | 28-11-2023 |
| Ing. Johanna Enith Aguilar Reyes, Mgs.<br>**DIRECTOR DEL TRABAJO DE TITULACIÓN** |  | 28-11-2023 |
| Dra. Jaqueline Elizabeth Balseca Castro, Mgs.<br>**ASESORA DEL TRABAJO DE TITULACIÓN** |  | 28-11-2023 |

## DEDICATORIA

Queridos padres, hermanos y amigos. Quiero expresarles mi más profunda gratitud por su apoyo constante durante este emocionante viaje de mi vida. Sin su apoyo incondicional, esta tesis no habría sido posible. Han sido mi fuente de inspiración y motivación en cada paso del camino, y estoy eternamente agradecido por todo lo que han hecho por mí.

A mis queridos padres, gracias por su amor incondicional, por creer en mí incluso cuando yo dudaba de mí mismo. Sus consejos, palabras de aliento y apoyo financiero han sido invaluables. Siempre han sido mi mayor apoyo y motivación en todo lo que hago. Esta tesis es un logro no solo para mí, sino también para ustedes, ya que su dedicación y sacrificio han sido fundamentales en mi éxito académico. A mis queridos hermanos, gracias por estar siempre a mi lado, por brindarme su aliento y por ser mis confidentes. Nuestros momentos de risa, nuestras discusiones y nuestras aventuras juntos han sido inolvidables. Su amor y apoyo me han dado la fortaleza para enfrentar los desafíos y alcanzar mis metas. A mis amigos, gracias por compartir conmigo el arduo camino de la preparación para esta tesis. Nuestros debates, nuestras discusiones y nuestro apoyo mutuo han sido fundamentales para mi crecimiento académico y personal. Han sido una verdadera familia durante estos años y estoy agradecido de tenerlos en mi vida.

En resumen, mi tesis no es solo un logro personal, sino también un logro compartido con todos ustedes. Gracias por ser parte de mi vida y por creer en mí. Dedico esta tesis a ustedes, con amor y gratitud.

**Sebastian**

El presente trabajo de investigación va dedicado a mi padre, por ser mi motor mi ejemplo a seguir por enseñarme a ser una persona persistente por enseñarme a seguir a delante a pesar de todas las adversidades y problemas que se me presentaron en todo este trayecto académico, por ser mi talón de Aquiles y contenerme en mis peores momentos. A mi abuelita por siempre brindarme su cariño y su amor por caminar siempre de mi mano y enseñarme que puedo lograr todo lo que me propongo. A mi hermana y cuñado por siempre estar ahí para mí por apoyarme incondicionalmente por ayudarme en los momentos difíciles por brindarme muchas oportunidades para poder seguir adelante con mi carrera por cuidarme y estar al pendiente de mí y nunca dejarme sola.

**Evelyn**

**AGRADECIMIENTO**

Quiero expresar mi más sincero agradecimiento a mis padres Luis Reinaldo Tenesaca y Gloria María Buenaño por su amor, apoyo y aliento incondicional a lo largo de mi proceso educativo. Su constante apoyo emocional y su confianza en mí han sido un motor importante para enfrentar los retos y superar los obstáculos en este camino académico. Además, quiero agradecer a la carrera de Estadística por brindarme una educación integral y de calidad. Los conocimientos adquiridos en esta carrera han sido fundamentales para el desarrollo de mi tesis y para mi formación profesional en el campo de la estadística. Agradezco especialmente a mis profesores y mentores de la carrera de Estadística por su dedicación y compromiso en mi formación académica. Sus enseñanzas, orientación y retroalimentación han sido de gran valor en el desarrollo de mi trabajo de tesis. Agradezco de todo corazón a mis amigos Leila Jibaja, Alison Pérez, Isaac Trejo y Andrés Mogro por su apoyo y consejos en este logro académico. Sin su aliento y amistad incondicional, esta tesis no habría sido posible. ¡Gracias por estar siempre ahí para apoyarme en este camino!

**Sebastian**

Quiero empezar agradeciendo a mi papá, abuelita, hermana y cuñado, quienes han estado apoyándome en todo este proceso, tanto económica como emocionalmente, son las personas que han confiado en mi me han alentado a seguir luchando por conseguir este sueño. A la Ing. Johanna Aguilar y a la Ing. Natalia Pérez quienes me guiaron y tutelaron en el desarrollo de este trabajo de investigación. A mis amigos y compañeros con quienes he pasado momentos amenos durante esta trayectoria

**Evelyn**

**ÍNDICE DE CONTENIDO**

**ÍNDICE DE TABLAS**

**ÍNDICE DE ILUSTRACIONES**

**RESUMEN**

La presente tesis tiene como objetivo principal determinar el mejor modelo para pronosticar la anemia en niños menores de 5 años atendidos en el Hospital Pediátrico Alfonso Villagómez Román durante el período 2020-2021. Para lograr este objetivo, se desarrolló un estudio utilizando dos enfoques de modelado: el Modelo Logístico y las Redes Neuronales. En el primer lugar, se llevó a cabo una revisión exhaustiva de la literatura para identificar los posibles factores asociados a la anemia infantil, lo cual permitió establecer un marco teórico pertinente para el análisis. Posteriormente, se recolectó información relevante sobre estos factores mediante el acceso a las historias clínicas de los niños menores de 5 años que presentaban anemia y que fueron atendidos en el hospital pediátrico. Con el conjunto de datos recopilados, se procedió a modelar la presencia de anemia en niños menores de 5 años utilizando tanto el Modelo Logístico como el Modelo de Redes Neuronales. Ambos modelos fueron aplicados y evaluados meticulosamente a través de ajuste de bondad como las métricas de la matriz de confusión y la curva de ROC para obtener pronósticos precisos.t Los resultados obtenidos fueron comparados para determinar cuál de los dos modelos ofrecía una mejor capacidad de predicción. En este análisis comparativo, se encontró que las Redes Neuronales superaron al Modelo Logístico, demostrando ser el modelo de predicción más efectivo para pronosticar la anemia en niños menores de 5 años. En conclusión, el estudio demuestra que el uso de Redes Neuronales para el pronóstico de anemia en niños menores de 5 años es más efectivo que el Modelo Logístico. Estos hallazgos tienen implicaciones significativas para la atención médica pediátrica, ya que un pronóstico preciso de la anemia en etapas tempranas puede contribuir a una intervención médica oportuna y mejorar el cuidado de la salud infantil.

**Palabras clave:** <ANEMIA>, <NIÑOS MENORES DE 5 AÑOS>, <MODELO LOGÍSTICO>, <REDES NEURONALES>, <PRONÓSTICO>.

2136-DBRA-UPT-2023

**SUMMARY**

The main objective of the present study was to determine the best model for predicting anemia in children under 5 years of age treated at the Alfonso Villagómez-Román Pediatric Hospital during the period 2020-2021. First, a comprehensive review of the literature was performed to identify the possible factors associated with childhood anemia, which allowed a relevant theoretical framework for analysis to be established. Subsequently, relevant information on these factors was collected by accessing the clinical records of children under 5 years of age with anemia who were treated at the pediatric hospital. With the data set collected, we proceeded to model the presence of anemia in children under 5 years of age using both the Logistic Model and the Neural Network Model. Both models were applied and meticulously evaluated through goodness of fit such as the confusion matrix and ROC curve metrics to obtain accurate forecasts. The results obtained were compared to determine which of the two models offered better predictive ability. In this comparative analysis, it was found that Neural Networks outperformed the Logistic Model, proving to be the most effective prediction model for predicting anemia in children under 5 years of age. In conclusion, the study demonstrates that the use of Neural Networks for the prognosis of anemia in children under 5 years of age is more effective than the Logistic Model. These findings have significant implications for pediatric health care, as accurate prognosis of anemia in early stages can contribute to timely medical intervention and improve child health care.

**Keywords:** <ANEMIA>, <CHILDREN UNDER 5 YEARS OF AGE>, <LOGISTIC MODEL>, <NEURAL NETWORKS>, <PROGNOSIS>.

## INTRODUCCIÓN

La anemia es un problema de salud pública que afecta a niños menores de 5 años en todo el mundo, siendo una de las principales causas de morbimortalidad infantil. En Riobamba, el Hospital Pediátrico Alfonso Villagómez Román, atiende a un gran número de pacientes con anemia, lo que representa un desafío para el sistema de salud y la atención médica especializada.

El pronóstico de la anemia en menores de 5 años es fundamental para su adecuada atención y tratamiento, ya que permite identificar tempranamente a aquellos niños que presentan mayor riesgo de complicaciones y requerirían intervenciones médicas más agresivas. Tradicionalmente, los modelos logísticos han sido utilizados para predecir la probabilidad de desarrollar anemia en niños, utilizando variables clínicas y de laboratorio. Sin embargo, con los avances en la tecnología y el uso de redes neuronales, se ha abierto una nueva posibilidad para mejorar la precisión y exactitud de los pronósticos con diferentes factores.

El objetivo de esta tesis fue desarrollar un modelo logístico y el uso de redes neuronales para el pronóstico de anemia en menores de 5 años en el Hospital Pediátrico Alfonso Villagómez Román durante el período 2020-2021. Se utilizó factores sociales de los pacientes atendidos en el hospital para entrenar y evaluar los modelos, con el fin de determinar su capacidad predictiva y comparar su desempeño. Además, se buscará identificar las variables clínicas más relevantes en la predicción de la anemia, con el objetivo de mejorar la detección temprana y el manejo de esta condición en los niños atendidos en el hospital.

El desarrollo de un modelo logístico y el uso de redes neuronales en el pronóstico de anemia en menores de 5 años en un hospital pediátrico es una herramienta potencialmente útil para apoyar la toma de decisiones clínicas, optimizar el uso de recursos y mejorar la atención médica a esta población vulnerable. Los resultados obtenidos de este estudio podrían tener aplicaciones clínicas importantes en la identificación temprana y manejo adecuado de la anemia en niños, contribuyendo así a la mejora de la salud infantil y el bienestar de la comunidad.

### Antecedentes investigativos

La medicina es una disciplina que se apoya en gran medida en la evidencia científica para tomar decisiones clínicas informadas y mejorar la atención médica. En este contexto, la estadística desempeña un papel fundamental al permitir el análisis y la interpretación de datos clínicos y epidemiológicos, así como la realización de investigaciones rigurosas para evaluar la efectividad de

intervenciones médicas. En este artículo, se revisará la importancia de la estadística en la medicina moderna, destacando su papel en la generación de evidencia científica confiable y en la toma de decisiones clínicas basadas en datos.

La estadística es esencial en la medicina para analizar y sintetizar datos clínicos y epidemiológicos. Como señalan (Smith, et al., 2019), la estadística permite resumir grandes cantidades de datos en medidas descriptivas como la media, la mediana y la moda, lo que facilita la comprensión de la distribución y variabilidad de los datos clínicos. Además, la estadística inferencial permite realizar estimaciones y pruebas de hipótesis para generalizar los hallazgos de un estudio a una población más amplia (Brown, et al., 2018). Estos métodos estadísticos son cruciales para determinar la significancia clínica y epidemiológica de los resultados de investigación en medicina.

En la medicina basada en la evidencia, la estadística también juega un papel central en la evaluación de la efectividad de intervenciones médicas. Por ejemplo, en un estudio de ensayo clínico aleatorizado, (Jones, et al., 2020) encontraron que la terapia de reemplazo hormonal redujo significativamente el riesgo de fracturas óseas en mujeres posmenopáusicas. Estos hallazgos se basaron en un análisis estadístico riguroso de los datos recolectados, que incluyó la comparación de grupos de tratamiento y control mediante pruebas estadísticas adecuadas.

Además, la estadística también es esencial en la evaluación de factores de riesgo en la medicina preventiva. Por ejemplo, un estudio de cohorte de largo plazo realizado por (Smith, et al., 2017) demostró que el tabaquismo se asociaba significativamente con un mayor riesgo de desarrollar enfermedades cardiovasculares. Estos hallazgos se basaron en análisis estadísticos de seguimiento a largo plazo de una gran muestra de individuos, lo que resalta la importancia de la estadística en la identificación de factores de riesgo en la medicina preventiva.

La anemia en menores de 5 años ha sido un problema de salud pública que ha sido objeto de investigación y estudio en diversos contextos y regiones a nivel mundial. A continuación, se presentan algunos antecedentes históricos relevantes relacionados con el tema de estudio:

Estudio en Etiopía: En un estudio llevado a cabo en Etiopía, se evaluó la prevalencia y los factores de riesgo asociados con la anemia en niños menores de 5 años. Los autores encontraron que la anemia estaba asociada con la edad, el estado nutricional, la presencia de parásitos intestinales y la infección por malaria (Gebreegziabher, et al., 2018)

Investigación en Bangladesh: Un estudio en Bangladesh investigó la prevalencia de la anemia

en niños menores de 5 años y su asociación con factores socioeconómicos, demográficos y nutricionales. Los resultados mostraron que la anemia estaba relacionada con la falta de acceso a servicios de salud, la pobreza, la falta de educación materna y la falta de una alimentación diversificada y adecuada (Rahman, et al., 2019)

Investigación en Perú: En un estudio realizado en Perú, se evaluó la prevalencia y los factores de riesgo de la anemia en niños menores de 5 años en una población rural. Los resultados mostraron que la anemia estaba asociada con la edad, el sexo, el nivel socioeconómico, la falta de acceso a servicios de salud y la desnutrición (Rosado-Mendoza, et al., 2017)

Investigación en India: En un estudio llevado a cabo en India, se desarrolló un modelo de predicción de la anemia en niños menores de 5 años utilizando redes neuronales artificiales. Los autores encontraron que el modelo de redes neuronales tenía una alta precisión en la predicción de la anemia, y sugerían su uso como una herramienta de apoyo en la toma de decisiones clínicas (Kaur, et al., 2018)

Estos antecedentes históricos destacan la importancia de la anemia en niños menores de 5 años como un problema de salud pública, y la necesidad de desarrollar modelos predictivos y utilizar tecnologías avanzadas como las redes neuronales para mejorar la detección temprana y el manejo de esta condición en la población infantil. Las referencias citadas proporcionan evidencia científica sobre la relevancia y el contexto de estudio del tema de investigación propuesto en la tesis.

**Antecedentes históricos**

La anemia infantil es un problema de salud pública que ha sido objeto de atención en diferentes épocas y regiones del mundo. Se ha evidenciado que la anemia en niños menores de 5 años puede tener consecuencias negativas en su crecimiento y desarrollo, así como en su salud a largo plazo.

En cuanto a la historia de los hospitales pediátricos, se remonta a la antigüedad, donde se evidencia la existencia de instituciones y médicos que atendían a niños enfermos. Sin embargo, los hospitales pediátricos como los conocemos hoy en día, especializados en la atención de niños y adolescentes, surgieron en el siglo XIX con el objetivo de brindar una atención médica especializada y centrada en las necesidades de los niños.

En Ecuador, la historia de la anemia infantil y los hospitales pediátricos ha evolucionado a lo largo del tiempo, con la implementación de políticas de salud y programas de atención que han buscado abordar este problema. Por ejemplo, el Ministerio de Salud Pública de Ecuador ha desarrollado

estrategias para la prevención, diagnóstico y tratamiento de la anemia en la población infantil, con énfasis en la promoción de una alimentación adecuada y la administración de suplementos de hierro y otros nutrientes.

En cuanto a los hospitales pediátricos en Ecuador, el Hospital Pediátrico Alfonso Villagómez Román es un referente en la atención especializada de niños y adolescentes en el país, brindando servicios médicos, quirúrgicos y de cuidados intensivos a pacientes pediátricos. Este hospital, ubicado en Ecuador, ha sido reconocido por su labor en la atención y cuidado de la salud de los niños, incluyendo la atención de enfermedades como la anemia.

# CAPÍTULO I

## 1. PROBLEMA DE INVESTIGACIÓN

### 1.1. Planteamiento del problema

La anemia es una complicación de salud mundial que afecta tanto a los países desarrollados como en desarrollo, y contribuye significativamente a la morbilidad y mortalidad infantil menores de cinco años. Alrededor del 43 % de los menores de cinco años son anémicos en todo el mundo, en el Ecuador 7 de cada 10 menores de 1 año sufren de anemia por deficiencia de hierro. Estas cifras casi se duplican en poblaciones rurales (Moyano Brito, et al., 2019).

Como menciona el manual de la Atención Integral a las Enfermedades Prevalentes de la Infancia (AIEPI), la anemia es una complicación de salud mundial que afecta tanto a los países desarrollados y los que están en desarrollo, contribuyendo significativamente a la morbilidad y mortalidad en niños menores de cinco años (Moyano Brito, et al., 2019). Sin embargo, la anemia es importante porque es muy común y tiene graves consecuencias, incluso la muerte en los casos más graves. La Constitución de la República del Ecuador establece que el estado brindará atención a los niños menores de seis años que garantice su nutrición, salud y cuidado diario. (Secretaría Técnica Ecuador Crece Sin Desnutrición, 2013).

En el 2018 se estudió los casos y controles pareado de base institucional. La población de estudio estuvo constituida por niños de 1 a 4 años que asistieron al Centro de Desarrollo Infantil “Los pitufos del Valle” entre enero y octubre de 2018. La fórmula de comparación de proporciones modificada se utilizó para elegir la muestra del estudio. Se estimó una seguridad del 95 %, un poder estadístico del 80 % y una frecuencia de exposición de casos del 31 % y una frecuencia de exposición de controles del 58 %. Además, se consideró una relación de riesgos (OR) de 2. (Moyano Brito, et al., 2019).

Para lo cual, se determinará cuáles de las dos técnicas, regresión logística y redes neuronales es el mejor modelo predictivo para poder pronosticar la Anemia en niños menores de 5 años atendidos en el hospital pediátrico Alfonso Villagómez Román periodo 2020-2021.

## 1.2. Limitaciones y delimitaciones

Las limitaciones en esta investigación son varias desde el desconocimiento de los factores conocidos en años posteriores de niños con anemia, ya que la mayor parte de ellos acude con otras intenciones, y mediante los procedimientos adecuados se les diagnostica con anemia de tipo no especificado u otro. También como la confidencialidad de las historias clínicas para la determinación de los factores clave, ya que no se puede acceder a ellas sin la debía autorización.

## 1.3. Problema general de investigación

¿Cuál es el mejor modelo para pronosticar la anemia en niños menores de 5 años en el hospital pediátrico Alfonso Villagómez Román periodo 2020-2021?

## 1.4. Problemas especificos de investigación

- ¿Cuáles son los factores asociados a la anemia infantil en menores de 5 años?
- ¿Cómo ejecutar la clasificación de los pacientes regresión logística y redes neuronales?
- ¿Cuáles son los mejores resultados predictivos obtenidos de la comparación entre los modelos?

## 1.5. Objetivos

### *1.5.1. Objetivo general*

- Determinar el mejor modelo que pronostique la anemia en niños menores de 5 años atendidos en el hospital pediátrico Alfonso Villagómez Román periodo 2020-2021

### *1.5.2. Objetivos específicos*

- Determinar a través de un marco teórico pertinente los posibles factores asociados a la anemia infantil.
- Recolectar la información de los posibles factores, usando las historias clínicas de los niños menores de 5 años que presenten anemia, y son atendidos en el hospital Pediátrico.
- Modelar la presencia de anemia en niños menores de 5 años mediante el modelo logístico.
- Modelar la presencia de anemia en niños menores de 5 años mediante el modelo de redes neuronales.

- Comparar los pronósticos obtenidos con los modelos de regresión logística y de redes neuronales

## 1.6. Justificación

### *1.6.1. Justificación teórica*

La deficiencia de hierro es el principal causante de anemia, la cual afecta a la población de los países de ingresos bajos, medios y altos (Health, Organization World, 2015). Es por ello que el Ministerio de salud junto con el estado Ecuatoriano ha desarrollado campañas de prevención contra la anemia incluso en conjunto con Perú, ya que es perjudicial para el desarrollo del niño menor de 5 años, pues afecta directamente al sistema inmunológico lo que causa que las defensas del sistema bajen, y así el paciente pueda adquirir con mayor facilidad diferentes infecciones, dando lugar a una mayor demanda de atención en los hospitales del Ecuador; pese a los esfuerzos que se han implementado no se ha podido contrarrestar de manera significativa.

Para ello se realizara este trabajo de investigación, haciendo uso aplicativo de una de las herramientas más importantes de la IA, las redes neuronales cuya base es una arquitectura perceptrón multicapa asociada al algoritmo de estimación "backpropagation error", la cual será comparada con la técnica paramétrica regresión logística binaria con el fin de evaluar, cuál de estos dos modelos puede pronosticar la anemia en los niños menores de 5 años en el Hospital Pediátrico Alfonso Villagómez Román., de tal modo que se implementen políticas de apoyo para las diferentes instituciones, creando campañas en diferentes comunidades de la región y la vez crear políticas de prevención para los niños que padecen esta enfermedad, así mismo este trabajo servirá como una fuente base, para otras futuras investigaciones que se realicen.

Los objetivos establecidos en este trabajo de investigación es determinar el mejor modelo para pronosticar la anemia en menores de 5 años, estimando el modelo logístico binario mediante la técnica estadística de máxima verosimilitud y el modelo de redes neuronales haciendo uso del algoritmo "backpropagation error".

### *1.6.2. Justificación metodológica*

Con el presente estudio de investigación, se utilizarán técnicas de redes neuronales y regresión logística para determinar los posibles factores asociados a la anemia en niños menores de 5 años registrados en el Hospital Pediátrico Alfonso Villagómez Román (HPAVR). A través de la revisión bibliográfica adecuada, se generará un modelo de clasificación que permitirá la predicción de la anemia en estos infantes.

