Universidad Nacional Mayor de San Marcos
Universidad del Perú. Decana de América
Facultad de Ciencias Matemáticas
Escuela Profesional de Estadística

Comparación de la regresión logística y redes neuronales en la predicción de la anemia en gestantes, ENDES 2022

TRABAJO DE SUFICIENCIA PROFESIONAL

Para optar el Título Profesional de Licenciada en Estadística

AUTOR

Ada Mariela SÁNCHEZ MEDINA

ASESOR

Dra. Ofelia ROQUE PAREDES

Lima, Perú

2023

Referencia bibliográfica

Sánchez, A. (2023). *Comparación de la regresión logística y redes neuronales en la predicción de la anemia en gestantes, ENDES 2022*. [Trabajo de suficiencia profesional de pregrado, Universidad Nacional Mayor de San Marcos, Facultad de Ciencias Matemáticas, Escuela Profesional de Estadística]. Repositorio institucional Cybertesis UNMSM.

Metadatos complementarios

Datos de autor	
Nombres y apellidos	Ada Mariela Sánchez Medina
Tipo de documento de identidad	DNI
Número de documento de identidad	77484549
URL de ORCID	https://orcid.org/0009-0006-4814-4814
Datos de asesor	
Nombres y apellidos	Ofelia Roque Paredes
Tipo de documento de identidad	DNI
Número de documento de identidad	06243124
URL de ORCID	https://orcid.org/0000-0001-8280-021X
Datos del jurado	
Presidente del jurado	
Nombres y apellidos	Zoraida Judith Huamán Gutiérrez
Tipo de documento	DNI
Número de documento de identidad	09890094
Miembro del jurado 1	
Nombres y apellidos	Hugo Marino Rodríguez Orellana
Tipo de documento	DNI
Número de documento de identidad	40162362
Datos de investigación	
Línea de investigación	A.3.2.6. Análisis de Datos y Modelamiento de Problemas de la Sociedad

Grupo de investigación	No aplica.
Agencia de financiamiento	Sin financiamiento.
Ubicación geográfica de la investigación	Universidad Nacional Mayor de San Marcos País: Perú Departamento: Lima Provincia: Lima Distrito: Cercado de Lima Coordenadas geográficas: Latitud: -12.0657518 Longitud: -77.0351583
Año o rango de años en que se realizó la investigación	Mayo 2023 – Setiembre 2023
URL de disciplinas OCDE	Estadísticas, Probabilidad https://purl.org/pe-repo/ocde/ford#1.01.03

ACTA DE SUSTENTACIÓN DEL TRABAJO DE SUFICIENCIA PROFESIONAL PARA LA OBTENCIÓN DEL TÍTULO PROFESIONAL DE LICENCIADA EN ESTADÍSTICA
(PROGRAMA DE TITULACIÓN PROFESIONAL 2023)

En la UNMSM – Ciudad Universitaria – Facultad de Ciencias Matemáticas, siendo las 19.... horas del sábado 21 de octubre del 2023, se reunieron los docentes designados como Miembros del Jurado Evaluador (PROGRAMA DE TITULACIÓN PROFESIONAL 2023): Dra. Zoraida Judith Huamán Gutiérrez (PRESIDENTE), Mg. Hugo Marino Rodríguez Orellana (MIEMBRO) y la Dra. Ofelia Roque Paredes (MIEMBRO ASESOR), para la sustentación del Trabajo de Suficiencia Profesional titulado: "**COMPARACIÓN DE LA REGRESIÓN LOGÍSTICA Y REDES NEURONALES EN LA PREDICCIÓN DE LA ANEMIA EN GESTANTES, ENDES 2022**", presentado por la señorita **Bachiller ADA MARIELA SÁNCHEZ MEDINA**, para optar el Título Profesional de Licenciada en Estadística.

Luego de la exposición del Trabajo de Suficiencia Profesional, la Presidente invitó a la expositora a dar respuesta a las preguntas formuladas.

Realizada la evaluación correspondiente por los Miembros del Jurado Evaluador, la expositora mereció la aprobación .....Sobresaliente....., con un calificativo promedio de .Diecisiete(17)

A continuación, los Miembros del Jurado Evaluador dan manifiesto que la participante **Bachiller ADA MARIELA SÁNCHEZ MEDINA**, en vista de haber aprobado la sustentación de su Trabajo de Suficiencia Profesional, será propuesta para que se le otorgue el Título Profesional de Licenciada en Estadística.

Siendo las 19:30. horas se levantó la sesión firmando para constancia la presente Acta.

Dra. Zoraida Judith Huamán Gutiérrez
PRESIDENTE

Mg. Hugo Marino Rodríguez Orellana
MIEMBRO

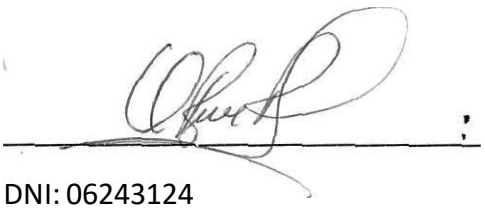
Dra. Ofelia Roque Paredes
MIEMBRO ASESOR

CERTIFICADO DE SIMILITUD

Yo, Ofelia Roque Paredes en mi condición de asesora acreditada con Resolución Decanal N° 001614-2023-D-FCM/UNMSM del Trabajo de Suficiencia Profesional, cuyo título es "COMPARACIÓN DE LA REGRESIÓN LOGÍSTICA Y REDES NEURONALES EN LA PREDICCIÓN DE LA ANEMIA EN GESTANTES, ENDES 2022", presentado por la bachillera ADA MARIELA SÁNCHEZ MEDINA, para optar el título de Licenciada en Estadística.

Certifico que se ha cumplido con lo establecido en la Directiva de Originalidad y de Similitud de Trabajos Académicos, de Investigación y Producción Intelectual. Según la revisión, análisis y evaluación mediante el software de similitud textual, el documento evaluado cuenta con el porcentaje de **19%** de similitud, nivel **PERMITIDO** para continuar con los trámites correspondientes y para su publicación en el repositorio institucional.

Se emite el presente certificado en cumplimiento de lo establecido en las normas vigentes, como uno de los requisitos para la obtención del título correspondiente.



DNI: 06243124

Dra. Ofelia Roque Paredes

Huellá Digital

Dedicatoria

A mi mayor inspiración, mis padres Leny y Daniel, que gracias a su amor incondicional y su sabiduría me guían cada día. A mi hermana Claudia, que es un ejemplo de perseverancia.

RESUMEN

La anemia es una enfermedad silenciosa que no presenta síntomas representativos hasta que se manifiestan signos importantes. Por este motivo, no es percibido fácilmente por quienes la padecen. Esta es de especial preocupación si se presenta en personas vulnerables, como es el caso de los niños y las personas de la tercera edad. Para esto, es de vital importancia la detección en las primeras etapas de la vida, en especial de las gestantes, ya que podría ayudar a tomar medidas de forma preventiva. Actualmente, en el área de salud, no sólo se recurren a pruebas diagnósticas para la detección de la anemia, sino que también se recurre a técnicas predictivas para determinar si se padece la enfermedad o no. Es por ello que se propone una comparación entre dos de las técnicas más usadas en el campo predictivo. En el proceso, se encontró que el conjunto de datos presentaba desbalance en la variable dependiente (presencia de anemia) por lo que se empleó la técnica de muestreo SMOTE. Luego, en el análisis de las variables, se encontró que el número de hijos, edad, grado de instrucción y nivel de riqueza tienen influencia en la presencia de anemia. Finalmente, se encontró que las redes neuronales mostraron un mejor desempeño que la regresión logística.