Una vez obtenidos los modelos de redes neuronales y regresión logística, se evaluarán y se compararán las medidas de bondad de ajuste, como el error de la matriz de confusión y el área bajo la curva ROC, para determinar el mejor modelo con la mayor capacidad de predicción. Este modelo será utilizado para predecir la anemia en niños menores de 5 años en el HPAVR.

### *1.6.3. Justificación práctica*

La anemia es una condición médica que afecta a niños menores de 5 años en todo el mundo, siendo un problema de salud pública significativo. La identificación temprana de la anemia en esta población es crucial para un manejo adecuado y la prevención de complicaciones a largo plazo. En este contexto, el uso de técnicas estadísticas avanzadas como redes neuronales y regresión logística puede ser de gran utilidad para predecir la presencia de anemia en niños.

Un tipo de modelo de aprendizaje automático que se basa en la estructura y funcionamiento del cerebro humano se conoce como redes neuronales. Estos modelos son capaces de aprender patrones complejos y no lineales en los datos, lo que los hace adecuados para el análisis de datos médicos con múltiples variables y relaciones no lineales. Al utilizar redes neuronales en la predicción de anemia en niños, se pueden aprovechar sus capacidades de aprendizaje profundo para identificar patrones sutiles en los datos clínicos, como niveles de hemoglobina, edad, género, antecedentes familiares, entre otros, que pueden no ser evidentes a simple vista. Esto puede ayudar a mejorar la precisión y la capacidad de generalización del modelo, lo que a su vez puede contribuir a una detección temprana y precisa de la anemia en niños.

Por otro lado, la regresión logística es una técnica estadística que se utiliza para modelar la probabilidad de un evento binario, como la presencia o ausencia de anemia en este caso. La regresión logística permite examinar la relación entre una o más variables predictoras, como el nivel de hemoglobina, y la probabilidad de tener anemia. Además, es una técnica relativamente simple y fácil de interpretar, lo que la hace atractiva para su aplicación en el campo médico.

El uso de redes neuronales y regresión logística en la predicción de anemia en niños en el ámbito clínico puede tener varias ventajas. Primero, puede ayudar a los médicos y profesionales de la salud a identificar a los niños en riesgo de anemia de manera temprana, lo que permitirá un manejo oportuno y adecuado. Segundo, puede contribuir a una mejor asignación de recursos y a una toma de decisiones informada en el cuidado de la salud infantil, optimizando la atención médica y previniendo complicaciones a largo plazo. Además, el uso de técnicas estadísticas avanzadas puede mejorar la precisión y la capacidad de generalización de los modelos de predicción, lo que puede

resultar en un mayor rendimiento en comparación con métodos convencionales.

### 1.7. Hipótesis

Las redes neuronales es el mejor modelo que permitirá pronosticar la anemia en niños menores de 5 años en el Hospital Pediátrico Alfonso Villagómez Román.

# CAPÍTULO II

## 2. MARCO TEÓRICO

### 2.1. Bases teóricas

#### *2.1.1. Datos atípicos*

Un valor outlier o atípico es una puntuación extrema dentro de una variable. Son observaciones aisladas cuyo comportamiento se diferencia claramente del comportamiento medio del resto de las observaciones. Pueden utilizarse herramientas de análisis exploratorio de datos para detectar casos atípicos en un contexto univariante. Por ejemplo, en el gráfico de caja y bigotes los valores atípicos se presentan como puntos aislados en los extremos de los bigotes (Pérez López, et al., 2007 págs. 333-334).

#### *2.1.2. Identificación de datos atípicos*

Para encontrar grupos atípicos, se eliminan todos los puntos sospechosos de la muestra. Esto nos permite evitar el enmascaramiento y calcular el vector de medias y la matriz de covarianzas sin distorsiones. El primer paso para encontrar observaciones sospechosas es identificar las que son evidentes en relación an una variable. Para ello se puede utilizar el histograma o los diagramas de caja. Una regla fundamental es considerar las observaciones sospechosas como

$$\frac{|x_i - \underline{med}(x)|}{Meda\ (x)} > 4,5,$$

donde med(x) es la mediana de las observaciones, que es una estimación confiable del centro de los datos, y med(x) es la mediana de las desviaciones absolutas |x_i-med(x)|, que es una medida confiable de la dispersión. Este método puede verse como una estandarización robusta de los datos (Peña, 2002 pág. 122).

#### *2.1.3. Depuración de datos atípicos*

La técnica para depurar los datos atípicos búsqueda de proyecciones es propuesta por Peña y Prieto (2001), en el cual se dice que este método funciona con pocas variables. La cual trata proyectar los datos sobre ciertas direcciones específicas, escogidas de manera que tengan alta probabilidad de mostrar los atípicos cuando existan. Se ha comentado que en muestras univariantes

una pequeña proporción de atípicos hace aumentar el coeficiente de kurtosis, lo que sugiere investigar las direcciones donde los puntos proyectados tengan máxima kurtosis univariante. Por otro lado, un grupo grande de atípicos puede producir bimodalidad y baja kurtosis, por lo que conviene también explorar las direcciones donde los puntos proyectados tengan mínima kurtosis. La idea del procedimiento es buscar p direcciones ortogonales de máxima kurtosis y p direcciones ortogonales de mima kurtosis, eliminar provisionalmente los datos extremos en estas direcciones, calcular la media y la matriz de covarianzas con los datos no sospechosos y después identificar los datos atípicos como aquellos que son extremos con la distancia de Mahalanobis calculada con las estimaciones no contaminadas. Dada la muestra multivariante ($x_1, \ldots, x_n$) el proceso se realiza como sigue (Peña, 2002):

- Sean $\overline{X}$ y $S$ el vector de medias y la matriz de covarianzas de los datos. Estandarizar los datos de forma multivariante y sean $z_i = S_x^{-\frac{1}{2}}(X_i - \underline{X})$ los datos estandarizados con media cero y matriz de covarianzas identidad. Tomar

- Calcular la dirección $d_j$ con norma unidad que maximiza el coeficiente de kurtosis univariante de los datos proyectados. Llamando $Y_i^{(j)} = d'_j z_i^{(j)}$, a los datos proyectados sobre la dirección $d_j$ , esta dirección se obtiene como solución de:

$$max \sum^{Ä} Y_i^{(j)} - \overline{Y}^{j\,ä_4} + \lambda(d'd - 1)$$

  que puede resolverse como se indica en el apéndice 4.1.

- Proyectar los datos sobre un espacio de dimensión $p - j$ definido como el espacio ortogonal a la dirección$d_j$. Para ello tomar $z^{(j+1)} = (I - d_j d'_j)z^{(j)}$. Hacer $j = j + 1$

- Repetir (2) y (3) hasta obtener las aa direcciones, $d_1, \ldots, d_p$.

- Repetir (2) y (3) pero ahora minimizando la kurtosis en lugar de maximizarla para obtener otras aa direcciones, $d_{p+1}, \ldots, d_{2p}$

- Considerar como sospechosos aquellos puntos que están claramente alejados del resto de estas dos direcciones, es decir,

$$\frac{.Y_i^{(j)} - med(Y^{(j)}).}{Meda\,(Y^{(j)})} > 5,$$

A continuación, se eliminan todos los valores sospechosos detectados y se vuelve a 2 para analizar los datos restantes. La estandarización multivariante ahora se realizar· con la nueva media y matriz de covarianzas de los datos restantes. Los pasos 2 a 6 se repiten hasta que no se detecten más datos

atípicos o se haya eliminado una proporción de datos prefijada, por ejemplo, un máximo del 40 % de los datos(Peña, 2002).

### *2.1.4. Datos faltantes*

Valores no disponibles que serían útiles o significativos para el análisis de los resultados se denominan datos faltantes. Hay una variedad de tipos de datos que pueden faltar y muchas razones por las que pueden ocurrir. Al enfrentar la ausencia de datos, estos dos factores son cruciales. Lo principal es decidir si la pérdida es aleatoria, es decir, afecta a todos los individuos de manera uniforme, o si puede ser causada por una razón o razones específicas que pueden introducir sesgos que invaliden los resultados.(Datos faltantes (missing values), 2014 págs. 332-334).

### *2.1.5. Imputación de datos faltantes*

Una técnica tradicional y muy conocida para el tratamiento de datos faltantes es la imputación. Las técnicas de imputación se pueden clasificar, en primer lugar, en dos grandes grupos: las técnicas de imputación simples y las de imputación múltiple (Muñoz Rosas, et al., 2009 pág. 3).

Técnica de imputación simple: Las técnicas simples de imputación presentan algunas ventajas frente a las técnicas de imputación múltiple. Por ejemplo, las técnicas simples tienen una implantación más sencilla sin que por el contrario sufran una importante pérdida de eficiencia en comparación con las técnicas de imputación múltiple. Por último, destacamos que las técnicas simples de imputación se pueden dividir en: Imputación por el método de medias no condicionadas, Imputación por medias condicionadas para datos agrupados, Imputación con variables ficticias, Imputación mediante un distribución no condicionada e Imputación por regresión (Muñoz Rosas, et al., 2009).

Técnica de imputación múltiple: La imputación múltiple requiere la construcción de M ($\geq$ 2) conjuntos de datos completos, los cuales se obtienen reemplazando cada dato faltante por M valores imputados, obtenidos mediante el mismo procedimiento de imputación. Aunque la imputación múltiple es una aproximación muy potente, sufre algunas limitaciones que no debemos pasar por alto. Por último, destacamos que las técnicas múltiples de imputación se pueden dividir en: Procedimiento de imputación múltiple (IM) (Medina, et al., 2007 pág. 31).

### *2.1.6. Regresión logística*

La regresión logística es un instrumento estadístico de análisis multivariado, de uso tanto explicativo como predictivo (Chitarroni, 2002). En estadística, la regresión logística es un tipo de análisis de

regresión que se utiliza para predecir los resultados de variables categóricas (variables que pueden contener un número limitado de categorías) a partir de predictores o variables independientes. Esto es útil para modelar la probabilidad de eventos en función de otros factores.

El análisis de regresión logística, parte del conjunto de modelos lineales generalizados (GLM), utiliza la función logit como función de enlace. Las probabilidades que describen los posibles resultados de los ensayos individuales se modelan como funciones de variables explicativas utilizando funciones logísticas. La regresión logística es ampliamente utilizada en medicina y ciencias sociales. Otros nombres para la regresión logística utilizados en varias áreas de aplicación incluyen modelo logístico, modelo logístico y clasificador de máxima entropía.

### *2.1.7. Objetivos de la regresión logística*

El objetivo principal es modelar matemáticamente la existencia de eventos, evaluar la probabilidad de su ocurrencia y proporcionar una imagen estadística de la influencia de varios factores y sus valores o niveles.

Los modelos de regresión logística en ciencias de la salud nos permiten analizar los resultados en términos explicativos y predictivos. Podemos conocer la fuerza de asociación mediante los OR de los factores de riesgo con el efecto estudiado y conocer el valor predictivo de cada uno de los factores de riesgo o del modelo en su conjunto. Los estudios prospectivos con finalidad pronóstica (epidemiología clínica), prospectivos con finalidad analítica (cohortes), ensayos caso-control (riesgo atribuible) y ensayos clínicos son los tipos de estudios en los que se utiliza más a menudo (Vindell, 2021).

### *2.1.8. Supuestos del modelo de regresión logísticas*

El análisis discriminante se basa en supuestos distribucionales, pero la regresión logística no. Sin embargo, si los predictores tienen una distribución normal multivariada, la solución puede ser más estable. La multicolinealidad entre los predictores también puede resultar en estimaciones sesgadas y errores estándar inflados, como sucede con otras formas de regresión. Cuando la pertenencia a grupos es una variable categórica auténtica, el procedimiento es más eficaz. Sin embargo, si la pertenencia a grupos se basa en valores de una variable continua (por ejemplo, CI alto en lugar de CI bajo), se debe considerar el uso de la regresión lineal para aprovechar la información mucho más rica proporcionada por la propia variable continua. (IBM, 2021).

Si se encuentra multicolinealidad, las variables se transforman promediando y también se pueden

incluir interacciones entre variables categóricas. Como se mencionó anteriormente, un modelo no requiere una relación lineal entre sus predictores y sus covariables, pero sí requiere que la relación entre los predictores y las probabilidades logarítmicas sea lineal. Otro requisito para usar un modelo es que el número de observaciones o el tamaño de la muestra sea lo suficientemente grande y suficiente (Calvo & Domínguez, 2002).

### *2.1.9. Regresión logística binaria*

Se utiliza principalmente cuando la variable de interés es una variable ficticia (solo dos clases), se puede agrupar (glogit), la variable dependiente sigue una distribución binomial o es una variable independiente (la variable dependiente sigue una distribución de Bernoulli) (Abanto Chavarri, 2022).

La regresión logística binaria es una herramienta estadística que puede ser univariante o multivariante y cae dentro de la categoría de MLG.

Su propósito es establecer causalidad en presencia de variables dicotómicas o ficticias. Dado que la variable de interés es una variable de dos clases con una distribución de probabilidad de Bernoulli, tenemos:

$$E\,(Y_i/X_i) = 1P_i + 0\,(1 - P_i) = P_i$$

Por lo tanto, la expectativa de que el evento ocurra o que la variable dependiente tome un valor igual a 1 es igual a la probabilidad estimada por el modelo(Abanto Chavarri, 2022).

### *2.1.10. Conceptos básicos de la regresión logística*

(Zaiontz, 2014) El enfoque básico es usar el siguiente modelo de regresión, empleando la notación de del método de mínimos cuadrados para regresión múltiple:

$$lnOdds(E) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

donde la función de probabilidades es como se indica en la siguiente definición.

*Función de probabilidad*

(Zaiontz, 2014) $odds(E)$ es la probabilidad de que ocurra el evento $E$ , a saber

$$\text{Odds}(E) = \frac{P(E)}{P\,(E')} = \frac{P(E)}{1 - P(E)}$$

Donde p tiene un valor $0 \leq p \leq 1$ (es decir, p es un valor de probabilidad), podemos definir la función de probabilidades como

$$\text{Odds}(p) = \frac{p}{1-p}$$

Para nuestros propósitos, la función de probabilidades tiene la ventaja de transformar la función de probabilidad, que tiene valores de 0 a 1, en una función equivalente con valores entre 0 y ∞. Cuando tomamos el logaritmo natural de la función de probabilidades, obtenemos un rango de valores de $-\infty$ a $\infty$ .

*Función de registro*

(Zaiontz, 2014) La función logit es el logaritmo de la función de probabilidades, a saber, $\text{logit}(E) = \ln Odds\,(E)$ , o

$$\text{logit}(p) = \ln Odds(p) = \ln \frac{p}{1-p} = \ln p \_ \ln(1 \_ p)$$

*Modelo logístico*

(Zaiontz, 2014) Con base en el modelo logístico descrito anteriormente, tenemos

$$\frac{P(E)}{1-P(E)} = Odds(E) = e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k+\varepsilon}$$

y entonces

$$p = P(E) = \frac{e^{b_0+b_1 x_1+b_2 x_2+\cdots+bx_k}}{1+e^{b_0+b_1 x_1+b_2 x_2+\cdots+bx_k}}$$
$$= \frac{1}{1+e^{-(b_0+b_1 x_1+b_2 x_2+\cdots+bx_k)}}$$
$$= \frac{1}{1+e^{-b_0-\sum_{j=1}^{k} b_j x_j}}$$

Aquí cambiamos al modelo basado en la muestra observada (y así el parámetro $\pi$ se reemplaza por su estimación de muestra $p$, los coeficientes $\beta_j$ se reemplazan por las estimaciones de muestra $b_j$ y el término de error $\varepsilon$ se elimina). Para nuestros propósitos, consideramos que el evento $E$ es que la variable dependiente y tiene un valor de 1. Si y toma solo los valores 0 o 1, podemos pensar en E como un éxito y en el complemento $E'$ de $E$ como un fracaso. Esto es como para los ensayos en una distribución binomial.

Una muestra consta de n elementos de datos de la forma ($y_i$, $x_{i1}$, $x$, . . . , $x_{ik}$), pero para la regresión logística, cada y $i$ solo toma el valor 0 o 1. Ahora sea $E_i$ el evento de que $y_i = 1$ y $p_i = P(E_i)$. Así

como la línea de regresión estudiada previamente proporciona una forma de predecir el valor de la variable dependiente y a partir de los valores de las variables independientes $x_1, \ldots, x_k$ en la regresión logística tenemos

$$p = P(y = 1) = \frac{1}{1 + e^{-b_0 - \sum_{j=1}^{k} b_j x_j}}$$
$$\text{logit}(p) = \ln\frac{p}{1-p} = \ln\frac{P(y=1)}{1-P(y=1)} = b_0 + \sum_{j=1}^{k} b_j x_j$$

En el caso donde k = 1, tenemos

$$p = \frac{1}{1 + e^{-b_0 - b_1 x}}$$

*Curva Sigmoidea*

(Zaiontz, 2014) Tal curva tiene forma sigmoidea:

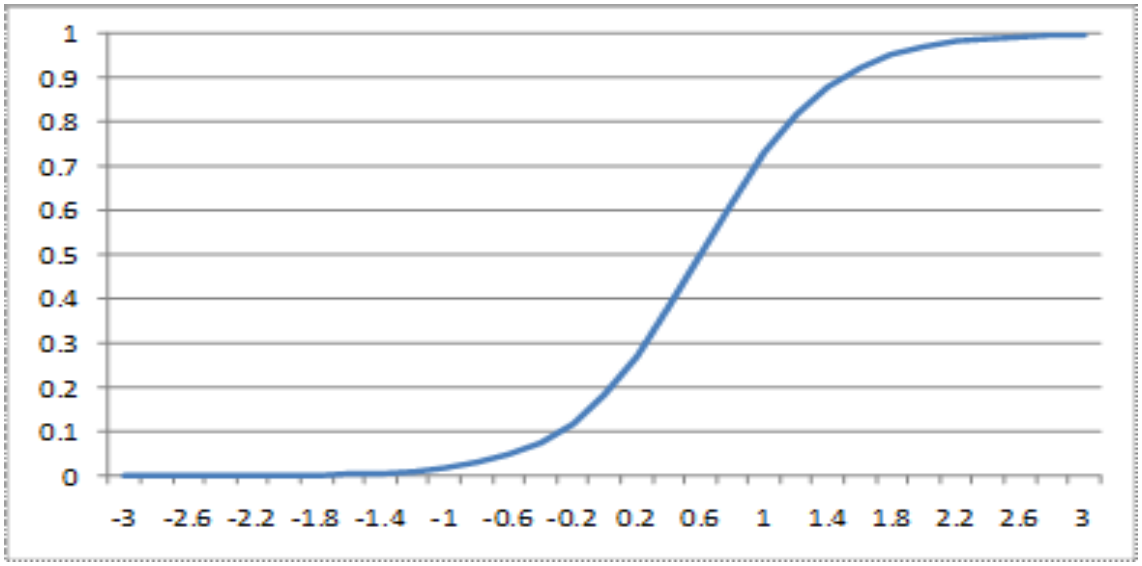



**Ilustración 2–1:** Curva Sigmoidea

**Obtenido de:** https://www.real- statistics.com/logistic-regression/basic-concepts-logistic-regression/

Los valores de $b_0$ y $b_1$ determinan la dirección de ubicación y la extensión de la curva. La curva es simétrica respecto al punto donde $x = -b_0/b_1$ . De hecho, el valor de $p$ es 0,5 para este valor de $x$.

### *2.1.11. Cuando usar regresión logística*

(Zaiontz, 2014) Se utiliza la regresión logística en lugar de la regresión múltiple ordinaria porque no se cumplen los supuestos necesarios para la regresión ordinaria. En particular por los siguientes casos:

- La suposición del modelo de regresión lineal de que los valores de *y* se distribuyen normalmente no se puede cumplir ya que y solo toma los valores 0 y 1.

- La suposición del modelo de regresión lineal de que la varianza de y es constante entre los valores de *x* (homogeneidad de las varianzas) tampoco se puede cumplir con una variable binaria. Como la varianza es $p(1-p)$ cuando el 50 por ciento de la muestra consta de 1, la varianza es .25, su valor máximo. A medida que avanzamos hacia valores más extremos, la varianza disminuye. Cuando $p = 0,10$ o $0,90$, la varianza es $(0,1)(0,9) = (0,09)$ y, por lo tanto, cuando p se acerca a 1 o 0, la varianza se acerca a 0.

- Con el modelo de regresión lineal, los valores pronosticados serán mayores que uno y menores que cero si se mueve lo suficiente en el eje *x*. Tales valores son teóricamente inadmisibles para las probabilidades.

Para el modelo logístico, no se puede utilizar el método de mínimos cuadrados para calcular los valores de los coeficientes $b_i$; en su lugar, se emplean las técnicas de máxima verosimilitud, como se describe a continuación, para encontrar estos valores.

### *2.1.12. Razón de probabilidades*

noindent (Zaiontz, 2014) La razón de posibilidades entre dos elementos de datos en la muestra se define de la siguiente manera:

$$R_{x_i, x_j} = \frac{\text{Odds}(x_{i1}, \ldots, x_{ik})}{\text{Odds}\ x_{jk}, \ldots, x_{jk}} = \frac{e^{b_0 + \sum_{m=1}^{k} b_m x_{im}}}{e^{b_0 + \sum_{m=1}^{k} b_m x_{jm}}} = e^{\sum_{m=1}^{k} b_m (x_{im} - x_{jm})}$$

Usando la notación $p(x) = P(X)$, el logaritmo de la razón de probabilidades de las estimaciones se define como

$$logit \left( \frac{p_{(x+1)}}{p_x} \right)$$

### *2.1.13. Log-verisimilitud máxima*

(Zaiontz, 2014) El modelo que usaremos se basa en la distribución binomial, es decir, la probabilidad de que los datos de la muestra ocurran como ocurre viene dada por

$$L = \prod_{i=1}^{n} p_i^{\ y_i} (1 - p_i)^{1 - y_i}$$

Tomando el logaritmo natural de ambos lados y simplificando obtenemos la siguiente definición: La estadística de probabilidad logarítmica se define de la siguiente manera:

$$LL = \ln L = \sum_{i=1}^{n} [y_i \ln p_i + (1 - y_i) \ln(1 - p_i)]$$

donde $y_i$ son los valores observados mientras que $p_i$ son los valores teóricos correspondientes. Nuestro objetivo es encontrar el valor máximo de $LL$ suponiendo que los $p_i$ son como en la definición del modelo logístico Esto nos permitirá encontrar los valores de las coordenadas $b_i$ .