Palabras clave: regresión logística, redes neuronales, SMOTE, anemia.

ABSTRACT

Anaemia is a silent disease that does not present representative symptoms until important signs are manifested. For this reason, it is not easily perceived by those who suffer from it. This is of special concern if it occurs in vulnerable people, such as children and elderly people. For this, it is of vital importance the detection in the early stages of life, especially of pregnant women, as it could help take preventive measures. Currently, in the health area, not only diagnostic tests are used to detect anaemia, but predictive techniques are also used to determine whether one suffers from the disease or not. That is why a comparison is proposed between two of the most used techniques in the predictive field. In the process, it was found that the data set presented an imbalance in the dependent variable (presence of anemia) so the SMOTE sampling technique was used. Then, in the analysis of the variables, it was found that the number of children, age, level of education and level of wealth have an influence on the presence of anemia. Finally, it was found that neural networks showed better performance than logistic regression.

Keywords: logistic regression, neural networks, SMOTE, anaemia.

Índice

Lista de tablas

Tabla de figuras

Tabla de anexos

I. INTRODUCCIÓN

La Escuela Nacional de Estadística e Informática es una institución pública que brinda servicios de capacitación especializada a los miembros del Instituto Nacional de Estadística e Informática y al público en general, que tiene por finalidad promover la formación y la capacitación continua, además de mantener actualizados a los profesionales con las últimas herramientas en tecnología.

La entidad cuenta con diversas áreas, siendo una de ellas es la Dirección Técnica. En esta área se desarrollan, principalmente, proyectos de Investigación. Uno de ellos es el "Impacto de la Inversión Pública en la Calidad de Vida en Lima Metropolitana entre los años 2010 y 2016", el cual fue elaborado en el periodo en el que fui colaboradora de dicha institución en mi periodo de Prácticas Pre Profesionales.

La línea de investigación del presente trabajo es el análisis de datos y modelamiento de problemas en la sociedad, ya que se empleará información de la Encuesta Demográfica y de Salud Familiar – ENDES, la cual será analizada, y posteriormente se le aplicarán dos modelos predictivos que serán comparados para determinar cuál es el mejor.

El objetivo de esta investigación es la comparación de dos modelos predictivos para encontrar el óptimo y poder predecir la anemia en gestantes.

II. DESCRIPCIÓN DE LA ACTIVIDAD

2.1 Breve Reseña de la Empresa

2.1.1 Antecedentes de la ENEI

Según el Instituto Nacional de Estadística e Informática (INEI, 2021):

> En 1969, mediante Decreto Ley N° 17532 "Ley Orgánica de la Presidencia de la República", se creó la Oficina Nacional de Estadística y Censos - ONEC, con dependencia de la Oficina del Primer Ministro.
>
> Mediante Decreto Legislativo N° 604 de mayo de 1990, se aprueba la "Ley de Organización y Funciones del lnstituto Nacional de Estadística e Informática" donde se precisa que el lnstituto Nacional de Estadística e Informática es un Organismo Público Descentralizado con personería jurídica de derecho público interno, con autonomía técnica y de gestión, dependiente de la Presidencia del Consejo de Ministros.
>
> En este mismo Decreto Legislativo N° 604, en el Título I Capítulo I Artículo 2°, se establece que: "Son objetivos de los Sistemas Nacionales de Estadística e Informática promover la capacitación, investigación y desarrollo de las actividades de Estadística e Informática". (p. 7)

2.1.2 Base Legal de la ENEI

Según el Instituto Nacional de Estadística e Informática (INEI, 2021):

- Constitución Política del Perú.
- Decreto Legislativo N° 604 "Ley de Organización y Funciones del lnstituto Nacional de Estadística e Informática.

- Decreto Supremo N° 018-96-PCM "Reglamento de Organización y Funciones del lnstituto Nacional de Estadística e Informática.
- Decreto Supremo N° 044-91 PCM "Modificar el Artículo 107 del Reglamento de Organización y Funciones del Instituto Nacional de Estadística e Informática.
- Ley del Sistema Estadístico Nacional.

2.1.3 Misión de la ENEI

La Escuela Nacional de Estadística e Informática tiene como misión (INEI, 2021):

> Proporcionar servicios de capacitación especializada en Estadística e Informática a los trabajadores del INEI, del Sistema Estadístico Nacional y público en general, incidiendo en la mejora e innovación de los procesos de trabajo y de gestión, para la obtención de productos y servicios de calidad, elevación de la cultura estadística, así como, promover la investigación y el intercambio conceptual y metodológico, con organismos nacionales e internacionales. (p. 7)

2.1.4 Visión de la ENEI

La Escuela Nacional de Estadística e Informática, tiene como finalidad (INEI, 2021):

> Ser un órgano moderno, con infraestructura adecuada, competitiva, promotora del aprendizaje continuo y permanente, con reconocido prestigio internacional, orientada a satisfacer las necesidades de capacitación de nuestros usuarios, internos y externos, incorporando nuevos enfoques de metodologías estadísticas, de administración y tecnologías modernas de información y comunicaciones. (p. 7)

2.2 Organigrama de la entidad

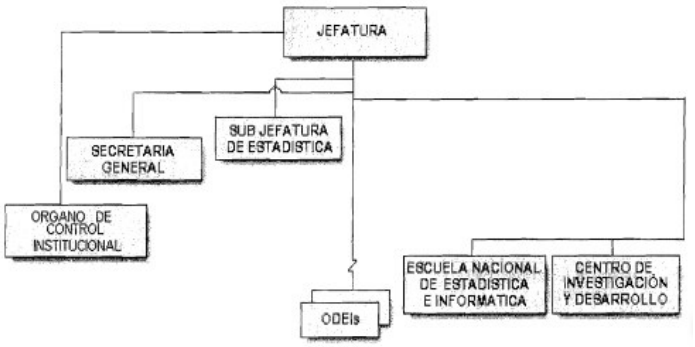
Estructura de la Escuela Nacional de Estadística e Informática

La Escuela Nacional de Estadística e Informática se divide en dos Direcciones Ejecutivas y son las siguientes:

- La Dirección Ejecutiva Académica, que se encarga del desarrollo y coordinación de capacitaciones para profesionales y público en general en el campo de la Estadística e Informática.
- La Dirección Ejecutiva Administrativa, que tiene como objetivo el apoyo en el desarrollo de las capacitaciones, así como de la distribución de recursos necesarios para llevar a cabo dichas actividades de enseñanza.

Figura 1

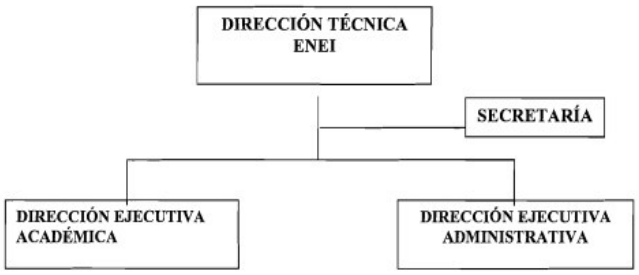
Organigrama de la entidad



Nota. Obtenido de la Escuela Nacional de Estadística e Informática

Figura 2

Diagrama estructural de la entidad



Nota. Obtenido de la Escuela Nacional de Estadística e Informática

2.3 Problemática

La anemia es una enfermedad que consiste en la disminución del número de glóbulos rojos, lo que ocasiona una serie de síntomas que hacen que una persona pueda desarrollar diversos problemas de salud, o incluso, la muerte. La anemia afecta principalmente a niños, gestantes y adultos mayores, por lo que se considera un problema de salud pública. Es considerado un problema de salud pública que afecta principalmente a niños, gestantes y adultos mayores.

Según la OMS (2020), la anemia afecta al 37% de embarazadas, al 30% de mujeres de 15 a 49 años y al 20% de niños de 6 a 59 meses de edad. En el mundo hay 2,000 millones de personas (cerca del 30% de la población) que padece anemia, en Latinoamérica, afecta aproximadamente al 22% de la población y en el Perú, podemos referir que afectó al 32% de la población.

A pesar de que los esfuerzos realizados por el estado para reducir estas cifras, no se muestran cambios significativos en los porcentajes de anemia. Este problema no solo requiere

atención de parte del estado, sino también una gestión pública adecuada, un sistema de salud capacitado y una población dispuesta a contribuir con la reducción de estas cifras.

> La gestación es una de las etapas en las que se requiere grandes cantidades de hierro, debido a que se necesitan nutrientes para el feto. Sin embargo, en esta etapa ocurre una disminución en la concentración de hemoglobina, que se puede observar a partir del tercer trimestre de embarazo. (Gonzales y Olavegoya, 2019)

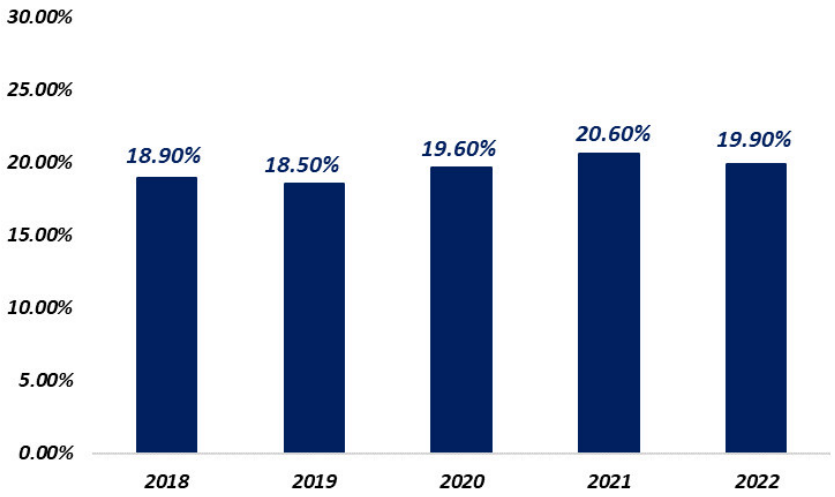
La regresión logística es uno de los modelos más usados para analizar datos epidemiológicos, cuyo objetivo principal es el de modelar cómo influye la probabilidad de ocurrencia de un suceso, frecuentemente dicotómico (Ortiz, Z. 2005).

En estos últimos años, las redes neuronales artificiales están teniendo una gran repercusión en distintos campos del conocimiento, en especial en la medicina (Sharpe y Caleb, 1994).

Los modelos predictivos de regresión logística y redes neuronales son utilizados hoy en día para realizar predicciones en el campo de la medicina humana. Por un lado, la regresión logística, que es uno de los modelos más utilizado en epidemiología y el de redes neuronales, que cada vez más va ganando terreno en el campo de las ciencias de la salud, gracias a su capacidad de aprendizaje y versatilidad.

Figura 3

Evolución de la proporción de anemia en el Perú por periodo



Nota. Proporción de anemia en gestantes que acuden a establecimientos de salud. Fuente: Instituto Nacional de Salud/Sistema de Información del Estado Nutricional

2.4 Objetivo general

- Comparar los modelos predictivos de regresión logística y las redes neuronales para la predicción de la anemia en gestantes.

2.5 Objetivos específicos

- Identificar las coincidencias de las variables que tienen influencia en la anemia en gestantes entre la regresión logística y las redes neuronales.
- Identificar la frontera de decisión de las redes neuronales y regresión logística.
- Identificar cuál de los dos métodos comete menos error.

2.6 Breve descripción de la metodología

A continuación, se describen los siguientes pasos para el análisis de los datos:

- Para el tratamiento de los datos se utilizarán los softwares R y Python.
- Análisis descriptivo de las variables
- Segmentación de la data en entrenamiento y prueba
- Ajuste de los modelos predictivos
- Evaluación
- Comparación de los modelos
- Elección e interpretación del mejor modelo

III. MARCO TEÓRICO

3.1 Modelos Lineales Generalizados

Los modelos lineales cumplen los siguientes supuestos:

- Los errores presentan distribución normal
- La varianza es constante
- Existe relación lineal entre la variable dependiente e independiente.

Sin embargo, en algunos casos, debido a la naturaleza de la variable respuesta, alguno de estos supuestos no se cumple. Este problema se puede solucionar transformando la variable respuesta. Pero aún con esta corrección, estas transformaciones podrían no cumplir con los supuestos, es por ello que, de forma alternativa podemos utilizar los modelos lineales generalizados.

Según McCullagh, P. y Nelder, J. (1989), "Los modelos lineales generalizados son una extensión de los modelos lineales clásicos" (p.26).

Los modelos lineales generalizados tienen los siguientes componentes:

- **Componente aleatoria:** Identifica la variable respuesta y su distribución de probabilidad.

Está formado por una v.a. Y con observaciones independientes $(y_1, y_2, \dots, y_n)$.

y_i puede ser una variable de recuento, binaria, o continua. Estos pueden incluirse dentro de la familia exponencial.

$$f(y_i|\theta_i) = a(\theta_i) * b(y_i) * \exp[y_i Q(\theta_i)]$$

- **Componente sistemática**: Especifica las variables explicativas (independientes o predictoras) utilizadas en la función predictora lineal. Estas variables se relacionan por medio de una combinación lineal:

$$\alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

Esta expresión se denomina predictor lineal, la cual se pude expresar como un vector $(\eta_1, \eta_2, \eta_3, \ldots, \eta_n)$ tal que:

$$\eta_i = \sum_j \beta_j x_{ij}$$

Donde:

$x_{ij}: j - esimo\ valor\ del\ i - esimo\ individuo$

- **Función de enlace:** Es una función expresada como una combinación lineal de las variables predictoras. Se tiene la esperanza de Y como $\mu = E(Y)$, entonces la función enlace relaciona mu con el predictor lineal mediante una función g(.) que se denota como:

$$g(\mu) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

Las funciones de enlace más comunes son:

Tabla 1

Funciones de enlace más comunes

Nombre de función	Función de enlace: $\eta = g(\mu)$	$\mu = g^{-1}(\eta)$
Identidad	μ	η
Logaritmo	$\log(\mu)$	$\exp(\eta)$
Logit	$\log(\frac{\mu}{1-\mu})$	$\frac{\exp(\eta)}{1+\exp(\eta)}$
Reciproca	$\frac{1}{\mu}$	$\frac{1}{\eta}$
Potencia	μ^k	$\eta^{\frac{1}{k}}$
Raíz cuadrada	$\sqrt{\mu}$	η^2
Probit	$\Phi^{-1}(\mu)$	$\Phi(\eta)$
Log-log	$\log(-\log(\mu))$	$\exp(-\exp(\eta))$
C-log-log	$\log(-\log(1-\mu))$	$1-\exp(-\exp(\eta))$

Nota. Extraído de: Madsen, H. y Thyregod, P. (2010, p. 103)

De este modo, la función enlace relaciona las componentes sistemática y aleatoria.