### *2.1.14. Redes neuronales*

(Izaurieta, et al., 200) Las redes neuronales son un tipo de modelo de aprendizaje automático que se inspira en la forma en que funciona el cerebro humano. Están formadas por capas de neuronas, que son unidades matemáticas que se encargan de procesar la información. Las redes neuronales pueden ser utilizadas para resolver una gran variedad de problemas, como el reconocimiento de patrones, la clasificación de imágenes o la predicción de valores futuros. Las redes neuronales se componen de una o varias capas de neuronas, que son unidades matemáticas que procesan la información y transmiten el resultado a otras neuronas en capas posteriores. Cada neurona se conecta con otras neuronas a través de enlaces, que tienen un peso asociado. El peso de un enlace determina la importancia de la señal que se transmite a través de él.

(Cornejo Ruiz & Quispe Gavino, 2008) Las entradas y salidas de una neurona pueden ser clasificadas en dos grandes grupos, binarias o continuas. Las neuronas binarias (digitales) sólo admiten dos valores posibles. Este tipo de neurona generalmente usa los alfabetos {0,1} o {-1,1}. Las neuronas continuas (analógicas), por otro lado, admiten valores dentro de un rango específico, que generalmente se define como [-1, 1]. El tipo de neurona a utilizar depende de la aplicación y del modelo a construir.

Las redes neuronales se pueden entrenar ajustando los pesos de los enlaces entre las neuronas de acuerdo con como de bien o de mal se ajusta el modelo a los datos de entrenamiento. Este proceso se llama aprendizaje supervisado, ya que se proporcionan ejemplos de entrada y salida deseados para cada ejemplo de entrenamiento.

### *2.1.15. Retropropagación*

El proceso de aprendizaje clásico de la retropropagación en una red neuronal sigue una serie de pasos que modifican los pesos que definen la red neuronal usando el error final cometido como objetivo a minimizar. Una observación de los pasos de este algoritmo nos muestra cómo y donde actúa la búsqueda aleatoria progresiva(Toda-Caraballo et al., 2010).

De forma esquemática, en cada paso del algoritmo de la retropropagación se han de calcular iterativamente las variaciones ($S^m$) de cada peso, usando la variación anterior para modificar éstos. Debido a que el objetivo es el de minimizar el error cuadrático de la salida de la red con los datos

de referencia (t), la secuencia de estas funciones queda como sigue:

$$S^M = -2F^M(n^M)(t-a)$$

$$S^m = F(n^m)\ W^{m+1}\ S^{m+1},\quad \forall m = M-1,\ldots,2,1$$

Dónde F es la matriz definida como

$$F^m(n^m) =$$

Con $f^m$ la función de la capa m-ésima y

$$n_i^m = \sum_{j=1}^{S^{m-1}} (W_i^m, j a_j^{m-1} + b_i^m)$$

En el que, $a^{m-1}$ son las salidas de las neuronas de la capa anterior. Finalmente se actualiza la matriz de pesos (w) y sesgos (b)

$$W^m\ (k+1) = W^m(k) - \alpha s^m (a^{m-1})'$$

$$b^m(k+1) = b^m(k) - \alpha s^m$$

Dónde $\alpha$ es el llamado índice de aprendizaje.

Si la red tiene varias capas, las variaciones $s^m$ pueden adoptar valores ínfimos, debido a que se obtienen de la derivación de funciones con variación casi nula en gran parte de su dominio. El algoritmo de la retropropagación calcula estas variaciones iterativamente usando su predecesora, por lo que si una de ellas roza valores nulos, éstos se propagan sucesiva-mente, haciendo que las variaciones en los pesos aso-ciados a estas zonas de la red neuronal sean pequeñas, promoviendo más variación en pesos que no se vean afectados (Toda-Caraballo et al., 2010).

### *2.1.16. Función de activación*

La salida de los nodos se propaga de una capa a la siguiente utilizando funciones de activación. Las funciones escalares an escalar activan la neurona. Este tipo de funciones permiten incorporar el modelado de datos de entrada no lineales a la red. (Ramírez, 2020)

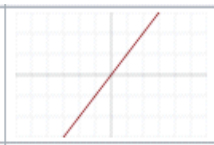
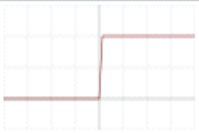
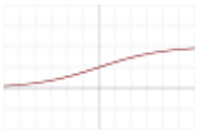
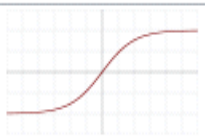
| | | |
|---|---|---|
| Identity |  | $f(x) = x$ |
| Binary step |  | $f(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 & \text{for } x \geq 0 \end{cases}$ |
| Logistic (a.k.a. Sigmoid or Soft step) |  | $f(x) = \sigma(x) = \frac{1}{1+e^{-x}}$ [1] |
| TanH |  | $f(x) = \tanh(x) = \frac{(e^{x} - e^{-x})}{(e^{x} + e^{-x})}$ |

**Ilustración 2–2:** Ejemplo de algunas funciones de activación

**Obtenido de:** https://empresas.blogthinkbig.com

La familia de funciones sigmoideas es la más importante y se utiliza en las funciones de activación; sin embargo, no es la única familia que se utiliza en estas funciones.

### *2.1.17. Algoritmo back propagation error*

los pesos de la red para minimizar el error entre la salida predicha y la salida real. La propagación posterior es un tipo de red de aprendizaje supervisado que utiliza un ciclo de propagación-adaptation de dos fases. Una vez que se ha aplicado un patrón a la entrada de la red como estímulo, este se propaga desde la primera capa a través de las capas superiores de la red, hasta generar una salida. La señal de salida se compara con la salida deseada y se calcula una señal de error para cada una de las salidas. Las salidas de error se propagan hacia atrás, partiendo de la capa de salida, hacia todas las neuronas de la capa oculta que contribuyen directamente a la salida (Bertona, 2005).

La importancia de este proceso consiste en su capacidad de auto adaptar los pesos de las neuronas intermedias para aprender la relación que existe entre un conjunto de patrones dados como ejemplo y sus salidas correspondientes. Después del entrenamiento, si una nueva entrada contiene un patrón que se asemeje a la característica que las neuronas individuales hayan aprendido a reconocer durante su entrenamiento, las neuronas de la capa oculta de la red responderán con una salida activa. Y a la inversa, las unidades de las capas ocultas tienen una tendencia a inhibir su salida si el patrón de entrada no contiene la característica para reconocer, para la cual han sido entrenadas (Cornejo Ruiz & Quispe Gavino, 2008).

### *2.1.18. Curvas ROC*

Es una herramienta estadística para evaluar la capacidad discriminativa de una prueba diagnóstica dicotómica. Son útiles para la elección del punto de corte más adecuado de una prueba, conocer el rendimiento global y comparar la capacidad discriminativa de dos o más pruebas diagnósticas (Martínez Pérez & Pérez Martin, 2023).

La curva ROC (Receiver Operating Characteristic curves) mide la bondad de ajuste mediante su forma gráfica comparando los positivos falsos frente a los positivos verdaderos, según el umbral de discriminación. Por lo cual, los modelos de predicción que se encuentran por encima de la línea discriminante mejor es el método de diagnóstico, los modelos que coinciden con la línea discriminante se clasifican como aleatorios y los que se encuentran debajo de la línea discriminante su método de diagnóstico es el peor (Lizares Castillo, 2017).

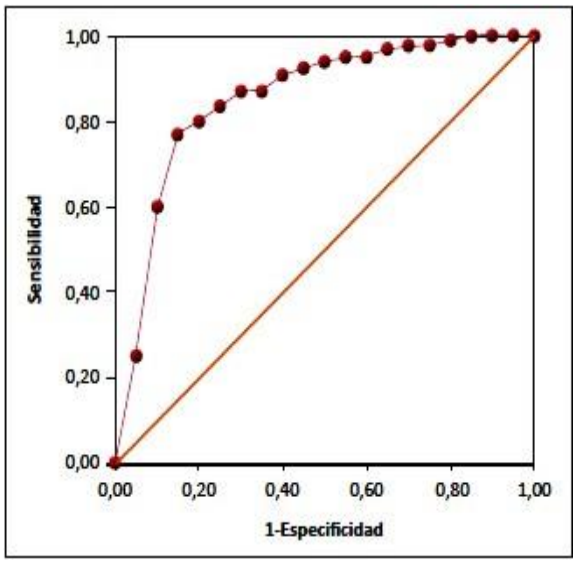



**Ilustración 2–3:** Ejemplo de la curva de ROC

**Fuente:** (Cerda y Cifuentes, 2012)

### *2.1.19. Matriz de confusión*

Una matriz de confusión, también conocida como matriz de error, es una tabla resumida utilizada para evaluar el desempeño de un modelo de clasificación. El número de predicciones correctas e incorrectas se resume con los valores de conteo y se destruye para cada clase. Esta matriz proporciona las predicciones de un algoritmo de aprendizaje monitoreado y los resultados correctos que debería haber demostrado. Por lo tanto, el rendimiento más grande o bajo se puede medir para determinar el tipo de errores y golpes de cada modelo en un proceso de aprendizaje para los datos

propuestos (Terence Shin, 2020).

Según (Terence Shin, 2020) la estructura de una matriz de confusión 2x2 es la siguiente:

Positivo (P): La observación es positiva.

Negativo (N): La observación no es positiva.

Verdadero Positivo (TP): Resultado en el que el modelo predice correctamente la clase positiva.

Verdadero Negativo (TN): Resultado donde el modelo predice correctamente la clase negativa.

Falso Positivo (FP): También llamado error de tipo 1, resultado donde el modelo predice incorrectamente la clase positiva cuando en realidad es negativa.

Falso Negativo (FN): También llamado error de tipo 2, un resultado en el que el modelo predice incorrectamente la clase negativa cuando en realidad es positiva.

### *2.1.20. Métricas de confusión*

Exactitud

Es la proporción de predicciones que el modelo clasificó correctamente. La matriz puede interpretarse como lo suficientemente exacta porque la coincidencia aumenta con la exactitud (Terence Shin, 2020).

$$Accuracy = \frac{\#of\ \ correct\ \ predictions}{total\ \#\ of\ \ predictions} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precisión

La precisión también se conoce como valor predictivo positivo y es la proporción de instancias relevantes entre las instancias recuperadas. En otras palabras, responde a la pregunta ¿Qué proporción de identificaciones positivas fue realmente correcta? (Terence Shin, 2020).

$$Precision = \frac{TP}{TP + FP}$$

Sensibilidad

La sensibilidad, tasa de aciertos o tasa positiva real (TPR), es la proporción de la cantidad total de instancias pertinentes que se recuperaron realmente. Responde a la pregunta ¿Qué proporción de positivos reales se identificó correctamente? (Terence Shin, 2020).

$$Recall = \frac{TP}{TP + FN}$$

Especificidad

La especificidad, también conocida como tasa negativa real (TNR), mide la proporción de negativos reales que se identifican correctamente como tales. Es lo opuesto a la sensibilidad (Terence Shin, 2020).

$$Specificity = \frac{TN}{TN + FP}$$

## 2.2. Bases conceptuales

### *2.2.1. Anemia*

La anemia es definida como una reducción de la concentración de la hemoglobina o de la masa global de hematíes en la sangre periférica por debajo de los niveles considerados normales para una determinada edad, sexo y altura sobre el nivel del mar. En la práctica, el diagnóstico de anemia se establece tras la comprobación de la disminución de los niveles de la hemoglobina y/o el hematocrito por debajo de -2 desviaciones estándar (DE) (o el percentil 3) (Merino et al., 2016).

Las principales causas de la anemia son la falta de vitaminas esenciales para el organismo, como la vitamina B12, la carencia de hierro, el ácido fólico y otras (Merino et al., 2016).

Los principales síntomas de la anemia son fatiga física, dolores de cabeza, problemas de memoria, pérdida de apetito, irritabilidad y hormigueo en manos y pies (Merino et al., 2016).

### *2.2.2. Hemoglobina*

La hemoglobina (Hb) es una proteína compleja formada por grupos hem que contienen hierro y una porción proteica llamada globina. El tetrámero de la molécula Hb está formado por dos pares de cadenas polipeptídicas (alfa y beta), cada una de las cuales tiene un grupo hem; las cadenas polipeptídicas alfa y beta son diferentes químicamente. La interacción dinámica de estos elementos confiere a la Hb propiedades específicas y exclusivas para el transporte reversible del oxígeno (Merino et al., 2016).

Se reconocen 3 tipos de hemoglobina: la Hb fetal (Hb F) y las Hb del adulto (A y A2). En los cromosomas 11 y 16, se encuentran los genes que regulan la síntesis de la Hb. A partir de los 3-6 meses de edad, solo quedan trazas de Hb F, y la relación entre las Hb A y A2 permanecerá ya estable alrededor de 30/1 a lo largo de toda la vida (Merino et al., 2016b).

### *2.2.3. Consecuencias de la anemia infantil*

Existen muchos estudios y revisiones sobre cómo la anemia en los niños impacta negativamente en el desarrollo psicomotor y, a pesar de corregirse la anemia, los niños con este antecedente presentan, a largo plazo, un menor desempeño en las áreas cognitiva, social y emocional. La anemia puede disminuir el desempeño escolar, y la productividad en la vida adulta, afectando la calidad de vida, y en general la economía de las personas afectadas (Zavaleta & Astete-Robilliard, 2017).

### *2.2.4. Clasificación de la anemia*

**Clasificación fisiopatológica**

Desde este punto de vista, las anemias se pueden clasificar según su respuesta reticulocitaria en anemias regenerativas e hiporregenerativas.

*Anemias regenerativas*: se observa una respuesta reticulocitaria elevada, lo cual indica incremento de la regeneración medula (Merino et al., 2016).

*Anemias no regenerativas*: en este caso se observa una respuesta reticulocitaria baja y traducen la existencia de una médula ósea hipo/inactiva (Merino et al., 2016).

**Según la forma de insaturación se clasifica en aguda o crónica**

*Anemia Aguda*: es cuando los valores de Hb y hematíes descienden en forma brusca por debajo de los niveles normales. Esta forma de anemia se presenta en dos situaciones bien definidas: hemorragia y por un aumento en la destrucción de los hematíes (hemólisis) (Merino et al., 2016).

*Anemia Crónica*: Se desarrolla lentamente y progresivamente y es el resultado de una variedad de enfermedades que causan una disminución en la producción de hematíes por la médula ósea o una disminución en la síntesis de hemoglobina, que puede ser hereditaria o adquirida. Este grupo incluye síndromes de insuficiencia medular, anemias carenciales (como ferropenia), anemias secundarias an enfermedades sistémicas (como nefropatías, infecciones crónicas, neoplasias, etc.) y anemias carenciales (Merino et al., 2016).

### *2.2.5. Tipos de anemia*

Existen diferentes tipos de anemia:

*Hemolíticas Inmunes:* En ciertas circunstancias, las inmunoglobulinas o algunos componentes del

complemento se adhieren a la membrana del hematíe, lo que provoca su destrucción temprana. La enfermedad hemolítica del recién nacido es un ejemplo común en la práctica en el que la hemólisis ocurre como resultado de la transferencia pasiva de anticuerpos maternos a los hematíes fetales (Merino et al., 2016).

*Talasemias:* Los síndromes talasémicos son un grupo heterogéneo de anemias hipocromas hereditarias de gravedad variable. El resultado final es la disminución o ausencia de los polipéptidos de las cadenas de la Hb; esta es estructuralmente normal por lo general megaloblásticas (Merino et al., 2016).

*Ferropénicas:* Es el tipo de anemia más prevalente en el mundo. La anemia por déficit de hierro representa un estadio avanzado en el déficit nutricional de hierro. La deficiencia de hierro incluye el déficit absoluto (sin hierro de depósito), el déficit funcional (cuando la demanda de hierro para la eritropoyesis excede el aporte).

### *2.2.6. Ferropenia*

La ferropenia (FeP) consiste en la deficiencia de los depósitos sistémicos de Fe, con potencial efecto nocivo, especialmente en la infancia. Si esta situación se agrava o se mantiene en el tiempo, se desarrollará anemia ferropénica (AFe), con mayor repercusión clínica. La AFe, la enfermedad hematológica más frecuente de la infancia, es la anemia producida por el fracaso de la función hematopoyética medular en la síntesis de Hb debido a la carencia de Fe (Merino et al., 2016).

### *2.2.7. Manifestaciones clínicas*

Según (Sociedad Argentina de Pediatría et al., 2017) las manifestaciones clínicas propias de la anemia ferropénica son:

- Palidez de piel y mucosas
- Decaimiento
- Taquicardia
- Hipotensión arterial
- cefalea
- sensación de mareo y vértigo
- visión nublada

- disminución de la capacidad de concentración
- cansancio precoz
- dolor muscular
- disnea
- hipersensibilidad al frio
- nauseas

### *2.2.8. Diagnóstico*

Según (Sociedad Argentina de Pediatría et al., 2017) el diagnostico debe basarse en:

**Interrogatorio**

Se debe prestar especial atención a estos puntos:

- Tipo de dieta: duración de la lactancia materna y/o de la ingesta de otras leches o fórmulas, ingesta de carne y alimentos ricos en hierro y otros nutrientes (vitaminas C, A y B12, ácido fólico, zinc), volumen de ingesta diaria de leche, exceso de carbohidratos, etc.
- Antecedentes de prematurez, embarazos múltiples y déficit de hierro en la madre.
- Antecedentes de patología perinatal.
- Pérdidas de sangre: color de heces, epistaxis, disnea, hematuria, hemoptisis, etc.
- Trastornos gastrointestinales: diarrea, esteatorrea, etc.
- procedencias geográficas: zonas de parasitosis (uncinariasis) endémicas.
- Suplemento con hierro: cantidad, tiempo, compuesto administrado (sulfato ferroso u otros).
- Trastornos cognitivos: bajo rendimiento escolar, déficit de atención, entre otros.

Según (Sociedad Argentina de Pediatría et al., 2017) el diagnostico debe basarse en:

**Examen físico**

La palidez cutáneo-mucosa es el signo principal y se puede también observar retardo del desarrollo pondoestatural, esplenomegalia leve, telangiectasias, alteración de tejidos epiteliales (uñas, lengua, cabello) y alteraciones óseas.

Según (Sociedad Argentina de Pediatría et al., 2017) el diagnostico debe basarse en:

**Estudios de laboratorio**

- Hemograma
- Hemoglobina y hematocrito: disminuidos.
- Recuento de reticulocitos: normal. Si está aumentado, se deben investigar pérdidas por hemorragia o posibilidad de otro diagnóstico.
- Recuento de plaquetas: normal o elevado.
- Recuento leucocitario: normal.
- Índices hematimétricos:
  - Volumen corpuscular medio (VCM): disminuido. Los valores normales durante la infancia son variables y distintos a los del adulto, por lo que, para definir microcitosis, deben tomarse como referencia los valores mostrados en la Tabla 9. 11,20
  - concentración de hemoglobina corpuscular media (CHCM): disminuida.
  - Amplitud de distribución eritrocitaria (ADE) o red blood cell distribution width (RDW): elevada.
- Morfología eritrocitaria: hipocromía, microcitosis, ovalocitosis, policromatofilia, punteado basófilo (eventualmente).

### *2.2.9. Factor de riesgo*

Según la OMS los factores de riesgo son condiciones, conductas, estilos de vida o situaciones que nos exponen a mayor riesgo de presentar una enfermedad o lesión, es decir, cada una de las características o factores de naturaleza, bilógicos, ambientales, socioculturales, económicos que modifican las posibilidades de contraer una enfermedad.

Un Factor de Riesgo puede ser específico para uno o varios daños (el alcoholismo es causa frecuente de accidentes del tránsito, arrestos policiales, suicidio y disfunción familiar), y a la vez varios Factores de Riesgo pueden incidir para un mismo daño (la obesidad, el sedentarismo, el hábito de fumar y la hiperlipidemia contribuyen a la aparición de Cardiopatía Isquémica. Son aquellas características y atributos (variables) que se presentan asociados diversamente con la enfermedad o el evento estudiado. Los factores de riesgo no son necesariamente las causas, solo sucede que están asociadas con el evento. En epidemiología es toda situación o circunstancia que aumenta la probabilidad de que una persona pueda contraer una enfermedad o algún problema de salud. Hay que diferenciar los factores de riesgo de los factores 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).También Hay factores de riesgo (edad, hipertensión arterial, raza) que cuando aparece la enfermedad son a su vez factores pronóstico (mayor probabilidad de que se desarrolle un evento) (Lema Punín & Inga Miguitama, 2018).