La elección de la función de enlace es sugerida por la forma funcional de la relación entre la respuesta y las variables explicativas.

3.1.1 Regresión Logística

Los modelos de regresión son una parte esencial en la mayoría de las tareas que involucre la relación entre una variable dependiente y otras independientes en el campo del análisis de datos. En muchas ocasiones, la variable respuesta es discreta y puede tomar uno o más valores.

El modelo de regresión logística se denota de la siguiente forma:

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$$

Convirtiendo la fórmula anterior, obtenemos la transformación logit. Esta es definida en términos de $\pi(x)$ como:

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

$$g(x) = \beta_0 + \beta_1 x$$

Según Fisher (1992), el método general de estimación que permite la obtención de la función de mínimos cuadrados en el modelo de regresión lineal es el método de máxima verosimilitud. Este genera valores para los parámetros desconocidos que maximizan la probabilidad de obtener el conjunto de datos observado. Para la aplicación de este método se debe construir una función llamada función de verosimilitud. Esta representa la probabilidad de los datos observados en función de los parámetros desconocidos. Los estimadores de máxima verosimilitud de estos parámetros son elegidos para maximizar esta función.

3.2 Redes Neuronales

- **Reseña Histórica**

Alan Turing (1936) fue un matemático Británico que estudió el cerebro y su relación con la computación, además de ser el primero en dar aporte a las redes neuronales.

> El primer modelo de red neuronal artificial se produce con McCulloch y Pitts (1943) mediante circuitos eléctricos. Seis años después, Donald Hebb (1949) intentó encontrar semejanzas entre el aprendizaje y la actividad nerviosa y, posteriormente sus investigaciones formaron las bases teóricas de las redes neuronales. En 1957, Frank Rosenblatt, psicólogo estadounidense que realizó estudios en el área de la inteligencia artificial desarrolló el perceptrón que tenía la capacidad de reconocer patrones luego de haber aprendido sin entrenamiento previo.

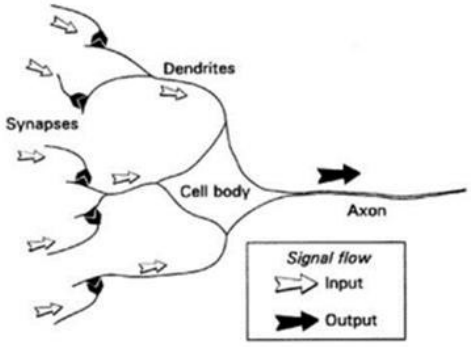
La primera red neuronal fue desarrollada por Windorf y Hoff (1960). Ellos crearon el modelo ADALINE y fue aplicado a un caso real. Luego, Minsky y Papert (1969), demostraron que el perceptrón no tenía la capacidad para solucionar problemas simples y, debido a ello se desencadenó la “etapa oscura” de las redes neuronales. Por otro lado, Paul Werbos (1974) impulsó la idea básica del algoritmo backpropagation, siendo este redescubierto en 1986. Luego de 1986 se continuaron con las investigaciones y el desarrollo de las redes neuronales, por lo que esta etapa es considerada como el renacimiento.

- **Definición**

Las redes neuronales son técnicas de Machine Learning que simulan el sistema nervioso humano mediante tareas de aprendizaje. Estas consisten en un grupo de unidades de procesamiento simple que se comunican mediante el envío de señales entre sí a través de un gran número de conexiones ponderadas.

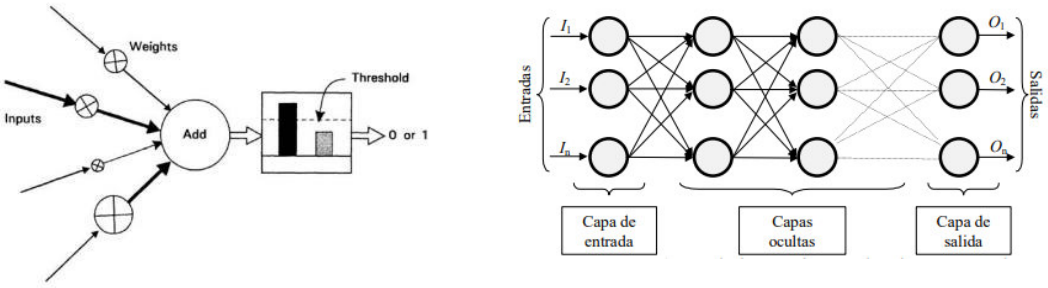
Figura 4

Neurona humana



Nota. Estructura de una neurona humana.
Fuente: Gurney, Kevin (1997).

Figura 5 *Estructura de una red neuronal artificial*



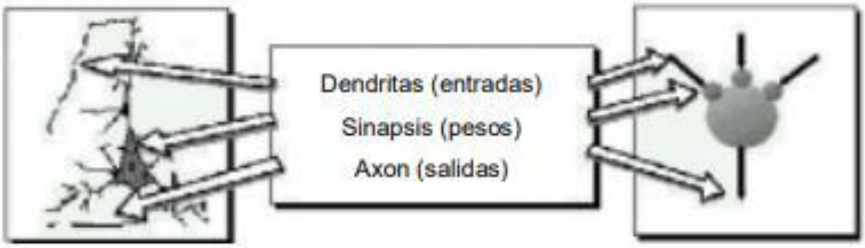
Nota. Estructura y elementos de una red neuronal artificial.
Fuente: Matich, D. (2001).

Está formada por neuronas distribuidas en 3 capas principales: capa de entrada, capa oculta y capa de salida. Los datos son ingresados por la capa de entrada, luego pasan por medio de la capa oculta (que podría contener una o más capas) y desembocan en la capa de salida.

Como se indicó inicialmente, la neurona artificial se asemeja a la neurona biológica, por ende, presentan componentes análogos:

Figura 6

Similitudes entre la neurona biológica y artificial



Nota. Similitudes entre una neurona humana y una neurona artificial.
Fuente: Matich, D. (2001).

Las redes neuronales están compuestas por:

- **Función de entrada**

Esta función permite combinar los valores de entrada junto con los pesos ingresados a la neurona. Algunas de estas pueden ser: función sumatoria, función productoria o función máxima.

- **Función de activación**

La función de activación se encarga de transferir la información producida por la combinación lineal de los pesos y las entradas. Debido a que el objetivo de la red neuronal es que esta tenga la capacidad de aprender temas complejos, las funciones de activación generarán la no linealidad de los modelos.

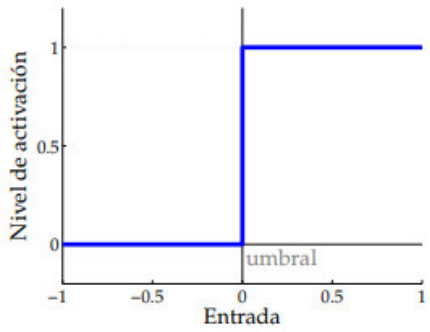
A continuación, se mencionan algunos ejemplos de funciones de activación:

- **Función identidad, escalón o threshold:**

Esta función contiene en el eje x el valor ya ponderado de los inputs y sus correspondientes pesos y, en el eje y, se tiene el valor de la función escalón.