# CAPÍTULO III

## 3. MARCO METODOLÓGICO

### 3.1. Tipo de investigación

Por el método de investigación el presente anteproyecto es de tipo cuantitativo, debido a que se analizará al número de pacientes menores a 5 años diagnosticados con anemia en el Hospital Pediátrico Alfonso Villagómez Román, según el objeto de estudio es aplicado debido a que se aplicaran diferentes análisis para comparar cuál de los dos métodos a emplear es mejor para realizar predicciones, según el nivel de profundización en el objeto de estudio es exploratorio e inferencial debido a que este estudio no ha sido realizado anteriormente, según la manipulación de las variables es no experimental ya que los datos fueron obtenidos mediante historias clínicas de los pacientes del Hospital pediátrico, según la inferencia es inductiva ya que se analizará los modelos logísticos y redes neuronales para predecir y se va a especificar cual método es el mejor, según el periodo temporal es longitudinal ya que se analizaron pacientes pediátricos que han sido diagnosticados en diferentes años calendario.

### 3.2. Diseño de investigación

Se utilizó 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 regresión logística binaria y redes neuronales donde a través de estos se obtiene un modelo matemático que mediante las medidas de bondad de ajuste se determina el "mejor" 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*

El proyecto de investigación planteado se lo llevó a cabo en el Hospital Pediátrico Alfonso Villagómez Román, en Riobamba en las calles Av. José Veloz & España.

### *3.2.2. Población de estudio*

La población de estudio corresponde a los pacientes menores de 5 años con anemia que son atendidos por el Hospital pediátrico Alfonso Villagómez Román.

### *3.2.3. Tamaño de la muestra*

La muestra corresponde a todos los pacientes menores de 5 años diagnosticados con anemia en el Hospital pediátrico Alfonso Villagómez Román.

### *3.2.4. Método de muestreo*

No se aplicó ningún método de muestreo ya que los datos son presentados por las historias clínicas del Hospital pediátrico Alfonso Villagómez Román.

### *3.2.5. Técnicas de recolección de datos*

No se aplicó una técnica de recolección de datos ya que los datos han sido proporcionados por el Hospital pediátrico Alfonso Villagómez Román.

### *3.2.6. Modelo estadístico*

Los modelos logísticos y redes neuronales estarán sujeta a la revisión previa del historial clínico de los pacientes que asisten al hospital pediátrico, de los cuales se obtuvó los datos para las variables independientes, las cuales fueron utilizadas para modelar la anemia a través de las técnicas planteadas.

## 3.3. Identificación de variables

### *3.3.1. Variable dependiente*

- Diagnóstico de alta (anemia)

### *3.3.2. Variables independientes*

- Sexo
- Edad
- Identificación étnica

- Tipo de seguro
- Código cantón
- Cantón
- Peso (Kg)
- Talla (cm)
- Perímetro encefálico
- PCTE_ULT_TALLA_EDAD_Z
- TALLAEDAD
- PCTE_ULT_PESO_EDAD_Z
- PESOEDAD
- PCTE_ULT_IMC_EDAD_Z
- IMC_EDAD
- PCTE_ULT_PESO_LONGTALLA_Z
- PESOTALLA
- ESTADO

### *3.3.3. Operacionalización de variables*

**Tabla 3–1:** Operacionalización de variables

| Variable | Descripción | Tipo | Escala de medición |
|---|---|---|---|
| Diagnóstico de alta | Estado de salud con el que el paciente es egresado del hospital (anemia u otro caso) | Cualitativa Dicotómica | Nominal |
| Sexo | Identidad sexual de los seres vivos, desde el punto de vista biológico (Femenino, Masculino) | Cualitativa Dicotómica | Nominal |

| | | | |
|---|---|---|---|
| Edad | Edad cumplidos en años (menores de 5 años) | Cuantitativa Discreta | Razón |
| Identificación étnica | Se entiende por identidad étnica el sentido de pertenencia e identificación a un determinado grupo étnico (ya sea, blanco, mestizo, indígena) | Cualitativa Politómica | Nominal |
| Tipo de seguro de salud | Registra el seguro que tiene el paciente (IESS, ISSPOL, ISFFA, otro o ninguno) | Cualitativa Politómica | Nomina |
| CODC | Registra el código el cual representa a cada cantón | Cuantitativa Politómica | Nominal |
| Cantón | Es una entidad territorial que subdividen a un municipio, una provincia un departamento u otro tipo de distrito | Cualitativa Politómica | Nominal |
| Peso (Kg) | Se refiere a la masa o el peso del paciente | Cuantitativa Continua | Razón |
| Talla (cm) | Hace referencia a la estatura del paciente | Cuantitativa Continua | Razón |
| Perímetro encefálico | Registra la medida de la cabeza de un niño en su parte más grande | Cuantitativa Continua | Razón |
| PCTE_ULT_TALLA_EDAD_Z | Métricas de evaluación cuantitativa del paciente entre la talla y edad | Cuantitativa Continua | Razón |
| TALLAEDAD | Valoración cualitativa de acuerdo con la valoración métrica y su rango determinado | Cualitativa | Nominal |
| PCTE_ULT_PESO_EDAD_Z | Métricas de evaluación cuantitativa del paciente entre el peso y edad | Cuantitativa Continua | Razón |

| | | | |
|---|---|---|---|
| PESOEDAD | Valoración cualitativa de acuerdo con la valoración métrica y su rango determinado | Cualitativa | Nominal |
| PCTE_ULT_IMC_EDAD_Z | Métricas de evaluación cuantitativa del paciente entre el imc y edad | Cuantitativa Continua | Razón |
| IMC_EDAD | Valoración cualitativa de acuerdo con la valoración métrica y su rango determinado | Cualitativa | Nominal |
| PCTE_ULT_PESO_LONGTA_MATRIZ | Métricas de evaluación cuantitativa del paciente entre el peso y talla | Cuantitativa Continua | Razón |
| PESOTALLA | Valoración cualitativa de acuerdo con la valoración métrica y su rango determinado | Cualitativa | Nominal |
| ESTADO | Si el paciente esta con estado presuntivo o definitivo en anemia | Cualitativa | Nominal |

**Fuente:**Matriz de datos del HPAVR 2020-2021

**Realizado por:**Sánchez Evelyn y Tenesaca Sebastian, 2023

### *3.3.4. Codificación de las variables*

**Tabla 3–2:** Codificación de variables

| Variable | Codificación |
|---|---|
| Sexo | Mujer = 0<br>Hombre = 1 |
| Etnia | Blanco = 0<br>Indígena = 1<br>Mestizo = 2<br>Mulato = 3<br>No aplica = 4 |

| | |
|---|---|
| Seguro | No aporta = 0<br>Aporta = 1<br>Desconocido = 2 |
| Cantón | Otro = 0<br>Riobamba = 1 |
| Código cantón | Otro = 0<br>Riobamba = 1 |
| TALLAEDAD | Otro = 0<br>Normal = 1 |
| PESOEDAD | Otro = 0<br>Normal = 1 |
| IMCEDAD | Otro = 0<br>Normal = 1 |
| PESOTALLA | Otro = 0<br>Normal = 1 |
| Estado | Presuntivo = 0<br>Definitivo = 1 |

**Fuente:** Matriz de datos del HPAVR 2020-2021

**Realizado por:** Sánchez Evelyn y Tenesaca Sebastian, 2023

# CAPÍTULO IV

## 4. RESULTADOS Y DISCUSIÓN

El análisis estadístico del conjunto de datos recopilado en el Hospital Pediátrico Alfonso Villagómez Román (HPAVR) de Riobamba durante el período comprendido entre 2020 y 2021 sirve como base para el presente informe. Este estudio examina a 2287 individuos o pacientes pediátricos de 0 a 5 años que recibieron atención en el servicio de consulta externa del hospital bajo el diagnostico presuntivo y definitivo de Anemia. Para ello debemos tomar en cuenta los cuadros de clasificación por percentiles donde se demonina si un menor tienen relaciones (bajas, normal, altas) entre su talla, peso y edad.

| Talla Edad | |
|---|---|
| <=-3 | Baja talla severa |
| >-3 y <-2 | Baja talla |
| -2 | Normal con intervención inmediata |
| >-2 y <-1.5 | Normal con seguimiento |
| >=-1.5 y <=2 | Normal |
| >2 | Talla alta para edad |

| Peso edad | |
|---|---|
| <=-3 | Baja peso severo |
| >-3 y <-2 | Bajo peso |
| -2 | Normal con intervención inmediata |
| >-2 y <-1.5 | Normal con seguimiento |
| >=-1.5 y <=2 | Normal |
| >2 | Peso alto para edad |

| IMC edad | |
|---|---|
| <=-3 | Severamente emaciados |
| >-3 y <-2 | Emaciados |
| >-2 y <2 | Normal con seguimiento/ Normal |
| >2 | Sobrepeso |
| >3 | Obesidad |

| Peso Talla | |
|---|---|
| <=-3 | Severamente emaciados |
| >-3 y <-2 | Emaciados |
| >-2 y <2 | Normal con seguimiento/ Normal |
| >2 | Sobrepeso |
| >3 | Obesidad |

**Ilustración 4–1:** Clasificación por percentiles relacionales

**Realizado por:** Sánchez Evelyn y Tenesaca Sebastian, 2023.

### 4.1. Análisis exploratorio univariado

En este estudio, se utilizaron 20 variables en total, que se dividieron equitativamente en 10 variables cuantitativas y 10 variables cualitativas. El análisis descriptivo para observar el comportamiento de las variables se muestra a continuación. Estas variables capturaron varios aspectos relevantes para el análisis estadístico y permitieron obtener una visión completa de los pacientes pediátricos atendidos en el servicio de emergencia HPAVR.

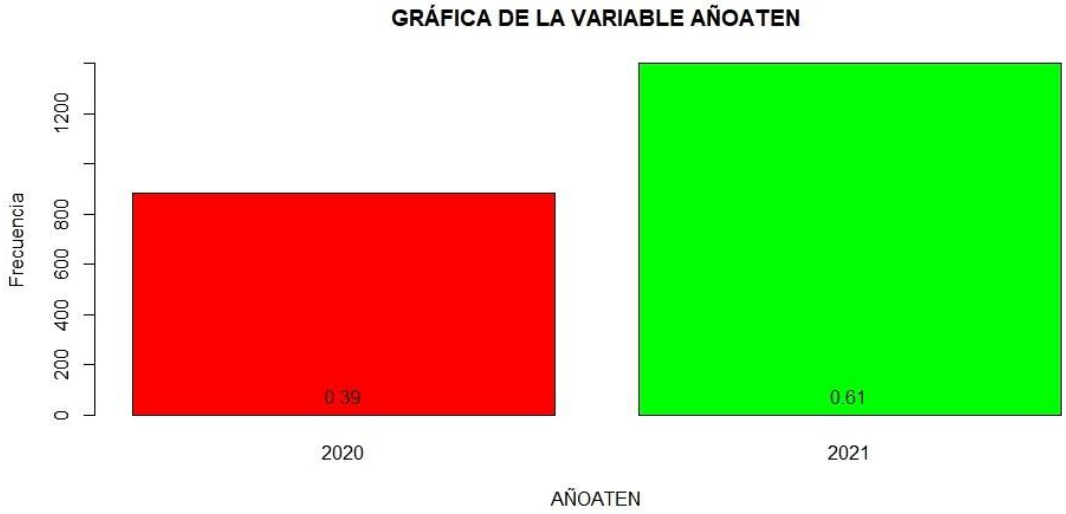



**Ilustración 4–2:** Distribución de la variable AÑOATEN

**Realizado por:** Sánchez Evelyn y Tenesaca Sebastian, 2023.

En el Hospital Pediátrico Alfonso Villagómez Román (HPAVR) de Riobamba se han registrado 2287 observaciones de pacientes atendidos, según la gráfica de barras. La variable examinada es AÑOATEN, que indica el año en que los pacientes recibieron atención en un centro de atención externa. Como se muestra en la gráfica, el 39 % de las observaciones se remontan a 2020, mientras que el 0,61 % se remonta a 2021. La pandemia de COVID-19 tuvo un impacto en la demanda de atención médica pediátrica, lo que puede causar una distribución desigual entre los años.


**Ilustración 4–3:** Distribución de la variable SEXO

**Realizado por:** Sánchez Evelyn y Tenesaca Sebastian, 2023.

En la gráfica se evidenció que el 53% de las observaciones pertenecen a pacientes masculinos,

mientras que el 47% restante pertenece a pacientes femeninos. Se evidencia leve atención mayoritaria en pacientes Hombres que en Mujeres.

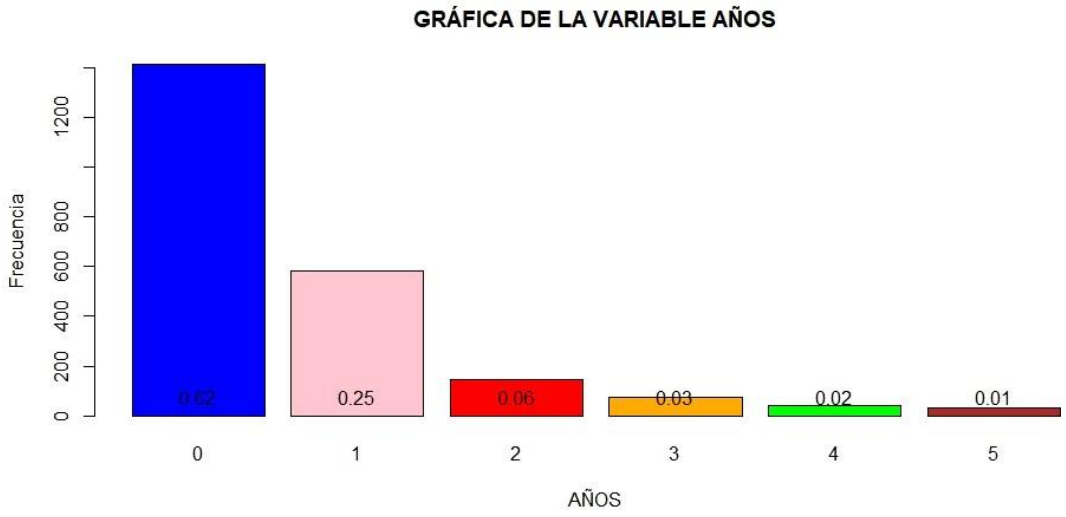



**Ilustración 4–4:** Distribución de la variable AÑOS

**Realizado por:** Sánchez Evelyn y Tenesaca Sebastian, 2023.

Con respecto a la variable AÑOS que se refiere a la edad de los infantes, se encontró que la mayoría están entre los 0 y 1 años, al ser un hospital pediátrico es normal que los pacientes atendidos sean recién nacidos e infantes en su primera etapa de crecimiento.

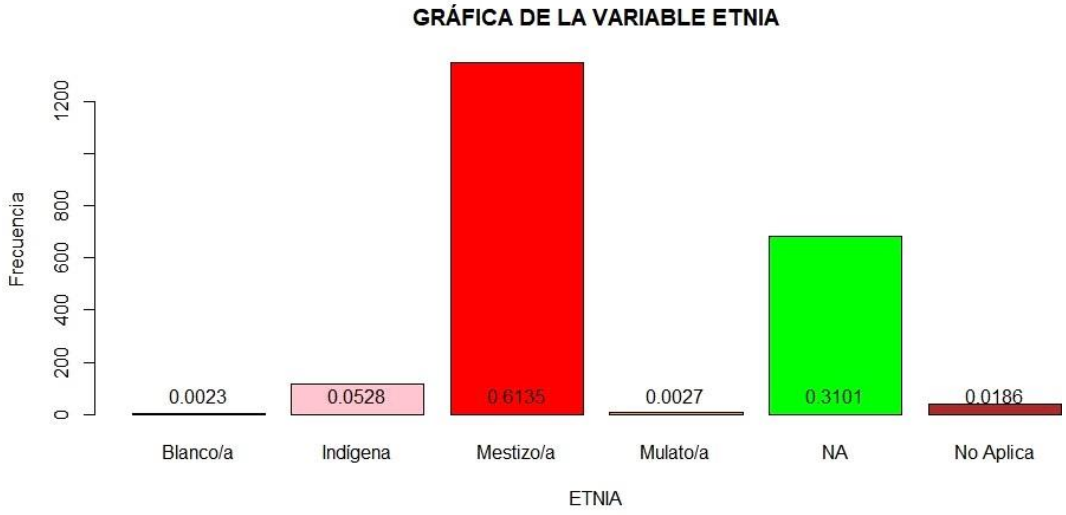



**Ilustración 4–5:** Distribución de la variable ETNIA

**Realizado por:** Sánchez Evelyn y Tenesaca Sebastian, 2023.

Según los datos estudiados muestran gran atención a pacientes que se autoidentifican como Mestizos/as, sin embargo, la métrica que le sigue es del 31% evidencia que no se tiene datos

al respecto de estos pacientes. Mientras que, solo pocos de ellos se autoidentifican como Blancos/as, Indígenas o Mulatos y los No aplica corresponde a pacientes extranjeros generalmente.

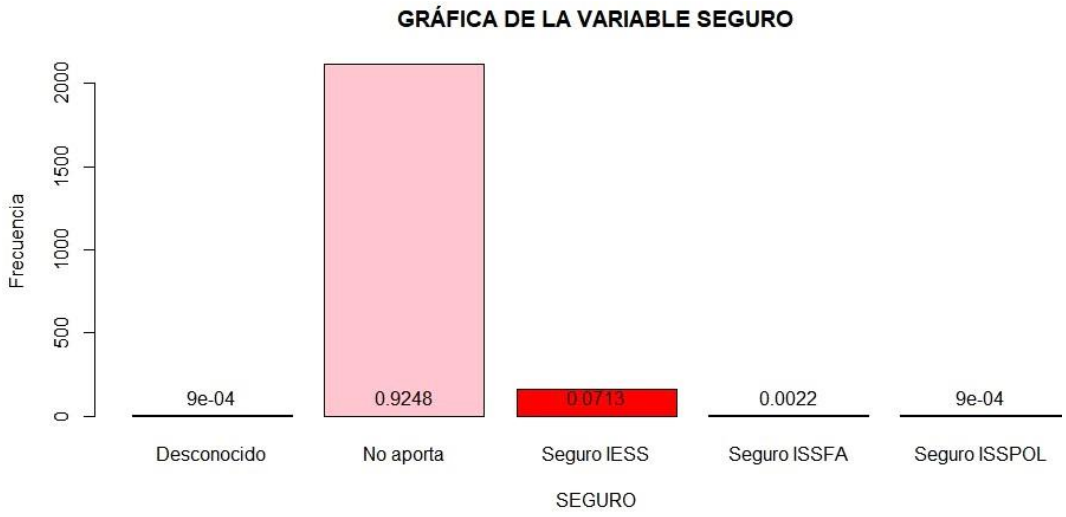



**Ilustración 4–6:** Distribución de la variable SEGURO

**Realizado por:** Sánchez Evelyn y Tenesaca Sebastian, 2023.

En esta grafica de la variable SEGURO se evidencio que la gran mayoría de los pacientes con el 92 % no aportan a un seguro, es lo más frecuente ya que al ser un hospital de salud pública los que cuentan con seguro como IESS, ISSPOL, ISSFA u otro son pocos los referidos en consulta externa. Puede ser un factor importante en el ámbito económico, ya que los infantes provenientes de familias no afiliadas tienen algún tipo de anemia.

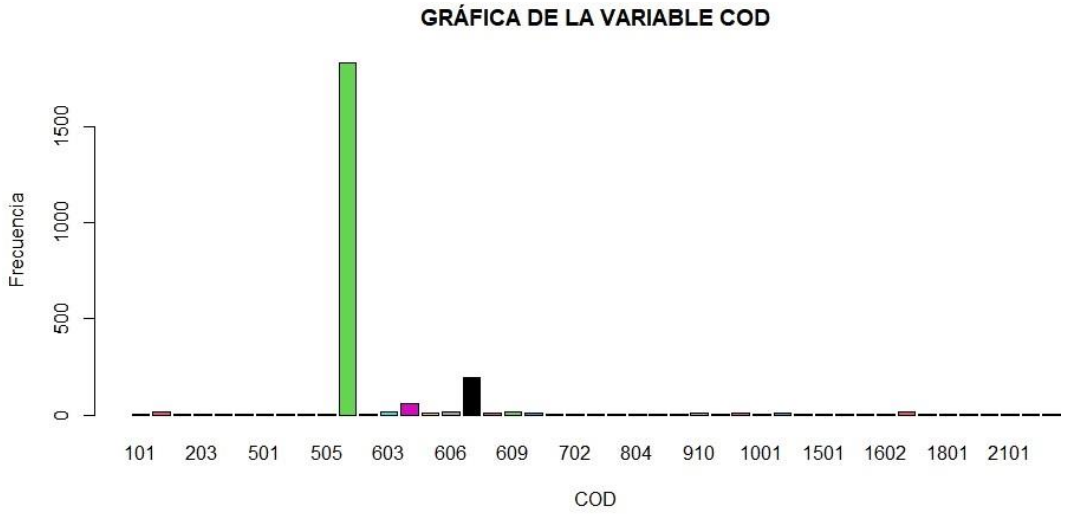



**Ilustración 4–7:** Distribución de la variable COD

**Realizado por:** Sánchez Evelyn y Tenesaca Sebastian, 2023.

La variable COD hace referencia al código del lugar de residencia de los pacientes pediátricos tenemos gran proporción de datos el código 601 que hace referencia al cantón Riobamba con el 80 % se observó que los demás códigos están repartidos en cantones de varias provincias del Ecuador, esto se debe a que en el 2020 por el COVID-19 varias familias tuvieron que trasladarse a diferentes partes del país para sobrellevar la situación sanitaria por lo cual son pocos los datos de estos pacientes o que a su vez son pacientes referidos de otros centros de salud que no tienen las especialidades necesarias para atender adecuadamente a los infantes

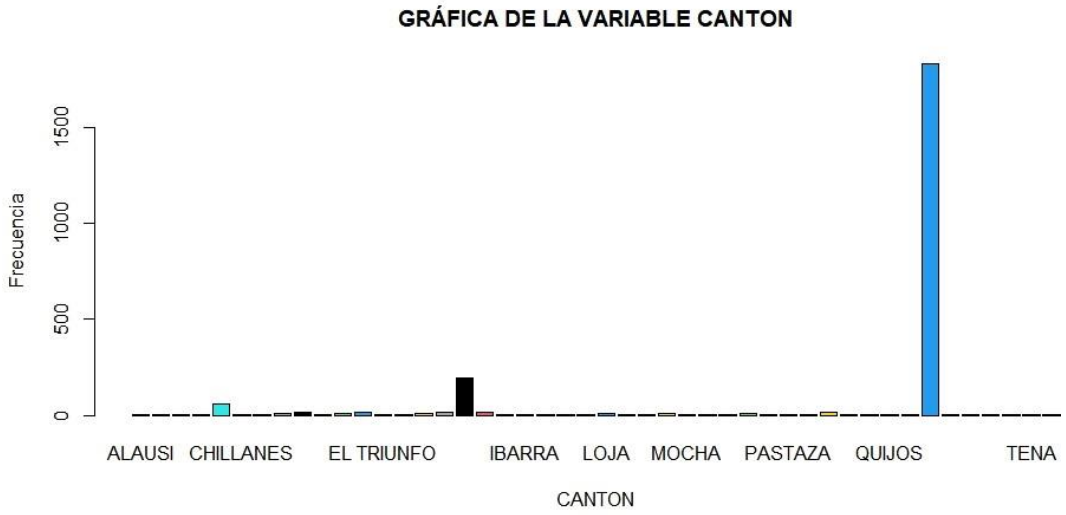



**Ilustración 4–8:** Distribución de la variable CANTÓN

**Realizado por:** Sánchez Evelyn y Tenesaca Sebastian, 2023.

La variable CANTÓN hace referencia a lo mismo que la variable COD, sin embargo, en esta tenemos los nombres de los cantones donde sé evidencia que Guano es otra de los indicadores levemente grande en comparación a los demás.