Figura 7

Función identidad



Nota: Extraído del libro Redes Neuronales de Berzal, F. (2018)

Donde, el nivel de activación está representado por:

$$y = \phi(x)$$

Además,

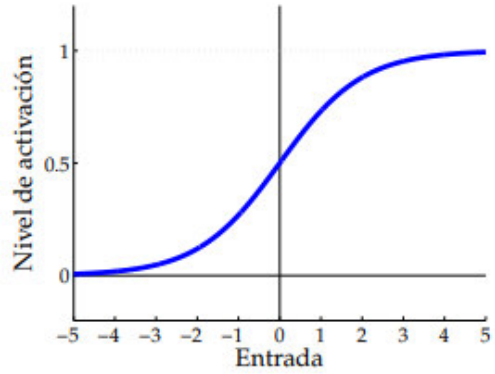
$$\phi(x) = \begin{cases} \mathbf{0}, & \boldsymbol{si\ x < 0} \\ \mathbf{1}, & \boldsymbol{si\ x \geq 0} \end{cases}$$

- **Función sigmoidal o sigmoide**

A diferencia de la función anterior, esta presenta cambios suaves (es decir, es derivable). Para valores positivos de x, la función se aproxima a 1 y para valores negativos de x, la función se aproxima a 0.

Figura 8

Función sigmoidal



Nota: Extraído del libro Redes Neuronales de Berzal, F. (2018)

Donde, el nivel de activación está representado por: $y = \phi(x)$

Además: $\phi(x) = \frac{\mathbf{1}}{\mathbf{1} + e^{-x}}$

Esta función tiene las siguientes características:

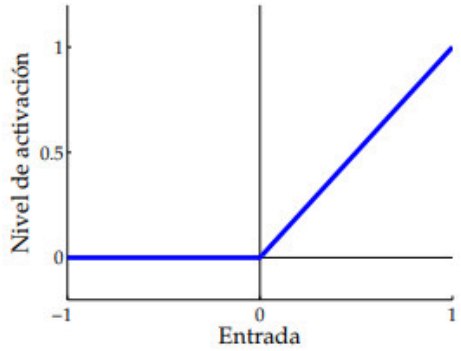
- Presenta un buen rendimiento en la última capa
- Está acotada entre 0 y 1
- Presenta lenta convergencia
- Satura el gradiente
- Suele ser utilizada en capas de salida para variables de respuesta binarios

❖ **Función rectificadora (ReLU)**

Esta función transforma los valores de entrada anulando los valores negativos y conservando los valores positivos.

Figura 9

Función ReLU



Nota: Extraído del libro Redes Neuronales de Berzal, F. (2018)

Donde, el nivel de activación está representado por: $y = \phi(x)$

Además: $\phi(x) = \boldsymbol{max}(\boldsymbol{x}, \mathbf{0})$

Esta función tiene las siguientes características:

- Presenta un buen desempeño en redes convolucionales

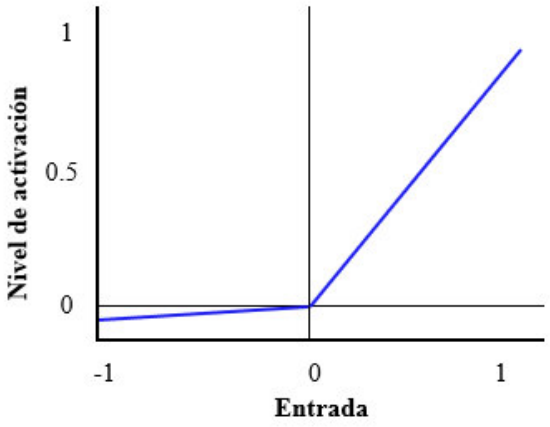
- Tiene buen comportamiento con imágenes
- Se pueden morir neuronas
- No está acotada
- Presenta activación sparse – sólo si son positivos

❖ **Función Leaky ReLU**

Esta función transforma los valores de entrada multiplicando los valores negativos por un coeficiente rectificador y los positivos se mantienen constantes.

Figura 10

Función Leaky ReLU



Nota: Extraído del libro Redes Neuronales de Berzal, F. (2018)

Donde, el nivel de activación está representado por: $y = \phi(x)$

Además: $\phi(x) = \begin{cases} \mathbf{0}, & \boldsymbol{si\ x < 0} \\ \boldsymbol{ax}, & \boldsymbol{si\ x \geq 0} \end{cases}$

Esta función tiene las siguientes características:

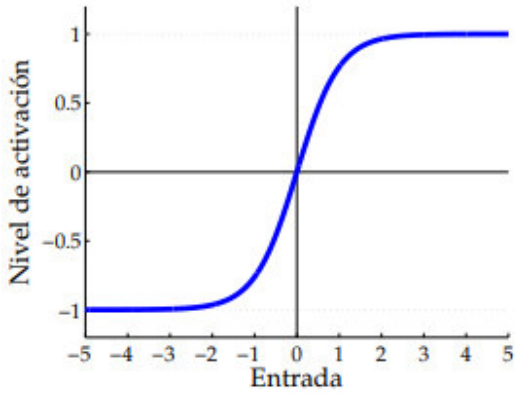
- Es similar a la función ReLU
- Tiene un buen desempeño en redes convolucionales
- Presenta buen comportamiento con imágenes
- No está acotada
- Penaliza los valores negativos a través de un coeficiente rectificador

❖ **Función tangente hiperbólica**

Esta función otorga valores de salida que oscilan entre -1 y +1.

Figura 11

Función tangente hiperbólica



Nota: Extraído del libro Redes Neuronales de Berzal, F. (2018)

Donde, el nivel de activación está representado por: $y = \phi(x)$

Además: $\phi(x) = \frac{1 - e^{-2x}}{1 + e^{-2x}}$

Esta función tiene las siguientes características:

- Presenta un buen desempeño en redes recurrentes
- Se emplea en la decision entre dos opciones
- Está acotada entre -1 y 1
- Es de lenta convergencia
- Es similar a la funcion sigmoide

- **Valores de salida y pesos**

Los valores de la red neuronal pueden ser continuos o categóricos. Para que el procesamiento se haga efectivo, cada sinapsis tendrá asignado un peso. Es necesario el ajuste de estos pesos para obtener una red neuronal que cumpla con la función para la cual fue creada.

- **Función de salida**

Esta función determina el valor transmitido a las neuronas conectadas. Si la función de activación está por debajo de cierto umbral, no se pasa ninguna salida a la siguiente neurona.

Las funciones de salida pueden ser:

- Función identidad: En este caso, la salida es la misma que la entrada.
- Función binaria: $\begin{cases} 1, \ si \ \ act_i \geq \xi_i \quad donde \ \xi_i : umbral \\ 0, \ caso \ contrario \end{cases}$

- **Función de pérdida:**

Esta función se encarga de evaluar la diferencia que existe entre los valores predichos y los valores reales. Cuanto menor es el resultado de esta función, la red neuronal es más eficiente. La reducción de la desviación entre la predicción y el valor real para cierta observación se hace ajustando los distintos pesos de la red neuronal. El objetivo es minimizar esta función, la cual mejora la precisión del modelo. Dicha función debe ser continua y derivable, ya que ciertos algoritmos de aprendizaje calculan la derivada de la función de pérdida. Existe una amplia gama de funciones de pérdida, pero solo vamos a mencionar a las siguientes:

- **Error cuadrático medio:** Esta métrica es usada frecuentemente en problemas de regresión. Aquí se calculan las distancias al cuadrado entre los valores el valor original y el predicho. Se puede utilizar en problemas de regresión.
- **Entropía cruzada categórica:** Esta se utiliza a menudo en modelos de redes cuya salida representa una probabilidad o en problemas de clasificación categórica.
- **Entropía cruzada binaria:** Se deriva de la entropía cruzada categórica pero usada en clasificación binaria y la función de activación es la sigmoide.
- **Entropía cruzada categórica dispersa:** Esta suele ser usada en casos en los que se utilicen números enteros.