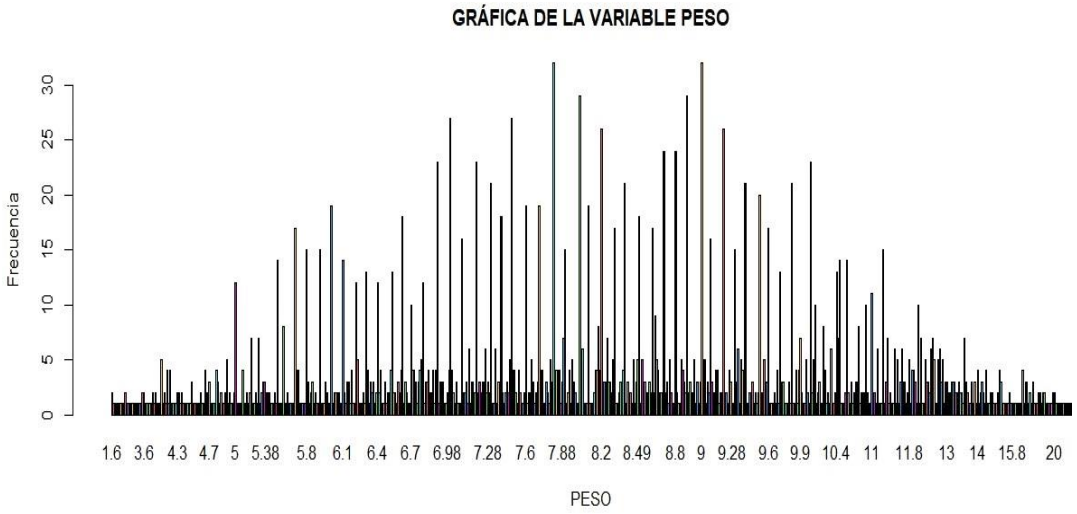



**Ilustración 4–9:** Distribución de la variable PESO

**Realizado por:** Sánchez Evelyn y Tenesaca Sebastian, 2023.

En la frecuencia de la variable PESO se observó métricas desde 1.6 kg hasta 78 kg de los infantes, siendo 7.8 la más frecuente. Sin embargo, se evidencia mala digitación de datos ya que existieron pacientes de 0 años tabulados con un peso de más de 70 kg o más, por otro lado, no se puede definir si estos pesos son adecuados o no ya que todo depende de la edad de los pacientes, pero si nos basamos que los niños entre 0 y 1 años son atendidos con mayor frecuencia los pesos si son adecuados.

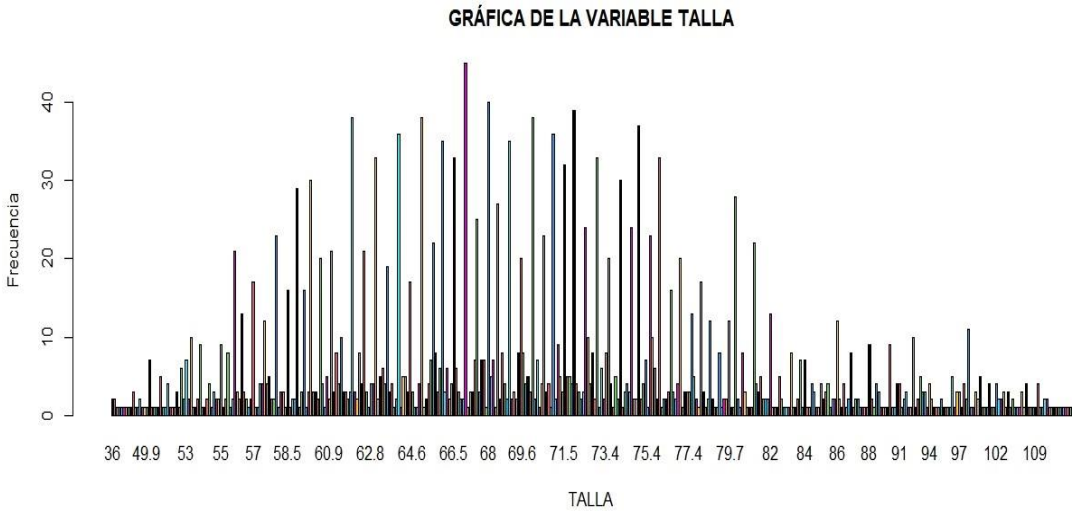



**Ilustración 4–10:** Distribución de la variable TALLA

**Realizado por:** Sánchez Evelyn y Tenesaca Sebastian, 2023.

En el análisis de la variable TALLA tenemos métricas desde los 36 cm hasta los 125 cm. Existe

gran variedad de datos en esta variable ya que tenemos 2287 individuos analizados y al igual que el peso no se puede definir si las métricas son adecuadas o no ya que todo depende de la edad de los pacientes, pero si nos basamos que los niños entre 0 y 1 años son atendidos con mayor frecuencia las tallas si son adecuadas.

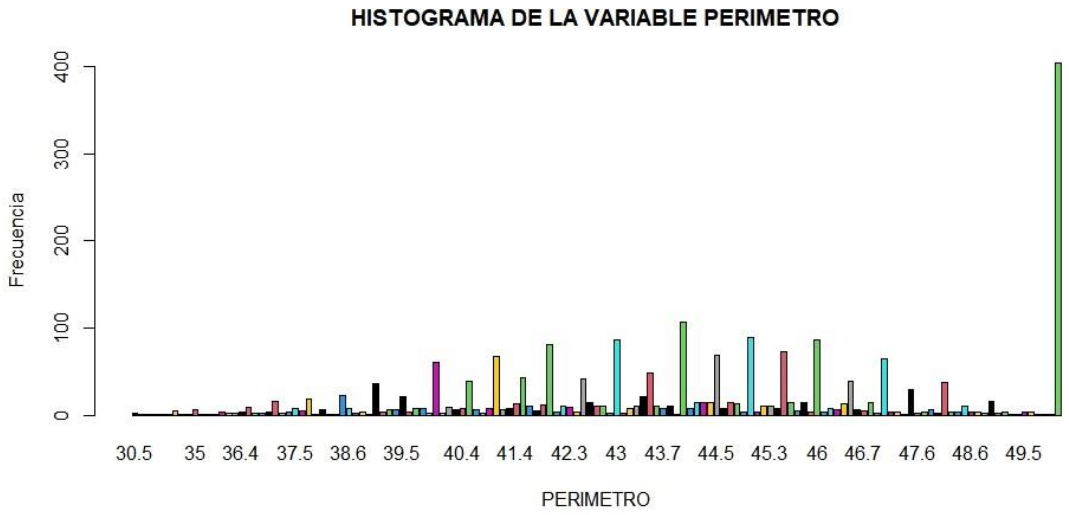



**Ilustración 4–11:** Distribución de la variable PERIMETRO

**Realizado por:** Sánchez Evelyn y Tenesaca Sebastian, 2023.

En la variable PERIMETRO hace referencia a el perímetro cefálico que se utiliza para evaluar el tamaño del cráneo y detectar posibles variaciones en el crecimiento, se evidencio métricas desde los 30.5 cm hasta los 51.8 cm lo cual es adecuado para niños de 0 a 1 años. Sin embargo, se encontró mayoría de NA es decir datos faltantes.

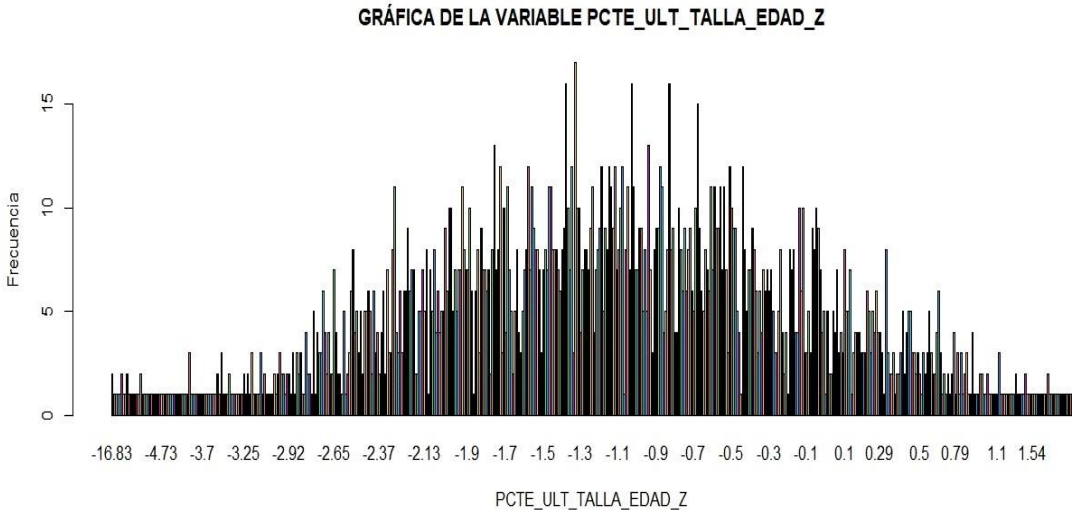



**Ilustración 4–12:** Distribución de la variable PCTE_ULT_TALLA_EDAD_Z

**Realizado por:** Sánchez Evelyn y Tenesaca Sebastian, 2023.

Con respecto a la variable PCTE_ULT_TALLA_EDAD_Z hace referencia a la métrica de los percentiles es la relación talla y edad. A los pacientes que tienen como percentil por debajo de -3 son pacientes con baja talla severa para su edad, los pacientes con métricas -3 y -2.01 son pacientes con baja talla para su edad, los que tienen -2 exactos se consideran normal, pero con intervención inmediata y los que están con métricas de -1.5 hasta -1.99 se consideran normal, pero con Seguimiento. Si un paciente esta entre -1.49 y 2 se considera normal, pero si el paciente está por encima de 2 es un paciente con alta talla para su edad. En la gráfica se evidencia que la mayoría entra en un rango normal, lo cual se demostró en la gráfica 4-12.

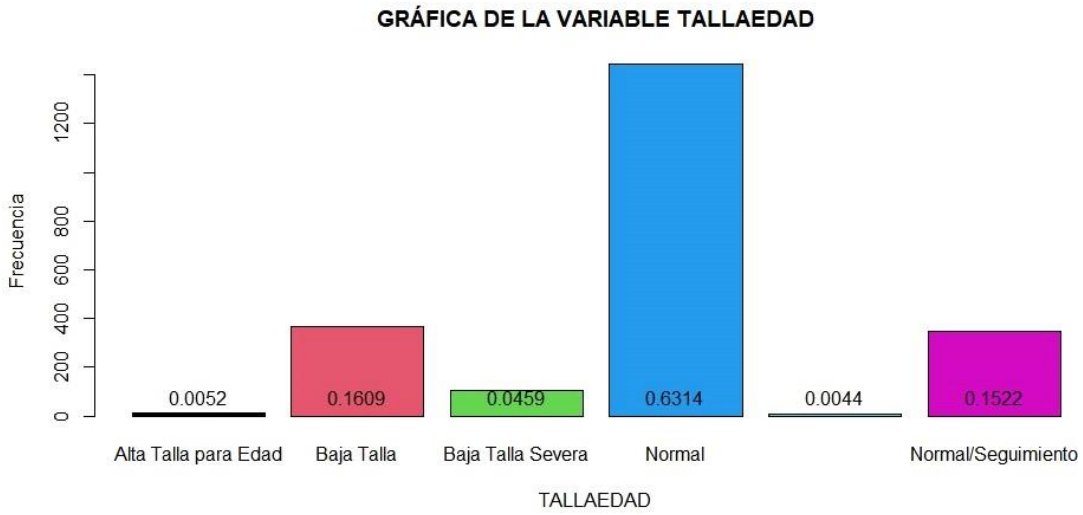



**Ilustración 4–13:** Distribución de la variable TALLAEDAD

**Realizado por:** Sánchez Evelyn y Tenesaca Sebastian, 2023.

La variable TALLAEDAD hace referencia a la categorización de las métricas de la variable PCTE_ULT_TALLA_EDAD_Z, la cual identifica en qué nivel se encuentra el infante. Se evidencia que el 63 % de los pacientes tienen una talla en relación con su edad normal, el 16 % esta con una condición de baja talla para su edad y tan solo el 4% se encontró con baja talla severa.

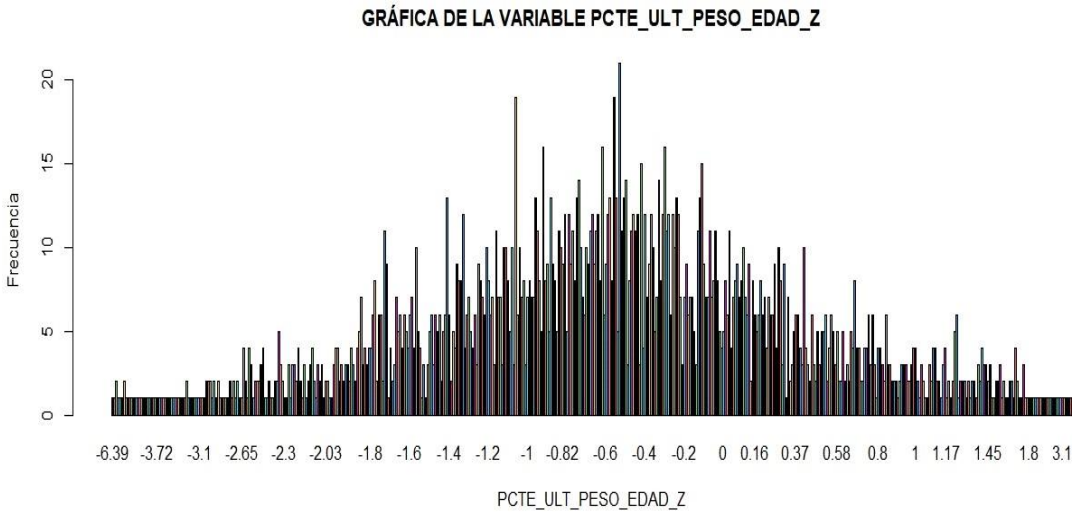



**Ilustración 4–14:** Distribución de la variable PCTE_ULT_PESO_EDAD_Z

**Realizado por:** Sánchez Evelyn y Tenesaca Sebastian, 2023.

Gracias a este análisis de la variable PCTE_ULT_PESO_EDAD_Z hace referencia a la métrica de los percentiles es la relación peso y edad. A los pacientes que tienen como percentil por debajo de -3 son pacientes con bajo peso severo para su edad, los pacientes con métricas -3 y -2 son pacientes con bajo peso para su edad y los que están con métricas de -1.5 hasta -1.99 se consideran normal, pero con seguimiento. Si un paciente esta entre -1.49 y 2 se considera normal, pero si el paciente está por encima de 2 es un paciente con peso elevado para su edad. En la gráfica se evidencia que la mayoría entra en un rango normal, lo cual se demostró en la gráfica 4-14.

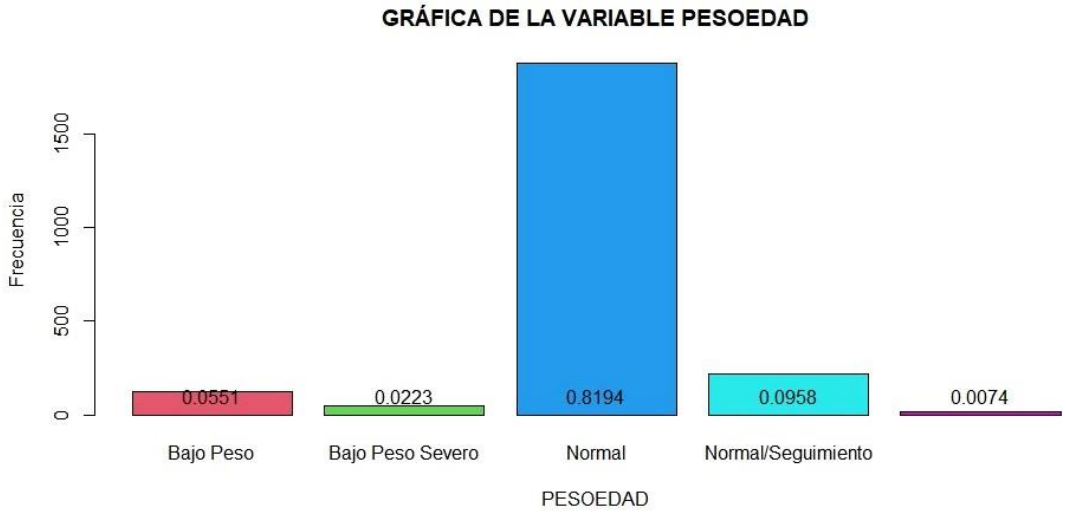



**Ilustración 4–15:** Distribución de la variable PESOEDAD

**Realizado por:** Sánchez Evelyn y Tenesaca Sebastian, 2023.

La variable PESOEDAD hace referencia a la categorización de las métricas de la variable PCTE_ULT_PESO_EDAD_Z, la cual identifica en qué nivel se encuentra el infante. Se evidencia que el 82 % de los pacientes tienen un peso en relación con su edad normal, el 6 % esta con una condición de bajo peso para su edad y tan solo el 2% se encuentro que están con bajo peso severo.

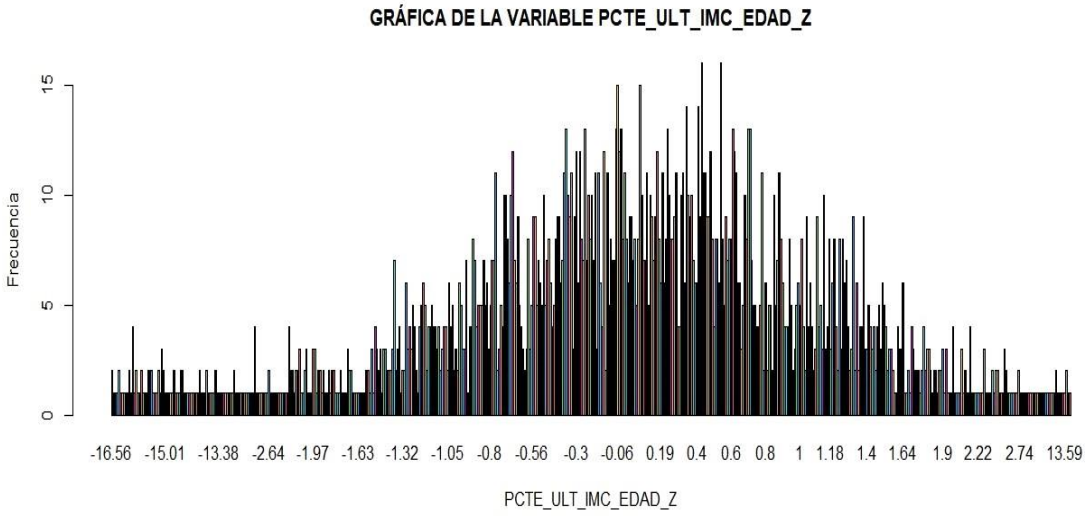



**Ilustración 4–16:** Distribución de la variable PCTE_ULT_IMC_EDAD_Z

**Realizado por:** Sánchez Evelyn y Tenesaca Sebastian, 2023.

La variable PCTE_ULT_IMC_EDAD_Z hace referencia a la métrica de los percentiles es la relación peso y edad. A los pacientes que tienen como percentil por debajo de -3 son pacientes severamente emaciados para su edad, los pacientes con métricas -3 y -2 son pacientes emaciados para su edad y los que están con métricas de -2 hasta 2 se consideran normal, algunos con seguimiento. Si el paciente está entre 2 a 3 es un paciente con sobrepeso para su edad y si está por encima de 3 se considera obesidad. En la gráfica se evidencia que la mayoría entra en un rango normal, lo cual se demostró en la gráfica 4-16.

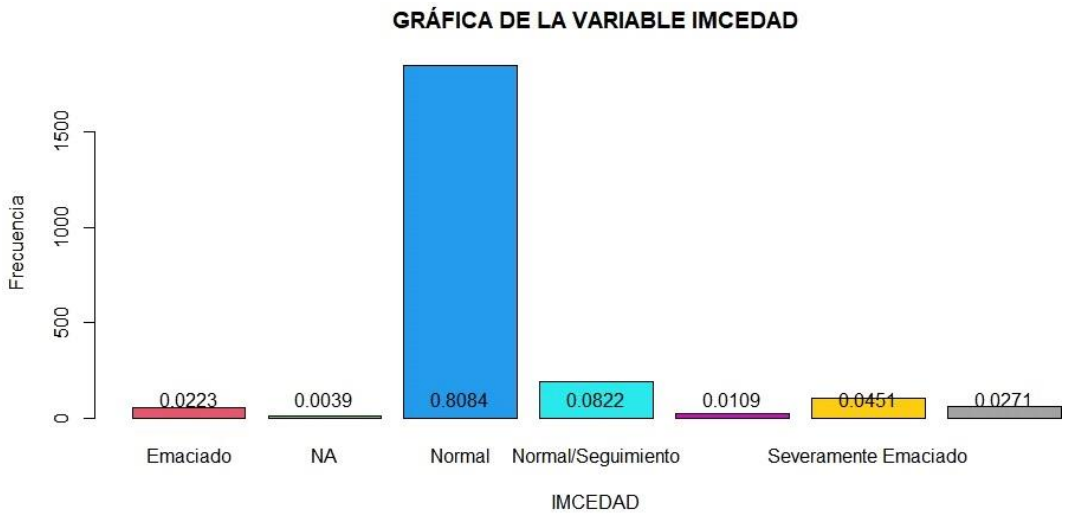



**Ilustración 4–17:** Distribución de la variable IMCEDAD

**Realizado por:** Sánchez Evelyn y Tenesaca Sebastian, 2023.

La variable IMCEDAD hace referencia a la categorización de las métricas de la variable PCTE_ULT_IMC_EDAD_Z, la cual identifica en qué nivel se encuentra el infante. Se evidencia que el 81 % de los pacientes tienen un IMC en relación con su edad normal, el 8 % esta con una condición de normal, pero con seguimiento y tan solo el 2% se encuentro que están emaciados y el 4% severamente emaciados, se encontraron también datos no existentes o NA.

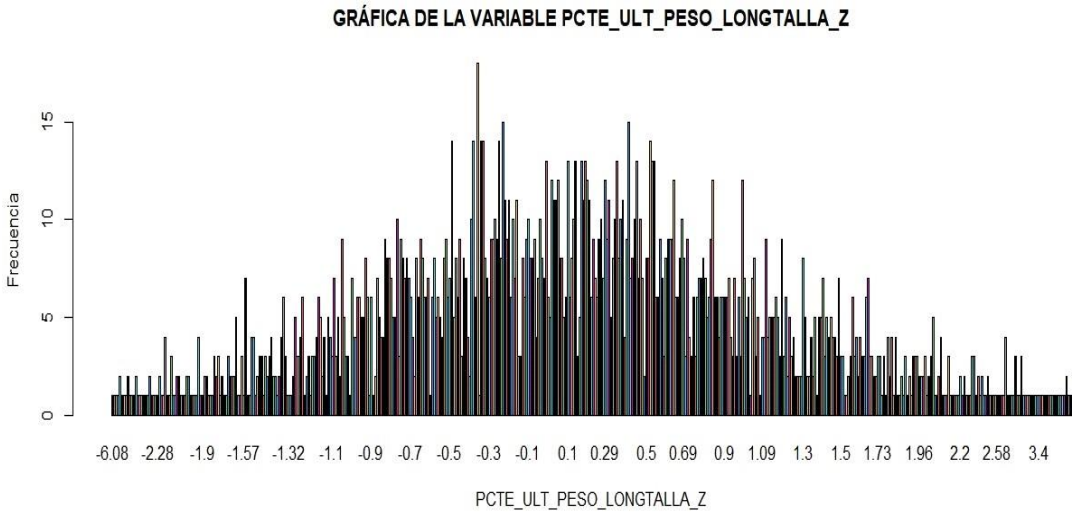



**Ilustración 4–18:** Distribución de la variable PCTE_ULT_PESO_LONGTALLA_Z

**Realizado por:** Sánchez Evelyn y Tenesaca Sebastian, 2023.

Gracias a la gráfica de la variable PCTE_ULT_PESO_LONGTALLA_Z hace referencia a la métrica de los percentiles es la relación peso y talla. A los pacientes que tienen como percentil por debajo

de -3 son pacientes severamente emaciados para su edad, los pacientes con métricas -2.99 y -2.01 son pacientes emaciados para su edad y 2 normal, pero con intervención inmediata. Los que están con métricas de -1.99 hasta -1.5 y de 1.5 a 1.99 se consideran normal, pero con seguimiento, a los que están entre -1.49 y 1.49 se considera normal. Si el paciente está entre 2 a 3 es un paciente con sobrepeso para su edad y si está por encima de 3 se considera obesidad. En la gráfica se evidencia que la mayoría entra en un rango normal, lo cual se demostró en la gráfica 4-18.

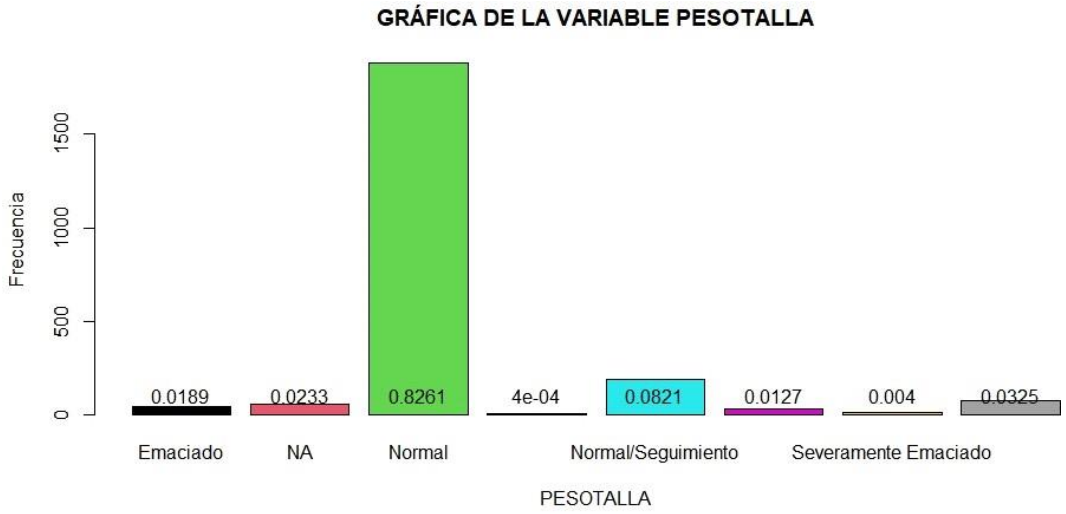



**Ilustración 4–19:** Distribución de la variable PESOTALLA

**Realizado por:** Sánchez Evelyn y Tenesaca Sebastian, 2023.

La variable PESOTALLA hace referencia a la categorización de las métricas de la variable PCTE_ULT_PESO_LONGTALLA_Z, la cual identifica en qué nivel se encuentra el infante. Se evidencia que el 83 % de los pacientes tienen un peso en relación con su talla normal, el 8 % esta con una condición de normal, pero con seguimiento y tan solo el 2 % se encuentro que están emaciados y se encontraron también datos no existentes o NA.

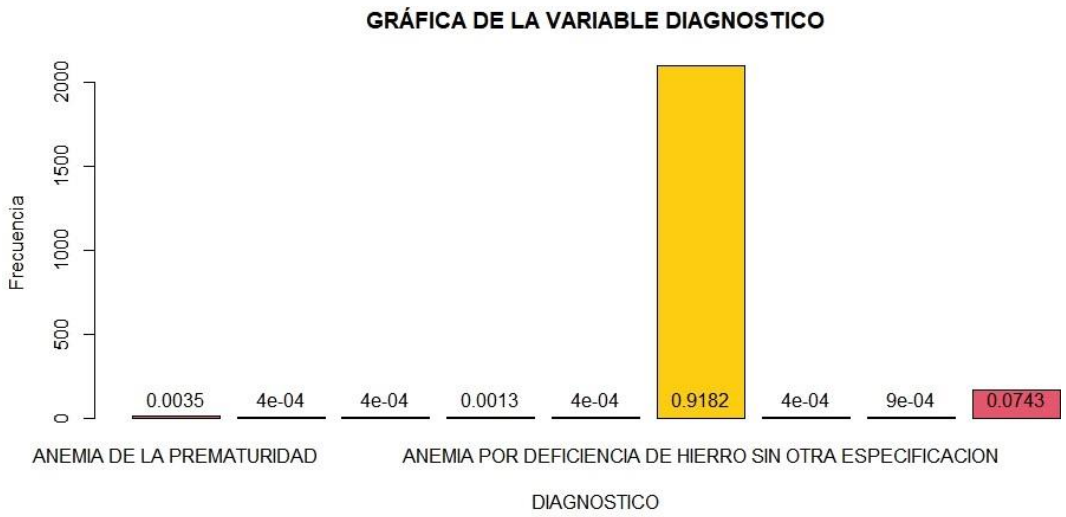



**Ilustración 4–20:** Distribución de la variable DIAGNOSTICO

**Realizado por:** Sánchez Evelyn y Tenesaca Sebastian, 2023.

noindent DIAGNOSTICO hace referencia al tipo de anemia que tiene el paciente, se evidencia que el 92 % de los infantes padecen de anemia por deficiencia de hierro si otra especificación. Mientras que el 8 % se reparte en las demás anemias como: anemia de la prematuridad, anemia hemolítica (adquirida, hereditaria y autoinmune), anemia no especificada y anemia refractaria.

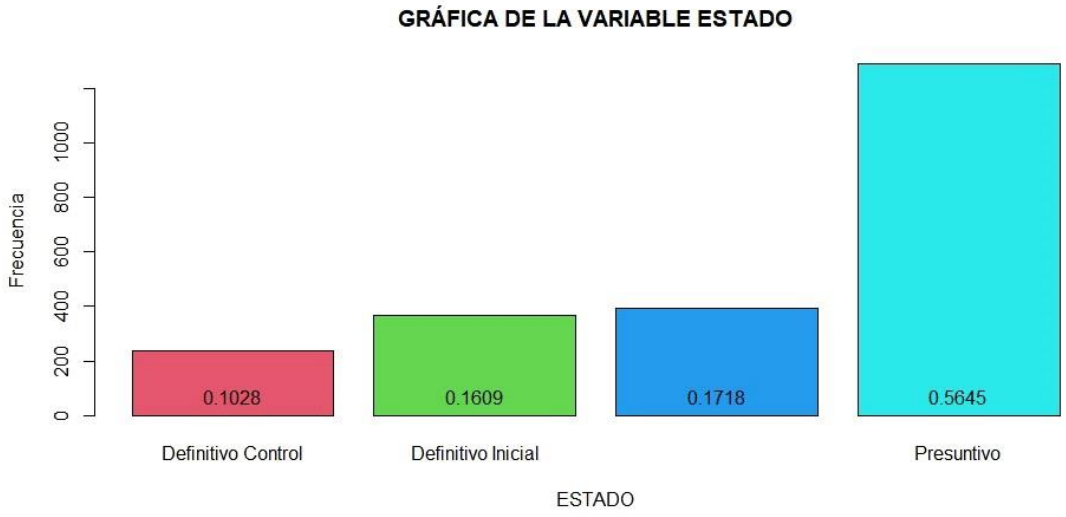



**Ilustración 4–21:** Distribución de la variable ESTADO

**Realizado por:** Sánchez Evelyn y Tenesaca Sebastian, 2023.

La variable ESTADO se refiere al estado concluyente del paciente, es decir, de la variable anterior si su diagnóstico es presuntivo (que solo se deduce más no se afirma) y el definitivo (que fue diagnosticado con pruebas de laboratorio o en otros controles anteriores), se observa que el

predominante con el 56 % es presuntivo y el restante ya es definitivo ya sea de control, inicial o por confirmación de laboratorio.

## 4.2. Depuración de la base de datos

Para la depuración de la base de datos, primero se realizó el rellenado de datos faltantes con la media con un filtro por edades (Imputación por grupos). Rellenar los valores faltantes con este método puede ayudar a preservar la distribución general de los datos y evitar distorsiones significativas en la media y la varianza. Si los valores faltantes no son numerosos y no representan una parte significativa de los datos, rellenarlos con la media puede ser preferible a eliminar las filas con datos faltantes, lo que podría resultar en la pérdida de información valiosa y simplificar el análisis, ya que no es necesario implementar técnicas más complejas para manejar los valores faltantes, como la imputación múltiple o el uso de modelos de imputación.
A continuación, se realizó la eliminación de datos atípicos de la siguiente manera:

- Se calcularon las medias de las variables numéricas filtradas por edad y se almacenaron en *medias*.
- Se reemplazaron los valores faltantes (NA) por las medias correspondientes en las variables numéricas.
- Se aplicó un filtro para eliminar datos atípicos, manteniendo solo los valores que se encuentran dentro del rango "media - 4 * desviación estándar" "media + 4 * desviación estándar." en cada columna numérica.

Los datos que estaban fuera de este rango, es decir, los datos atípicos, fueron eliminados del dataframe *datos*. Como resultado, se quedaron 2190 pacientes en el dataframe después de eliminar los datos atípicos.

## 4.3. Técnicas de modelado

### *4.3.1. Modelo de clasificación: regresión logística binaria*

El objetivo de la regresión logística binaria es proporcionar un modelo estadístico que explique la relación entre una variable binaria de respuesta (Anemia presuntiva o definitiva) y las variables predictoras; esto permite la predicción de la probabilidad de éxito (1) en función de las variables predictoras. Se utiliza para una variedad de propósitos, incluida la clasificación de pacientes en grupos de riesgo y la investigación de factores de riesgo.

En este caso los modelos de regresión logística se ajustan utilizando la función glm en R. Esta función permite seleccionar el tipo de distribución y la función de enlace apropiados para el análisis. En este caso, se utilizan la distribución binomial y la función de enlace logit (logaritmo de las probabilidades) mediante el argumento en la función "glm", "familia = "binomial"(enlace = "logit").

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)               -1.93998    1.70933  -1.135 0.256402
SEXO                       0.03767    0.09676   0.389 0.697052
AÑOS                      -0.24869    0.14227  -1.748 0.080467 .
ETNIA                     -0.12781    0.19563  -0.653 0.513563
SEGURO                    -0.49059    0.17292  -2.837 0.004552 **
CODC                       0.10224    0.12400   0.825 0.409638
CANTON                     0.02898    0.11961   0.242 0.808543
PESO                      -0.12242    0.17635  -0.694 0.487564
TALLA                      0.07200    0.04432   1.625 0.104203
PERIMETRO                 -0.05149    0.02747  -1.874 0.060902 .
PCTE_ULT_TALLA_EDAD_Z     -0.52290    0.15561  -3.360 0.000779 ***
TALLAEDAD                  0.02369    0.14651   0.162 0.871542
PCTE_ULT_PESO_EDAD_Z       0.63466    0.24990   2.540 0.011096 *
PESOEDAD                  -0.40690    0.20806  -1.956 0.050506 .
PCTE_ULT_IMC_EDAD_Z       -0.01439    0.02141  -0.672 0.501459
IMCEDAD                    0.05285    0.26575   0.199 0.842362
PCTE_ULT_PESO_LONGTALLA_Z -0.32791    0.17369  -1.888 0.059039 .
PESOTALLA                  0.30307    0.26171   1.158 0.246852
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2998.1  on 2190  degrees of freedom
Residual deviance: 2941.6  on 2173  degrees of freedom
AIC: 2977.6

Number of Fisher Scoring iterations: 4
```

**Ilustración 4–22:** Variables significativas del modelo I de regresión lineal en el software R

**Realizado por:** Sánchez Evelyn y Tenesaca Sebastian, 2023.

En el resultado del primer modelo nos arroja que de las variables seleccionadas las que son significativas son: AÑOS, SEGURO, PERIMETRO, PCTE_ULT_TALLA_EDAD_Z, PCTE_ULT_PESO_EDAD_Z, PESOEDAD PCTE_ULT_PESO_LONGTALLA_Z. Esto se definió con el valor p asociado a cada coeficiente estimado se utiliza para evaluar la significancia estadística. Se considera que el coeficiente es estadísticamente significativo si el valor p es menor que un umbral predeterminado, en este caso de (Signif. Codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1.) como se puede observar nos arroja un Criterio de Información de Akaike (AIC: 2977.6) que sé comparo con los demás modelos al eliminar los factores o variables no significativas.

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -0.327591   0.945999  -0.346  0.72912
AÑOS                      0.148506   0.053012   2.801  0.00509 **
SEGURO                   -0.472220   0.170939  -2.763  0.00574 **
PERIMETRO                 0.004036   0.021210   0.190  0.84910
PCTE_ULT_TALLA_EDAD_Z    -0.421316   0.142497  -2.957  0.00311 **
PCTE_ULT_PESO_EDAD_Z      0.603280   0.226842   2.659  0.00783 **
PESOEDAD                 -0.246234   0.186087  -1.323  0.18576
PCTE_ULT_PESO_LONGTALLA_Z -0.444883   0.158313  -2.810  0.00495 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2998.1  on 2190  degrees of freedom
Residual deviance: 2956.4  on 2183  degrees of freedom
AIC: 2972.4

Number of Fisher Scoring iterations: 4
```

**Ilustración 4–23:** Variables significativas del modelo II de regresión lineal en el software R

**Realizado por:** Sánchez Evelyn y Tenesaca Sebastian, 2023.

El segundo modelo II nos arroja que ya solo quedan dos variables no significativas que son PERIMETRO Y PESOEDAD, al igual que el primer modelo que se examinó se determina que tienen un Criterio de Información de Akaike (AIC: 2972.4) que se evaluó con otro modelo sacando las variables sin relieve para el estudio.

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -0.40087    0.06894  -5.815 6.06e-09 ***
AÑOS                      0.14985    0.05275   2.841 0.004502 **
SEGURO                   -0.45689    0.17028  -2.683 0.007292 **
PCTE_ULT_TALLA_EDAD_Z    -0.42137    0.11932  -3.531 0.000413 ***
PCTE_ULT_PESO_EDAD_Z      0.57164    0.17731   3.224 0.001264 **
PCTE_ULT_PESO_LONGTALLA_Z -0.44191    0.12795  -3.454 0.000553 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2998.1  on 2190  degrees of freedom
Residual deviance: 2958.2  on 2185  degrees of freedom
AIC: 2970.2