- **Optimizadores:**

Estos se encargan de ajustar los parámetros de la red neuronal para minimizar la función de pérdida. Usan el gradiente calculado durante el backpropagation para hacer los ajustes. Entre los ejemplos se encuentran la gradiente descendente estocástica y otras avanzadas como ADAM

y RMSProp que ajusta los ratios de aprendizaje para cada parámetro que permiten un entrenamiento más eficiente y estable.

3.3 Anemia

Según la Organización Mundial de la Salud (OMS, 2017):

> La anemia se define como un trastorno en el cual la concentración de la hemoglobina se encuentra por debajo de cierto valor de corte, reduciendo de este modo la capacidad de sangre para el transporte de oxígeno en el organismo.

Para el diagnóstico de esta enfermedad se necesitan evaluar los niveles de hemoglobina, proteína que se encuentra en la sangre y contiene el mayor porcentaje de hierro del cuerpo humano.

Según el INSN (2021), “Una de las principales causas de la anemia en el Perú es el déficit de consumo de hierro necesario para la formación de hemoglobina. En consecuencia, la gestante podría tener un bebé prematuro o con bajo peso al nacer”.

En la etapa de gestación, esta enfermedad tiene un efecto importante para la madre y para el feto, siendo para este último un impacto mayor, ya que podría presentar serios problemas de salud desde su crecimiento hasta su vida adulta.

Es importante mencionar que esta enfermedad es la que presenta más frecuencia de diagnóstico durante la etapa de gestación, ya que, se evidencian cambios físicos y hormonales en el cuerpo materno al expandirse el volumen corporal para que sea ocupado por el feto.

El riesgo de contraer anemia presenta una tendencia creciente a medida que van transcurriendo los meses de embarazo, lo cual es una preocupación para los países subdesarrollados, en los que la alimentación basada en hierro es muy pobre.

Tipos de anemia

En el periodo de gestación pueden producirse diversos tipos de anemia, entre ellos se encuentran:

- **Anemia del embarazo:** Este tipo de anemia no es considerada muy riesgosa, ya que durante el embarazo las mujeres suelen tener más cantidad de sangre de la normal, lo que hace que los glóbulos rojos en su organismo se diluyan. Pero si estos niveles son muy bajos podría ser preocupante.
- **Anemia ferropénica:** Este tipo de anemia es el más frecuente en las gestantes. Durante la etapa de embarazo, el bebé utiliza los glóbulos rojos de la madre que almacenó antes de quedar embarazada. Sin embargo, la madre puede contraerla si no tiene las reservas necesarias de hierro.
- **Anemia por deficiencia de vitamina B12:** Durante el embarazo, es necesario el consumo de alimentos que prevengan la deficiencia de vitamina B12 tales como alimentos de origen animal y vegetal, ya que es fundamental para la formación de glóbulos rojos.
- **Anemia por deficiencia de folato:** El folato – o también llamado ácido fólico - es una vitamina que, junto con el hierro, colaboran en el crecimiento celular. Si el organismo no cuenta con suficiente ácido fólico podría tener deficiencia de hierro.

Tabla 2

Valores establecidos de concentración de hemoglobina en gestantes y puérperas

	Con anemia (Según niveles de Hemoglobina (g/dL)			Sin anemia
	Severa	Moderada	Leve	
Mujer gestante de 15 años a más	< 7.0	7.0 - 9.9	10.0 - 10.9	> 11.0
Mujer puérpera	< 8.0	8.0 - 10.9	11.0 - 11.9	> 12.0

Nota. Niveles de concentración de hemoglobina.
Fuente: Extraído del Plan Nacional para la reducción y control de la anemia Materno Infantil y la desnutrición crónica en el Perú: 2017-2021

3.4 Desbalanceo de datos

El desbalanceo de datos es un problema que se puede presentar en muchos conjuntos de datos y que puede repercutir en los resultados obtenidos dando falsas conclusiones. Este concepto se puede definir como la desproporción significativa en una de las categorías de la variable dependiente. Es frecuente encontrar este tipo de casos en la detección de fraudes o en la detección de enfermedades raras, en los cuales se puede observar un 0.1% del total de casos. Por ejemplo, si se tiene un conjunto de datos desbalanceados que contiene el 1% de la clase de interés y el 99% de la otra clase, el algoritmo podría predecir todos los casos como pertenecientes a la clase mayoritaria, por ejemplo, en el caso de la detección de cáncer, podría llevar a conclusiones erradas, como indicar al paciente que no tiene la enfermedad cuando en realidad sí la tiene, poniendo en riesgo su vida.

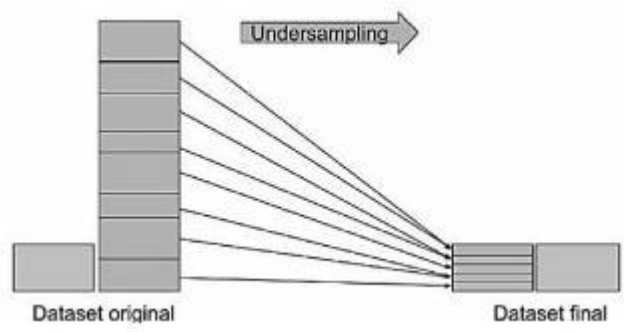
Para solucionar esta dificultad, existen muchas técnicas de muestreo con la finalidad de equilibrar la asignación de categorías del conjunto de datos.

3.4.1 Submuestreo

Según Martinelli (2022), “consiste en la reducción del número de casos de la clase mayoritaria dejando intactos los casos de la clase minoritaria, obteniendo de esta manera un balance” (p. 28). Si bien esta es una solución, ocurre el riesgo de eliminar registros que pueden ser importantes para el proceso de predicción.

Figura 12

Técnica undersampling



Nota: Extraído de Clasificación de Datos Desbalanceados. Martinelli, J. (2022)

La disminución de la clase de mayor frecuencia puede ser efectuada de distintas maneras, para esto se tienen algunos algoritmos de undersampling.

3.4.1.1 Submuestreo aleatorio (RUS).

También llamado *Random Undersampling*. Mediante este método se excluyen muestras de la clase de mayor frecuencia para equilibrar el conjunto de datos. La principal desventaja de este método es que en el proceso de eliminación de algunos registros se podría eliminar

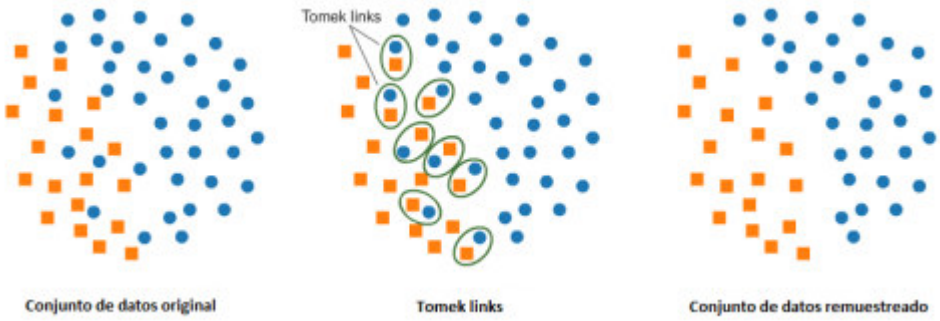
información potencialmente útil para la clasificación. Por otro lado, esta técnica es muy útil en conjuntos de datos que no son tan desequilibrados.