Number of Fisher Scoring iterations: 4
```

**Ilustración 4–24:** Variables significativas del modelo III de regresión lineal en el software R

**Realizado por:** Sánchez Evelyn y Tenesaca Sebastian, 2023.

En el tercer modelo evaluado los factores analizados son todos significativos con un Criterio de

Información de Akaike (AIC: 2970.2). Dado el conjunto de variables predictoras consideradas, el modelo con el menor valor de AIC es el que proporciona el mejor ajuste. Teniendo en cuenta la cantidad de parámetros estimados, un valor de AIC más bajo indica que el modelo logra un mejor ajuste a los datos. Esto significa que el modelo III tiene el menor AIC es el más parsimonioso, lo que significa que puede explicar mejor los datos con el menor número de parámetros.

$$odds = e^{-0.40087} * e^{0.14985X_1} * e^{-0.45689X_2} * e^{-0.42137X_3} * e^{0.57164X_4} * e^{-0.44191X_5}$$

Donde:

**Tabla 4–1:** Variables $X_i$

| $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ |
|---|---|---|---|---|
| AÑOS | SEGURO | PCTE_ULT_ TALLA_EDAD_Z | PCTE_ULT_ PESO_EDAD_Z | PCTE_ULT_PESO_ LONGTALLA_Z |

**Fuente**: Matriz de datos del HPAVR 2020-2021

**Realizado por**: Sánchez Evelyn y Tenesaca Sebastian, 2023

### *4.3.2. Modelo de redes neuronales: retropropagación mejorada (Rprop+)*

Los modelos de redes neuronales se utilizan principalmente para aprender a reconocer patrones y usar los datos para hacer predicciones o tomar decisiones. Estos modelos se basan en el funcionamiento del cerebro humano y en un conjunto interconectado de unidades llamadas neuronas artificiales. Durante el proceso de entrenamiento, las redes neuronales aprenden a ajustar los pesos y las conexiones entre las neuronas para ajustarse a los datos de entrada y producir la salida deseada.

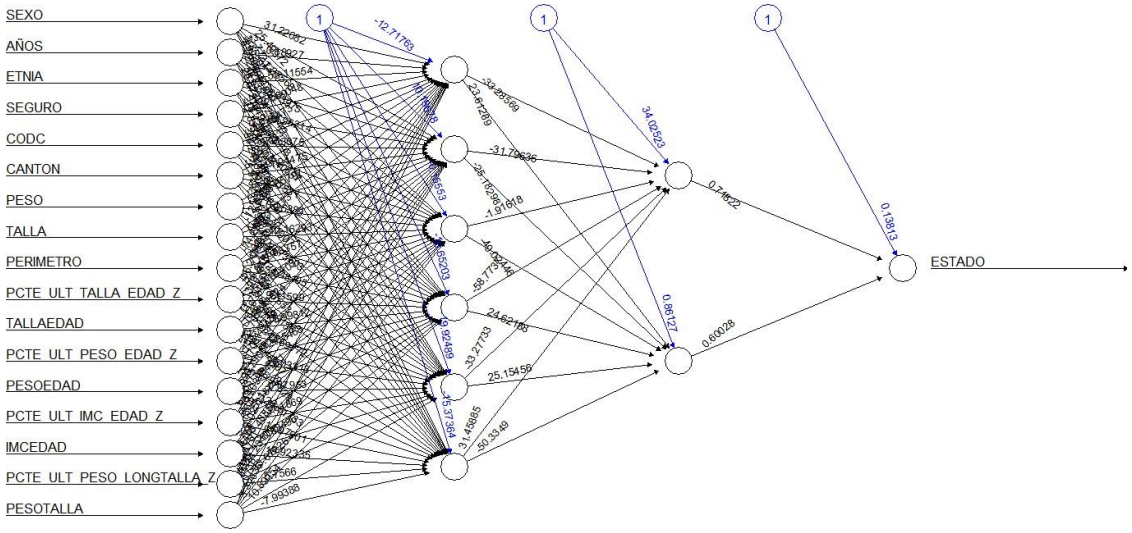



**Ilustración 4–25:** Red Neuronal en R

**Realizado por:** Sánchez Evelyn y Tenesaca Sebastian, 2023.

En el gráfico de modelado de red neuronal para predecir la anemia infantil, se puede observar que las 17 variables de entrada o predictoras se entrelazan con las capas ocultas de 6 y 2 neuronas, lo que hace es dar información a cada nodo y a la vez esta crea un modelo matemático que se entrena con la información simulada en cada neurona, esto genera interacciones encada parte de la red y se arroja un peso significativo, en este caso las 17 variables interactúan con las 6 neuronas la cual tendrá como objetivo dar información a las otras dos y creando una sola salida de un resultado de ESTADO de la anemia si es definitivo o presuntivo. Se observo que el peso del nodo de salida es positivo, lo cual indica un peso significativo para la predicción de anemia. La red neuronal se entrena mediante el algoritmo de retropropagación mejorada con incrementos proporcionales (Rprop+). Este algoritmo es una variante del algoritmo de retropropagación estándar y se utiliza para ajustar los pesos de las conexiones entre neuronas en función del error de predicción del entrenamiento.

### *4.3.3. Curvas de ROC de los modelos analizados*

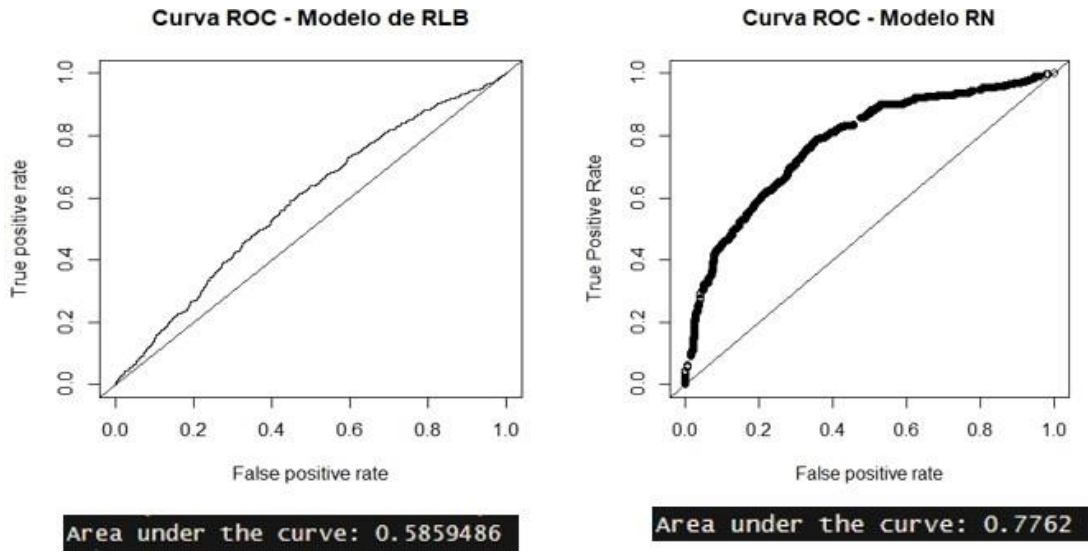



**Ilustración 4–26:** Curvas de ROC del modelo de regresión logística binaria y de redes neuronales

**Realizado por:** Sánchez Evelyn y Tenesaca Sebastian, 2023.

El modelo de redes neuronales tiene una AUC más alta que el modelo de regresión logística. Esto demuestra que el modelo de redes neuronales puede predecir mejor la clasificación de clases positivas y negativas, es decir, el estado presuntivo o definitivo de los pacientes. También puede detectar patrones o relaciones no lineales más complejos en los datos que el modelo de regresión logística no puede. Esto podría deberse al hecho de que el modelo de redes neuronales puede aprender representaciones de datos más complejas y no lineales. La capacidad discriminativa se mide por la AUC, y un AUC más alto en el modelo de redes neuronales indica que, en comparación con el modelo de regresión logística, el modelo tiene un mayor poder predictivo para clasificar las clases positivas y negativas.

### *4.3.4. Mosaicos de confusión de los modelos*

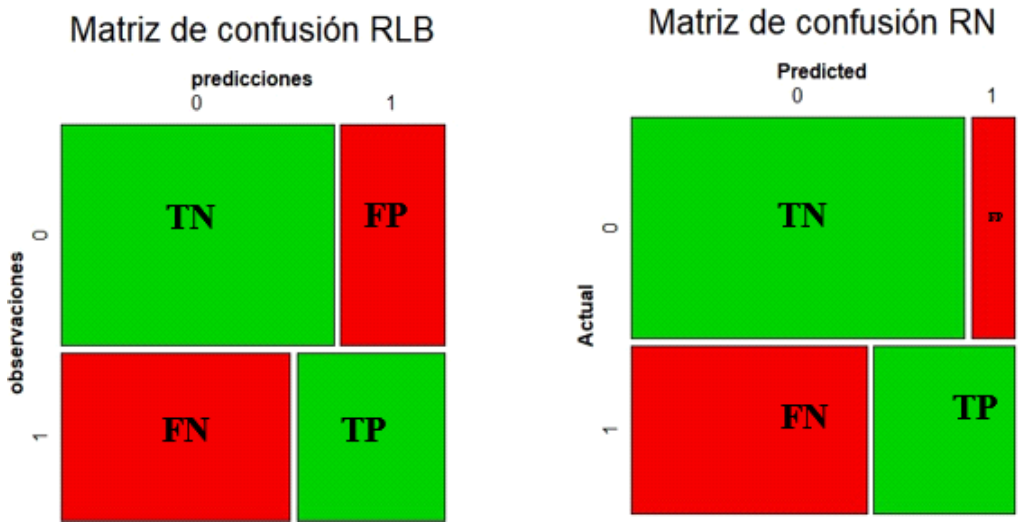



**Ilustración 4–27:** Mosaicos de confusión de regresión logística binaria y de redes neuronales

**Realizado por:** Sánchez Evelyn y Tenesaca Sebastian, 2023.

Los mosaicos nos dieron una vista preliminar, lo que se pudo deducir es que tenemos muchos valores de Verdaderos Negativos (TN) y de Falsos Negativos (FN), La alta cantidad de Verdaderos Negativos (TN) indica que el modelo es capaz de reconocer correctamente la mayoría de los casos negativos. Se evidencio que en el modelo de redes neuronales este valor es más grande. Esto indica que el modelo puede distinguir correctamente entre muestras positivas y negativas, lo que lo hace positivo. Para un mejor análisis se identificó las métricas correspondientes de: Sensibilidad, Especificidad, Exactitud y Precisión.

Verdadero positivo (**TP**): El valor real y la predicción son positivos. O bien la prueba indica que una persona tiene anemia definitiva.

Verdadero negativo (**TN**): El resultado de la prueba fue negativo, así como el valor real. O bien la prueba indica que la persona tiene anemia presuntiva.

Falso negativo (**FN**): La prueba predijo un resultado negativo a pesar de que el valor real es positivo. Aunque la persona tiene anemia definitiva, la prueba indica incorrectamente que está en estado presuntivo. Se conoce como error tipo II.

Falso positivo (**FP**): La prueba predijo un resultado positivo a pesar de que el valor real es negativo.

Aunque la persona tiene anemia presuntiva, la prueba indica incorrectamente que está en estado definitivo. Se conoce como error tipo I.

### *4.3.5. Matrices de confusión de los modelos*

Una vez obtenidos los resultados en R, podemos construir la siguiente tabla:

**Tabla 4–2:** Matriz de confusión del modelo de regresión logística binaria

| **REGRESIÓN LOGÍSTICA BINARIA** | | **ESTADO** | |
|---|---|---|---|
| | | **PRESUNTIVO** | **DEFINITIVO** |
| **ESTADO** | **PRESUNTIVO** | 899 | 343 |
| | **DEFINITIVO** | 577 | 372 |

**Fuente**: Matriz de datos del HPAVR 2020-2021

**Realizado por**: Sánchez Evelyn y Tenesaca Sebastian, 2023

- a=Verdaderos negativos (TN): 899
- b= Falsos positivos (FP): 343
- c=Falsos negativos (FN): 577
- d=Verdaderos positivos (TP): 372

**Tabla 4–3:** Métricas de la matriz de confusión del modelo de RLB

| **Métricas de la matriz de confusión RLB** | | | |
|---|---|---|---|
| **Sensibilidad** | **Especificidad** | **Exactitud** | **Precisión** |
| $\frac{d}{d+c}$ | $\frac{a}{a+b}$ | $\frac{a+d}{a+b+c+d}$ | $\frac{d}{b+d}$ |
| 0.391 | 0.723 | 0.580 | 0.520 |

**Fuente**: Matriz de datos del HPAVR 2020-2021

**Realizado por**: Sánchez Evelyn y Tenesaca Sebastian, 2023

Los resultados de la matriz de confusión muestran que el modelo de regresión logística detecta casos positivos con baja sensibilidad y precisión y casos negativos con moderada especificidad. El modelo también tiene una precisión moderada en general.

**Tabla 4–4:** Matriz de confusión del modelo de redes neuronales

| **REDES NEURONALES** | **ESTADO** | |
|---|---|---|
| | **PRESUNTIVO** | **DEFINITIVO** |

| | | | |
|---|---|---|---|
| **ESTADO** | **PRESUNTIVO** | 1100 | 142 |
| | **DEFINITIVO** | 594 | 355 |

**Fuente**: Matriz de datos del HPAVR 2020-2021

**Realizado por**: Sánchez Evelyn y Tenesaca Sebastian, 2023

- a=Verdaderos negativos (TN): 1100
- b= Falsos positivos (FP): 142
- c=Falsos negativos (FN): 594
- d=Verdaderos positivos (TP): 355

**Tabla 4–5:** Métricas de la matriz de confusión del modelo de RN

| **Métricas de la matriz de confusión RLB** | | | |
|---|---|---|---|
| **Sensibilidad** | **Especificidad** | **Exactitud** | **Precisión** |
| $\frac{d}{d+c}$ | $\frac{a}{a+b}$ | $\frac{a+d}{a+b+c+d}$ | $\frac{d}{b+d}$ |
| 0.141 | 0.885 | 0.664 | 0.714 |

**Fuente**: Matriz de datos del HPAVR 2020-2021

**Realizado por**: Sánchez Evelyn y Tenesaca Sebastian, 2023

Los resultados de la matriz de confusión muestran que el modelo de redes neuronales en su forma de entrenamiento tiene una sensibilidad baja, lo que significa que no es muy eficaz en detectar casos positivos. Sin embargo, tiene una alta especificidad, lo que indica que puede identificar casos negativos con precisión. La exactitud general del modelo es moderada y la precisión en la clasificación de casos positivos es también moderada.

## CONCLUSIONES

- En conclusión, basándose en los resultados de este estudio, se puede decir que el modelo de redes neuronales es más efectivo que el modelo de regresión logística para predecir la anemia en niños menores de cinco años. El modelo de redes neuronales tiene una mayor AUC, lo que indica un mayor poder predictivo para clasificar correctamente los casos positivos y negativos.
- Se determino que los posibles factores enfocados en nuestras variables disponibles a través de un análisis descriptivo fueron: Sexo, Edad, Etnia, Seguro, Cantón, Peso, Talla, perímetro cefálico, las relaciones entre (Talla y Edad, Peso y Edad, IMC y edad, Peso y Talla) y se puede verificar que lo son, dado las variables significativas en el resultado de regresión logística.
- Con la base de datos recolectada por el Hospital Pediátrico Alfonso Villagómez Román, se pudo definir los factores asociados que fueron categorizadas la mayoría en binarios ya que de esa manera se evitó la depuración de varios datos que son de gran relevancia para este tipo de estudios, consolidando una matriz adecuada para este trabajo.
- Se logro modelar de manera adecuada el modelo de regresión logística, a pesar de que es poco preciso en la detección de casos negativos, pero es poco sensible y preciso en la detección de casos positivos. Por lo tanto, es ineficaz para clasificar a pacientes de manera precisa.
- El modelo de redes neuronales presenta una alta especificidad, y a la vez puede capturar relaciones no lineales más complejas en los datos, lo que le permite encontrar patrones más sutiles y hacer predicciones más precisas siendo el mejor modelo y método para la predicción de anemia en niños menores de 5 años.

## RECOMENDACIONES

- Las matrices obtenidas a través del Hospital Pediátrico Alfonso Villagómez Román son de gran ayuda como base para este tipo de estudios, sin embargo, se recomienda corregir datos erróneos de tipeo, ya que existen varios datos erróneos compresibles por el error humano al manejar gran cantidad de datos en cortos periodos de tiempo.

- Se deben elegir cuidadosamente las variables que se consideran más relevantes y se deben eliminar las variables que puedan tener un impacto mínimo en el modelo. Una evaluación exhaustiva es necesaria después de ajustar el modelo de regresión logística para determinar su rendimiento y robustez. Para evaluar la capacidad predictiva y la calidad del modelo, utiliza métricas de evaluación como la AUC, la sensibilidad, la especificidad y la precisión.

- Se debe comprender que las redes neuronales son modelos complejos y que con frecuencia se las considera çajas negras". Aunque son muy predictivos, la interpretación de los resultados puede ser difícil. Para obtener una mejor comprensión de cómo toma decisiones la red neuronal, se debe aprender técnicas como la importancia de las características o la visualización de activaciones.

- Se debe considerar el uso de modelos de aprendizaje automático para mejorar la capacidad de predicción y comprensión de la anemia en esta población. Estas herramientas pueden ayudar a identificar patrones y factores de riesgo con mayor precisión, lo que permitirá una intervención más efectiva.

**BIBLIOGRAFÍA**


**BERTONA, Luis Federico** *ENTRENAMIENTO DE REDES NEURONALES BASADO EN ALGORITMOS EVOLUTIVOS*. Buenos Aires: s.n., 2005.

**CORNEJO RUIZ, Daniel & QUISPE GAVINO, Giancarlo** *Aplicación del algoritmo Backpropagation de redes neuronales para determinar los niveles de morosidad en los alumnos de la Universidad Peruana Unión.* Perú: s.n., 2008.

**Secretaría Técnica Ecuador Crece Sin Desnutrición.** *Primera Infancia*. Quito: s.n., 2013.

**ABANTO CHAVARRI, Antony.** *Modelo logístico y redes neuronales para pronóstico*. Universidad Nacional de Trujillo, Trujillo, Perú: 2022.

**BROWN, A., et al.** *Estadística para médicos: Guía práctica para la interpretación de estudios clínicos y epidemiológicos*. Editorial Médica Internacional, 2018.

**CHITARRONI, Horacio.** *La regresión logística*. Buenos Aires: s.n., 2002.

**DANIGNO, Jorge.** *s.l.* Revista chilena de anestesia, 2014, Revista chilena de anestesia, Vol. 43, págs. 332-334.

**GEBREEGZIABHER, T., et al.** *Prevalence and risk factors of anemia among children aged 6-59 months in rural Ethiopia*. Pediatric Health, Medicine and Therapeutics, 2018, págs. 61-68.

**Health, Organization World.** *The Global Prevalence of Anaemia in 2011*. 2015.

**IZAURIETA, Fernando & SAAVEDRA, Carlos.** *Redes Neuronales Artificiales.* Chile: s.n., 200.

**JONES, E, SMITH, M. & DAVIS, R.** *Terapia de reemplazo hormonal y riesgo de fracturas óseas en mujeres posmenopáusicas: Un ensayo clínico aleatorizado*. Revista de Medicina Basada en la Evidencia, 2020, págs. 145-160.

**KAUR, H., et al.** *Artificial neural network model for prediction of anemia in under-five children using classification techniques*. International Journal of Applied Engineering Research, 2018, págs. 902-907.

**MEDINA, Fernando & GALVÁN, Marco.** *Imputación de datos teórica y práctica*. 2007.

**MOYANO BRITO, Edison Gustavo, et al.** *Factores asociados a la anemia en niños ecuatorianos de 1 a 4 años*. s.l.: Archivos Venezolanos de Farmacología y Terapéutica, 2019, págs. 695-699.

**MUÑOZ ROSAS, Juan Francisco & ÁLVAREZ VERDEJO, Encarnación.** *Métodos de imputación para el tratamiento de datos faltantes: aplicación mediante R/Splus*. 2009, pág. 30.

**PEÑA, Daniel.** *ANÁLISIS DE DATOS MULTIVARIANTES*. Madrid: McGraw, 2002, Vol. 24.

**PÉREZ LÓPEZ, César and SATIN GONZÁLES, Daniel.** *Minería de datos. Técnicas y herramientas: técnicas y herramientas*. s.l.: Editorial Paraninfo, 2007.

**RAHMAN, M. S., et al.** *Prevalence of anemia in Bangladesh: a systematic review andmeta-analysis*. BMC Hematology, 2019, pág. 5.

**RAMÍREZ, Fran.** *Las matemáticas del Machine Learning: Funciones de activación*. 2020.

**ROSADO-MENDOZA, E., et al.** *Anemia en niños menores de 5 años de una población rural del Perú: prevalencia y factores asociados*. Revista Peruana de Medicina Experimental y Salud Pública, 2017, págs. 234-241.

**SERNA PINEDA, Sandra Carolina.** *Comparación de árboles de regresión y clasificación y regresión logística*. [En línea] 2009. [Citado el: 2 de 2 de 2023.] https://repositorio.unal.edu.co/handle/unal/2421.

**SMITH, J. & JONES, K.** *Métodos estadísticos en investigación médica: Conceptos y aplicaciones*. s.l.: Editorial Médica Internacional., 2019.

**SMITH, R, JOHNSON, T. & BROWN, A.** *Asociación entre tabaquismo y enfermedades cardiovasculares: Resultados de un estudio de cohorte de largo plazo*. Revista de Medicina Preventiva, 2017, págs. 75-89.

**VINDELL, Juan José.** *Regresión Logística en Salud Pública*. s.l.: RPubs, 2021.

**ZAIONTZ, Carlos.** *ESTADÍSTICAS REALES USANDO EXCEL*. 2014.

**ANEXOS**

```
# Analisis descriptivo
library(readxl)
#LECTURA DE DATOS

DATOS <- read_excel("C:/Users/Bastianse/Documents/
BASE_ANEMIAORGINAL.xlsx")
ANEMIA<-data.frame(DATOS)
ANEMIA
str(ANEMIA)
ANEMIA$ANIOATEN<-as.numeric(ANEMIA$ANIOATEN)
str(ANEMIA)


# ANIOATEN
tabla_ANIOATEN <- table(ANEMIA$ANIOATEN)
prop_ANIOATEN <- prop.table(tabla_ANIOATEN)
colores <- c("red", "green")
# Graficar las barras con colores diferentes y mostrar la frecuencia en
cada barra
text(x = barplot(tabla_ANIOATEN, col = colores, main = "GRAFICA DE LA
VARIABLE ANIOATEN", xlab = "ANIOATEN", ylab = "Frecuencia"),
     y = prop_ANIOATEN, labels = round(prop_ANIOATEN,digits = 2)
, pos = 3,)


# SEXO
tabla_sexo <- table(ANEMIA$SEXO)
prop_sexo <- prop.table(tabla_sexo)
colores <- c("blue", "pink")
# Graficar las barras con colores diferentes y mostrar la frecuencia en
cada barra
text(x = barplot(tabla_sexo, col = colores, main = "GRAFICA DE LA
VARIABLE SEXO", xlab = "SEXO", ylab = "Frecuencia"),
     y = prop_sexo, labels = round(prop_sexo,digits = 2),
pos = 3,)

```

```r
# ANIOS
tabla_ANIOS <- table(ANEMIA$ANIOS)
prop_ANIOS<- prop.table(tabla_ANIOS)
colores <- c("blue", "pink","red","orange","green","brown")
# Graficar las barras con colores diferentes y mostrar la frecuencia en
cada barra
text(x = barplot(tabla_ANIOS, col = colores, main = "GRNIFICA DE LA
VARIABLE ANIOS", xlab = "ANIOS", ylab = "Frecuencia"),
     y = prop_ANIOS, labels = round(prop_ANIOS,digits = 2),
pos = 3)

# ETNIA
tabla_ETNIA <- table(ANEMIA$ETNIA)
prop_ETNIA<- prop.table(tabla_ETNIA)
colores <- c("blue", "pink","red","orange","green","brown")
# Graficar las barras con colores diferentes y mostrar la frecuencia en
cada barra
text(x = barplot(tabla_ETNIA, col = colores, main = "GRAFICA DE LA
VARIABLE ETNIA",
                 xlab = "ETNIA", ylab = "Frecuencia"),
     y = prop_ETNIA, labels = round(prop_ETNIA,digits = 4),
pos = 3)