3.4.1.2 Tomek links.

Esta técnica, descrita por Tomek, I. (1976), "Consiste en eliminar muestras de la clase de mayor frecuencia que estén próximas a muestras de la clase de menor frecuencia" (p. 769).

Figura 13

Funcionamiento de la técnica Tomek Links



Nota: Extraído de Clasificación de Datos Desbalanceados. Martinelli, J. (2022)

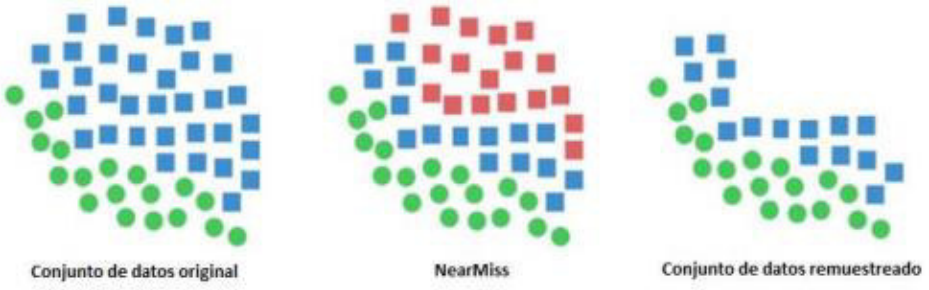
3.4.1.3 Vecinos cercanos.

Esta técnica, descrita por Zhang y Mani (2003), tiene sus bases en el algoritmo del vecino más cercano. Si las muestras de dos clases distintas están próximas entre sí, las muestras de la clase mayoritaria son eliminadas para aumentar los espacios que se encuentran entre las dos clases. Los pasos que se siguen para el funcionamiento de este algoritmo son los siguientes:

1. Se buscan las distancias entre las muestras de las clases de mayor y menor frecuencia.
2. Luego, se seleccionan “n” muestras de la clase con mayor frecuencia que tienen las distancias más pequeñas a las de la clase con menor frecuencia.
3. Si hay “k” muestras de la clase de menor frecuencia, la técnica del vecino más cercano dará como resultado kxn muestras de la clase mayoritaria.

Figura 14

Funcionamiento de la técnica NearMiss



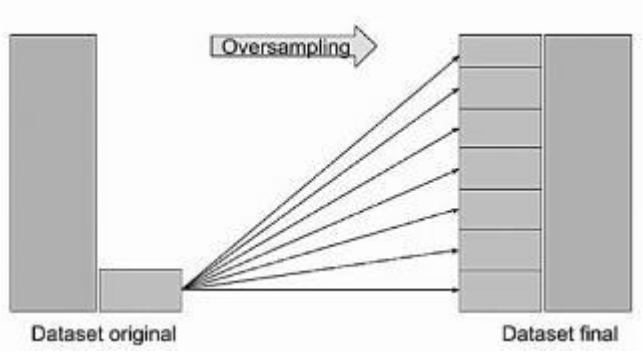
Nota: Extraído de Clasificación de Datos Desbalanceados. Martinelli, J. (2022)

3.4.2 Sobremuestreo

Esta técnica (o en inglés, *Oversampling*), consiste en equilibrar la distribución de los datos aumentando el número de muestras de menor frecuencia dejando intactas las muestras de mayor frecuencia. Sin embargo, la técnica de sobremuestreo tiene una particularidad: añade muestras a la clase de menor frecuencia para equiparar el conjunto de datos y estas muestras son ficticias. De esta manera, se genera ruido para los clasificadores lo que podría resultar en pérdida de rendimiento en cuanto a clasificación.

Figura 15

Funcionamiento de la técnica Oversampling



Nota: Extraído de Clasificación de Datos Desbalanceados. Martinelli, J. (2022)

Dentro de las ventajas de esta técnica, se encuentran las siguientes:

- No se pierde información, ya que mantiene la información inicial y se crean valores ficticios.
- Es recomendable utilizarlo cuando se cuente con un conjunto de datos pequeño.

Dentro de las desventajas, se encuentran:

- Emplea mucho tiempo de procesamiento de datos.
- Es posible que genere muestras ruidosas
- Es probable que ocasione sobreajuste

Se puede incrementar la clase de mayor frecuencia de diversos modos, es por ello que se presentan algunos algoritmos de oversampling.

3.4.2.1 Sobremuestreo aleatorio (ROS).

El Sobremuestreo Aleatorio (o por sus siglas en inglés *Random OverSampling*), descrito por Van-Hulse (2007), "Consiste en el equilibrio del conjunto de datos replicando las muestras de la clase con menor frecuencia" (pp. 935-942). Una de las ventajas de este método es que no es posible la pérdida de información. Por otro lado, un obstáculo que se puede presentar es el sobreajuste, además de un gran coste computacional si se observa una gran diferencia en el equilibrio de los datos.

3.4.2.2 Sobremuestreo sintético (SMOTE).

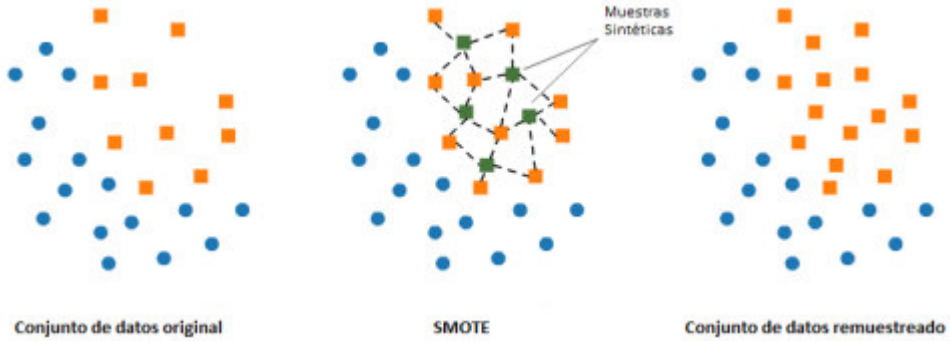
El Sobremuestreo Sintético (o por sus siglas en inglés *Synthetic Minority Oversampling Technique*), descrito por Chawla (2002):

> Consiste en la generación de muestras artificiales para equiparar los datos mediante la regla del vecino más cercano. La creación de estas muestras se realiza a través de la transpolación de muestras nuevas. Luego, selecciona una muestra de la clase de menor frecuencia y junto con los vecinos más cercanos elije uno y crea otra nueva muestra. (p. 321)

Una de las ventajas de esta técnica, en comparación con otras es que evita el sobreajuste, que ocurre cuando se crean copias idénticas de muestras minoritarias al conjunto de datos general.