#SEGURO
tabla_SEGURO <- table(ANEMIA$SEGURO)
prop_SEGURO<- prop.table(tabla_SEGURO)
colores <- c("blue", "pink","red","orange","green","brown")
# Graficar las barras con colores diferentes y mostrar la frecuencia en
cada barra
text(x = barplot(tabla_SEGURO, col = colores, main = "GRAFICA DE LA
VARIABLE SEGURO",
                 xlab = "SEGURO", ylab = "Frecuencia"),
     y = prop_SEGURO, labels = round(prop_SEGURO,digits = 4),
 pos = 3)

#COD
tabla_COD <- table(ANEMIA$COD)
prop_COD<- prop.table(tabla_COD)
colores <- c(1:20)
```

```r
# Graficar las barras con colores diferentes y mostrar la frecuencia en cada barra
barplot(tabla_COD, col = colores, main = "GRAFICA DE LA VARIABLE COD",
        xlab = "COD", ylab = "Frecuencia")
prop.table(table(ANEMIA$COD))


#CANTON
tabla_CANTON <- table(ANEMIA$CANTON)
prop_CANTON<- prop.table(tabla_CANTON)
colores <- c(1:20)
# Graficar las barras con colores diferentes y mostrar la frecuencia en cada barra
barplot(tabla_CANTON, col = colores, main = "GRAFICA DE LA VARIABLE CANTON",
                  xlab = "CANTON", ylab = "Frecuencia")
prop.table(table(ANEMIA$CANTON))

#PESO
tabla_PESO <- table(ANEMIA$PESO)
prop_PESO<- prop.table(tabla_PESO)
colores <- c(1:20)
# Graficar las barras con colores diferentes y mostrar la frecuencia en cada barra
barplot(tabla_PESO, col = colores, main = "GRAFICA DE LA VARIABLE PESO",
        xlab = "PESO", ylab = "Frecuencia")
prop.table(table(ANEMIA$PESO))



#TALLA
tabla_TALLA <- table(ANEMIA$TALLA)
prop_TALLA<- prop.table(tabla_TALLA)
colores <- c(1:20)
# Graficar las barras con colores diferentes y mostrar la frecuencia en cada barra
barplot(tabla_TALLA, col = colores, main = "GRAFICA DE LA
```

```r
 VARIABLE TALLA",
        xlab = "TALLA", ylab = "Frecuencia")
prop.table(table(ANEMIA$TALLA))

#PERIMETRO
tabla_PERIMETRO <- table(ANEMIA$PERIMETRO)
prop_PERIMETRO<- prop.table(tabla_PERIMETRO)
colores <- c(1:20)
# Graficar las barras con colores diferentes y mostrar la frecuencia en
cada barra
barplot(tabla_PERIMETRO, col = colores, main = "GRAFICA DE LA
 VARIABLE PERIMETRO",
        xlab = "PERIMETRO", ylab = "Frecuencia")
prop.table(table(ANEMIA$PERIMETRO))
max(prop_PERIMETRO)

#PCTE_ULT_TALLA_EDAD_Z
tabla_PCTE_ULT_TALLA_EDAD_Z <- table(ANEMIA$PCTE_ULT_TALLA_EDAD_Z)
prop_PCTE_ULT_TALLA_EDAD_Z<- prop.table(tabla_PCTE_ULT_TALLA_EDAD_Z)
colores <- c(1:20)
# Graficar las barras con colores diferentes y mostrar la frecuencia en
cada barra
barplot(tabla_PCTE_ULT_TALLA_EDAD_Z, col = colores, main = "GRAFICA DE
LA VARIABLE PCTE_ULT_TALLA_EDAD_Z",
        xlab = "PCTE_ULT_TALLA_EDAD_Z", ylab = "Frecuencia")
prop.table(table(ANEMIA$PCTE_ULT_TALLA_EDAD_Z))

#TALLAEDAD
tabla_TALLAEDAD <- table(ANEMIA$TALLAEDAD)
prop_TALLAEDAD<- prop.table(tabla_TALLAEDAD)
colores <- c(1:20)
# Graficar las barras con colores diferentes y mostrar la frecuencia en
 cada barra
text(x = barplot(tabla_TALLAEDAD, col = colores, main = "GRAFICA DE LA
 VARIABLE TALLAEDAD",
                 xlab = "TALLAEDAD", ylab = "Frecuencia"),
     y = prop_TALLAEDAD, labels = round(prop_TALLAEDAD,digits = 4)
, pos = 3)

```

```r

#PCTE_ULT_PESO_EDAD_Z
tabla_PCTE_ULT_PESO_EDAD_Z <- table(ANEMIA$PCTE_ULT_PESO_EDAD_Z)
prop_PCTE_ULT_PESO_EDAD_Z<- prop.table(tabla_PCTE_ULT_PESO_EDAD_Z)
colores <- c(1:20)
# Graficar las barras con colores diferentes y mostrar la frecuencia
 en cada barra
barplot(tabla_PCTE_ULT_PESO_EDAD_Z, col = colores, main = "GRAFICA DE
 LA VARIABLE PCTE_ULT_PESO_EDAD_Z",
        xlab = "PCTE_ULT_PESO_EDAD_Z", ylab = "Frecuencia")
prop.table(table(ANEMIA$PCTE_ULT_PESO_EDAD_Z))

#PESOEDAD
tabla_PESOEDAD <- table(ANEMIA$PESOEDAD)
prop_PESOEDAD<- prop.table(tabla_PESOEDAD)
colores <- c(2:20)
# Graficar las barras con colores diferentes y mostrar la frecuencia en
cada barra
text(x = barplot(tabla_PESOEDAD, col = colores, main = "GRAFICA DE LA
VARIABLE PESOEDAD",
                 xlab = "PESOEDAD", ylab = "Frecuencia"),
     y = prop_PESOEDAD, labels = round(prop_PESOEDAD,digits = 4),
pos = 3)


#PCTE_ULT_IMC_EDAD_Z
ANEMIA$PCTE_ULT_IMC_EDAD_Z<-as.numeric(ANEMIA$PCTE_ULT_IMC_EDAD_Z)
tabla_PCTE_ULT_IMC_EDAD_Z <- table(ANEMIA$PCTE_ULT_IMC_EDAD_Z)
prop_PCTE_ULT_IMC_EDAD_Z<- prop.table(tabla_PCTE_ULT_IMC_EDAD_Z)
colores <- c(1:20)
# Graficar las barras con colores diferentes y mostrar la frecuencia
en cada barra
barplot(tabla_PCTE_ULT_IMC_EDAD_Z, col = colores, main = "GRAFICA DE LA
VARIABLE PCTE_ULT_IMC_EDAD_Z",
        xlab = "PCTE_ULT_IMC_EDAD_Z", ylab = "Frecuencia")
prop.table(table(ANEMIA$PCTE_ULT_IMC_EDAD_Z))

#IMCEDAD
tabla_IMCEDAD <- table(ANEMIA$IMCEDAD)
```

```r
prop_IMCEDAD<- prop.table(tabla_IMCEDAD)
colores <- c(2:20)
# Graficar las barras con colores diferentes y mostrar la frecuencia en cada barra
text(x = barplot(tabla_IMCEDAD, col = colores, main = "GRAFICA DE LA VARIABLE IMCEDAD",
                 xlab = "IMCEDAD", ylab = "Frecuencia"),
     y = prop_IMCEDAD, labels = round(prop_IMCEDAD,digits = 4), pos = 3)

#PCTE_ULT_PESO_LONGTALLA_Z
ANEMIA$PCTE_ULT_PESO_LONGTALLA_Z<-as.numeric(ANEMIA$PCTE_ULT_PESO_LONGTALLA_Z)
tabla_PCTE_ULT_PESO_LONGTALLA_Z <- table(ANEMIA$PCTE_ULT_PESO_LONGTALLA_Z)
prop_PCTE_ULT_PESO_LONGTALLA_Z<- prop.table(tabla_PCTE_ULT_PESO_LONGTALLA_Z)
colores <- c(1:20)
# Graficar las barras con colores diferentes y mostrar la frecuencia en cada barra
barplot(tabla_PCTE_ULT_PESO_LONGTALLA_Z, col = colores, main = "GRAFICA DE LA VARIABLE PCTE_ULT_PESO_LONGTALLA_Z",
        xlab = "PCTE_ULT_PESO_LONGTALLA_Z", ylab = "Frecuencia")
prop.table(table(ANEMIA$PCTE_ULT_PESO_LONGTALLA_Z))

#PESOTALLA
tabla_PESOTALLA <- table(ANEMIA$PESOTALLA)
prop_PESOTALLA<- prop.table(tabla_PESOTALLA)
colores <- c(1:20)
# Graficar las barras con colores diferentes y mostrar la frecuencia en
 cada barra
text(x = barplot(tabla_PESOTALLA, col = colores, main = "GRAFICA DE LA VARIABLE PESOTALLA",
                 xlab = "PESOTALLA", ylab = "Frecuencia"),
     y = prop_PESOTALLA, labels = round(prop_PESOTALLA,digits = 4), pos = 3)
prop.table(table(ANEMIA$PESOTALLA))


#DIAGNOSTICO
tabla_DIAGNOSTICO <- table(ANEMIA$DIAGNOSTICO)
```

```r
prop_DIAGNOSTICO<- prop.table(tabla_DIAGNOSTICO)
colores <- c(2:20)
# Graficar las barras con colores diferentes y mostrar la frecuencia en
cada barra
text(x = barplot(tabla_DIAGNOSTICO, col = colores, main = "GRAFICA DE LA
VARIABLE DIAGNOSTICO",
                 xlab = "DIAGNOSTICO", ylab = "Frecuencia"),
     y = prop_DIAGNOSTICO, labels = round(prop_DIAGNOSTICO,digits = 4),
 pos = 3)
prop.table(table(ANEMIA$DIAGNOSTICO))


#ESTADO
tabla_ESTADO <- table(ANEMIA$ESTADO)
prop_ESTADO<- prop.table(tabla_ESTADO)
colores <- c(2:20)
# Graficar las barras con colores diferentes y mostrar la frecuencia en
 cada barra
text(x = barplot(tabla_ESTADO, col = colores, main = "GRAFICA DE LA
 VARIABLE ESTADO",
                 xlab = "ESTADO", ylab = "Frecuencia"),
     y = prop_ESTADO, labels = round(prop_ESTADO,digits = 4), pos = 3)
prop.table(table(ANEMIA$ESTADO))



_____________________________________________________________
# Paso 1: Instalar y cargar la librerIa "readxl"
install.packages("readxl")
library(readxl)
library(openxlsx)

# Paso 2: Leer el archivo XLSX
datos <- read_xlsx("C:/Users/Bastianse/Documents/BASEANEMIA.xlsx")

datos$ETNIA<-as.numeric(datos$ETNIA)
datos$PERIMETRO<-as.numeric(datos$PERIMETRO)
datos$PCTE_ULT_IMC_EDAD_Z<-as.numeric(datos$PCTE_ULT_IMC_EDAD_Z)
datos$IMCEDAD<-as.numeric(datos$IMCEDAD)
```

```r
datos$PCTE_ULT_PESO_LONGTALLA_Z<-as.numeric(datos$PCTE_ULT_PESO
_LONGTALLA_Z)
datos$PESOTALLA<-as.numeric(datos$PESOTALLA)
datos$ESTADO<-as.numeric(datos$ESTADO)

str(datos)

# Paso 3: Calcular la media de las variables numUricas
medias <- round(colMeans(datos, na.rm = TRUE))

# Paso 4: Reemplazar los valores NA por la media en cada variable
for (variable in names(datos)) {
  if (is.numeric(datos[[variable]])) {
    datos[[variable]][is.na(datos[[variable]])] <- medias[variable]
  }
}

# Paso opcional: Guardar los datos modificados en un nuevo archivo XLSX


write.xlsx(datos, "C:/Users/Bastianse/Documents/BDDANEMIA1.xlsx",
 row.names = FALSE)


_____________________________________________________________________
# Elimina los datos atIpicos
library(dplyr)
# Paso 2: Leer el archivo XLSX
datos <- read_xlsx("C:/Users/Bastianse/Documents/BDDANEMIA1.xlsx")

datos <- datos %>%
  filter_if(is.numeric, ~. > mean(.) - 4*sd(.) & . < mean(.) + 4*sd(.))
View(datos)
# Guarda los datos limpios en un nuevo archivo XLSX
ruta_salida <- "C:/Users/Bastianse/Documents/BDD_ANEMIA3.xlsx"
write.xlsx(datos, ruta_salida, rowNames = FALSE)
_____________________________________________________________________

#regresion logistica binaria
```

```r

library(readxl)
bd<-read_xlsx("C:/Users/Bastianse/Documents/BDD_ANEMIA3.xlsx")
bd<-as.data.frame(bd)
head(bd)

modelo1<- glm(ESTADO~ SEXO+ANIOS+ETNIA+SEGURO+CODC+CANTON+PESO+TALLA
+PERIMETRO+ PCTE_ULT_TALLA_EDAD_Z +TALLAEDAD +PCTE_ULT_PESO_EDAD_Z +
PESOEDAD+ PCTE_ULT_IMC_EDAD_Z +IMCEDAD +PCTE_ULT_PESO_LONGTALLA_Z
+PESOTALLA,
              data = bd,family="binomial"(link="logit"))
summary(modelo1)
#AIC: 2981.9

modelo2<- glm(ESTADO~ ANIOS+SEGURO+PERIMETRO+
                PCTE_ULT_TALLA_EDAD_Z +PCTE_ULT_PESO_EDAD_Z +PESOEDAD
               +PCTE_ULT_PESO_LONGTALLA_Z,
              data = bd,family="binomial"(link="logit"))
summary(modelo2)

#AIC: 2970.8

modelo3<- glm(ESTADO~ ANIOS+SEGURO+
                PCTE_ULT_TALLA_EDAD_Z +PCTE_ULT_PESO_EDAD_Z
              +PCTE_ULT_PESO_LONGTALLA_Z,
              data = bd,family="binomial"(link="logit"))
summary(modelo3)

#AIC: 2070.1



# REALIZACION DE PREDICCIONES

pred<-predict(modelo3,type ="response")
head(pred)

#Probar con un elemento de la poblaciOn
#define dos individuos
```

```
349
350
351
352 ### Curva ROC
353 library(ROCR)
354 ROCpred<-prediction(pred,bd$ESTADO)
355 ROCpref<-performance(ROCpred,"tpr","fpr")
356 plot(ROCpref,print.cutoffs.at=seq(0.1,by=0.4))
357
358
359 plot(ROCpref,type="l",main = "Curva ROC - Modelo de RLB")
360 abline(a=0,b=1)
361 # Area bajo la curva
362 AUC       <- performance(ROCpred,measure="auc")
363 AUCaltura <- AUC@y.values
364
365 # Punto de corte Optimo
366 cost.perf <- performance(ROCpred, measure ="cost")
367 opt.cut   <- ROCpred@cutoffs[[1]][which.min(cost.perf@y.values[[1]])]
368 #coordenadas del punto de corte Optimo
369 x<-ROCpref@x.values[[1]][which.min(cost.perf@y.values[[1]])]
370 y<-ROCpref@y.values[[1]][which.min(cost.perf@y.values[[1]])]
371 points(x,y, pch=20, col="red")
372 cat("Area under the curve:", AUCaltura[[1]])
373 cat("Punto de corte Optimo:",opt.cut)
374 X<-opt.cut
375
376 predicciones <- ifelse(test = modelo3$fitted.values >opt.cut , yes = 1,
377 no = 0)
378 matriz_confusion <- table(modelo3$model$ESTADO, predicciones,
379                           dnn = c("observaciones", "predicciones"))
380 matriz_confusion
381 library(vcd)
382 mosaic(main="Matriz de confusiOn RLB",matriz_confusion, shade = T,
383  colorize = T,
384        gp = gpar(fill = matrix(c("green3", "red2", "red2", "green3"),
385  2, 2)))
386
387
```

```r
a<-899
b<-343
c<-577
d<-372

#presiciOn
d/(d+b)
#exactitud
(d+a)/(a+b+c+d)
#sensibilidad
d/(d+c)
#especificidad
a/(a+b)
#valor de prediccion +
d/(d+b)
#valor de predicciOn -
a/(a+c)

(b+c)/(a+b+c+d)
_____________________________________________________________
library(readxl)
library(caTools)
library(neuralnet)
library(NeuralNetTools)

datos    <- read_excel("C:/Users/Bastianse/Documents/BDD_ANEMIA3.xlsx")
View(datos)
dataset<-as.data.frame(cbind(datos$ESTADO,scale(datos[,c(1:17)])))
names(dataset)[1]='ESTADO'
View(dataset)

sample<-sample.split(dataset$ESTADO,SplitRatio = 0.7)
train<-subset(dataset,sample==T)
test<-subset(dataset,sample==F)

modelo.nn=neuralnet(
  formula = ESTADO ~ SEXO + ANIOS + ETNIA + SEGURO + CODC + CANTON
+ PESO +  TALLA + PERIMETRO + PCTE_ULT_TALLA_EDAD_Z + TALLAEDAD +
PCTE_ULT_PESO_EDAD_Z + PESOEDAD + PCTE_ULT_IMC_EDAD_Z + IMCEDAD
```

```
+ PCTE_ULT_PESO_LONGTALLA_Z +  PESOTALLA
,data = train, hidden = c(6,2),act.fct = 'logistic',algorithm
 = "rprop+")
modelo.nn

plot(modelo.nn)

library(pROC)

# Obtener las probabilidades de clase predichas en conjunto
#de entrenamiento y prueba

train$probabilidad <- predict(modelo.nn, train)
test$probabilidad <- predict(modelo.nn, test)
train$probabilidad <- as.numeric(as.vector(train$probabilidad))
test$probabilidad <- as.numeric(as.vector(test$probabilidad))


# Calcular la curva ROC en conjunto de entrenamiento y prueba
roc_train <- roc(train$ESTADO, train$probabilidad)
roc_test <- roc(test$ESTADO, test$probabilidad)

# Graficar la curva ROC
plot(1-roc_train$specificities, roc_train$sensitivities, main =
"Curva ROC - Modelo RN", xlab = "False positive rate", ylab
 = "True Positive Rate")
plot(1-roc_test$specificities, roc_test$sensitivities, main =
"Curva ROC - Modelo RN", xlab = "False positive rate", ylab
 = "True Positive Rate")
abline(a = 0, b = 1, lty = 1)
# Encontrar el punto de corte Optimo en conjunto de entrenamiento
#y prueba
corte_optimo_train <- coords(roc_train, "best")
corte_optimo_test <- coords(roc_test, "best")

# Mostrar el punto de corte Optimo en conjunto de entrenamiento
#y prueba
abline(v = corte_optimo_train$specificity, h =
corte_optimo_train$sensitivity, col = "red", lty = 2)
```

```
text(corte_optimo_train$specificity, corte_optimo_train$sensitivity,
 paste0
("(", round(corte_optimo_train$specificity, 2), ",",
round(corte_optimo_train$sensitivity
, 2), ")"), pos = 3)

abline(v = corte_optimo_test$specificity,
h = corte_optimo_test$sensitivity,
 col = "red", lty = 2)
text(corte_optimo_test$specificity, corte_optimo_test$sensitivity
, paste0("(",
 round(corte_optimo_test$specificity, 2), ","
, round(corte_optimo_test$sensitivity,
 2), ")"), pos = 3)

# Calcular el AUC en conjunto de entrenamiento y prueba
auc_train <- auc(train$ESTADO, train$probabilidad)
auc_test <- auc(test$ESTADO, test$probabilidad)
__________

# Realizar predicciones en conjunto de entrenamiento y prueba
train$prediccion <- predict(modelo.nn, train)
test$prediccion <- predict(modelo.nn, test)

# Establecer umbral y asignar etiquetas de clase
umbral <- 0.7
train$prediccion <- ifelse(train$prediccion >= umbral,yes = 1, no = 0)
test$prediccion <- ifelse(test$prediccion >= umbral,yes = 1, no = 0)

# Crear matriz de confusiOn para el conjunto de entrenamiento
#y prueba unido
matriz_confusion <- table(Actual = c(train$ESTADO, test$ESTADO)
, Predicted
= c(train$prediccion, test$prediccion))
print(matriz_confusion)

library(vcd)
mosaic(main="Matriz de confusiOn RN",matriz_confusion, shade = T
, colorize = T,
```

```
       gp = gpar(fill = matrix(c("green3", "red2", "red2", "green3")
, 2, 2)))

a<-1991
b<-0
c<-684
d<-0


#presiciOn
p<-d/(d+b);p
#exactitud
(d+a)/(a+b+c+d)
#sensibilidad
s<-d/(d+c);S
#especificidad
a/(a+b)
#valor de prediccion +
d/(d+b)
#valor de prediccIOn -
a/(a+c)
```

**UNIDAD DE PROCESOS TÉCNICOS Y ANÁLISIS BIBLIOGRÁFICO Y DOCUMENTAL**

**REVISIÓN DE NORMAS TÉCNICAS, RESUMEN Y BIBLIOGRAFÍA**

**Fecha de entrega:** 20 / 12 / 2023

| **INFORMACIÓN DEL AUTOR/A (S)** |
|---|
| **Nombres – Apellidos:** EVELYN MISHEL SÁNCHEZ BARRIGA<br>SEBASTIAN ISRAEL TENESACA BUENAÑO |
| **INFORMACIÓN INSTITUCIONAL** |
| **Facultad:** CIENCIAS |
| **Carrera:** ESTADÍSTICA |
| **Título a optar:** INGENIERA ESTADÍSTICA<br>INGENIERO ESTADÍSTICO |
| **f. Analista de Biblioteca responsable:** Ing. Rafael Inty Salto Hidalgo |

2136-DBRA-UPT-2023