Figura 16

Funcionamiento de la técnica SMOTE



Nota: Extraído de Clasificación de Datos Desbalanceados. Martinelli, J. (2022)

3.4.2.3 Muestreo sintético adaptativo (ADASYN).

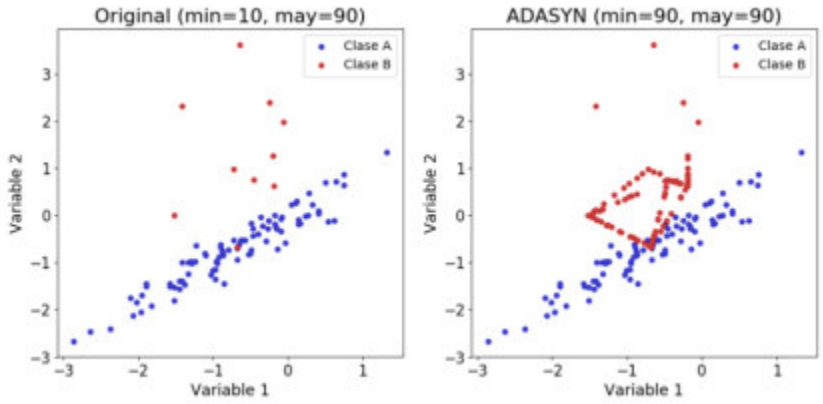
Este tipo de muestreo es una extensión del Sobremuestreo Sintético. Según He y García (2009):

> Se basa en la generación de muestras sintéticas adicionales en áreas fronterizas para mitigar el inconveniente de la superposición de capas, pero propone otro método para generar diferentes números de muestras sintéticas dependiendo de la ubicación de cada muestra de la clase de menor frecuencia. (pp. 1322-1328)

Los objetivos principales de este método son: crear muestras ficticias adicionales por medio de la interpolación lineal, esto para disminuir el desequilibrio con la clase de mayor frecuencia y el cambio de la frontera de decisión por medio de la adición de muestras en la zona de la clase con menor frecuencia, realizada por medio de una distribución de densidad.

Figura 17

Funcionamiento de la técnica ADASYN



Nota. Extraído de Clasificación de Datos Desbalanceados. Martinelli, J. (2022)

3.5 Métricas de evaluación del desempeño del modelo

El entrenamiento del modelo es una de las etapas fundamentales del aprendizaje automático en los proyectos de ciencia de datos. Sin embargo, otro aspecto importante es la evaluación de dicho modelo, para conocer cuán confiable es. Para esto, contamos con varias métricas, dentro de las cuales se encuentran:

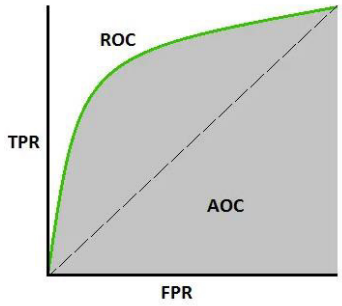
- **Curva ROC**

La curva ROC es una herramienta estadística que se utiliza para la evaluación de la capacidad de discriminación de una prueba diagnóstica dicotómica. En esta se observan curvas en las que se presenta la sensibilidad en función de los falsos positivos para distintos puntos de corte. Es útil para la elección del punto de corte más apropiado de una prueba y conocer el rendimiento de esta.

En la curva ROC existe una región denominada área bajo la curva (o AUC) que mide la capacidad de discriminación de la prueba. El AUC revela qué tan buena es la prueba que estamos utilizando para discriminar.

Figura 18

Curva ROC



Nota. Representación de la curva ROC. Fuente: Obtenido de Understanding AUC-ROC Curve de Narkhede, S. (2018) https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

Interpretación de los valores obtenidos

Con la curva ROC, se obtiene el área bajo la curva, cuyo valor oscila entre 0.5 y 1. Mediante este valor, podemos saber qué tan buena es la clasificación del modelo.

Tabla 3

Valores del área bajo la curva

Valores de AUC	Interpretación
0.5	No hay poder de discriminación
[0.5 - 0.6>	La discriminación es baja
[0.6 - 0.8>	La discriminación es aceptable
[0.8 - 0.9>	La discriminación es muy buena
Mayor a 0.9	La discriminación es muy alta

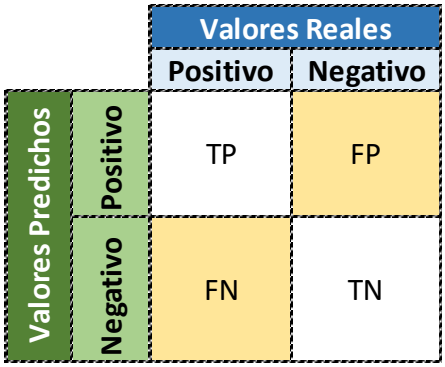
Nota. Adaptado de Llinás, H. (2021).

- **Matriz de confusión**

Es una herramienta que ayuda a resumir el desempeño de los modelos de clasificación. (Chelliah, I. 2020).

Figura 19

Matriz de confusión



Nota. Elaboración propia obtenida a partir de Suresh, A. (2020)
Fuente: What is a confusion matrix?
https://medium.com/analytics-vidhya/what-is-a-confusion-matrix-d1c0f8feda5

Los componentes de la matriz de confusión se enumeran a continuación:

- **Verdaderos positivos (TP):** Se tienen los casos cuyos valores predichos son verdaderos y en realidad son también verdaderos. (ejemplo: que un paciente tenga cierta enfermedad y que la prueba diagnóstica determine que no padece la enfermedad).
- **Verdaderos negativos (TN):** Estos se presentan cuando el valor predicho es falso y el valor real también es falso. (ejemplo: que un paciente no tenga cierta enfermedad y que el diagnóstico de la prueba sea negativo).

- **Falsos positivos (FP):** El valor real es falso pero el valor predicho es verdadero. Este es conocido también como el error de tipo I. (ejemplo: que el paciente no padezca de cierta enfermedad pero que la prueba diagnóstica determine que sí la tiene).
- **Falsos negativos (FN):** El valor real es verdadero pero el valor predicho es falso. Este es conocido también como error de tipo II. (Ejemplo: que el paciente padezca cierta enfermedad, pero la prueba diagnóstica indique que no hay enfermedad).

De los dos últimos casos, podría afirmarse que el error de tipo II es el más grave. Esto porque, si el paciente tiene la enfermedad y no se realiza el tratamiento para combatirla, podría llevarlo a complicaciones en su salud, e incluso, a la muerte. Por otro lado, si se comete el error de tipo I, es decir, si se diagnostica la enfermedad al paciente cuando en realidad está sano, se realizarían más exámenes y se descubrirá que en realidad no está enfermo.

Además, se tienen algunos indicadores de rendimiento:

- **Exactitud (accuracy):** Indica el número de predicciones correctas sobre el total de predicciones.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Tasa de error:** Indica cuántos casos fueron clasificados incorrectamente del total de casos.

$$Tasa\ de\ error = \frac{FP + FN}{TP + TN + FP + FN}$$

- **Precisión:** De todos los valores predichos como positivos, indica cuántos eran en realidad positivos.

$$Precisión = \frac{TP}{TP + FP}$$

- **Exhaustividad (recall):** De todos los casos que eran en realidad positivos, indica cuántos fueron predichos como positivos.

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

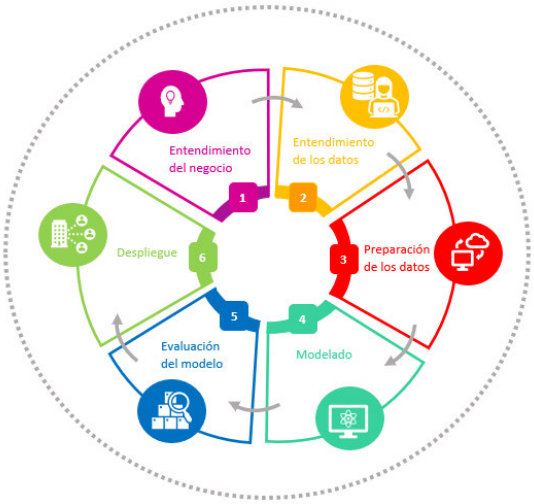
Un buen modelo es aquel que tiene altos ratios de verdaderos positivos y verdaderos negativos y bajos ratios de falsos positivos y falsos negativos.

3.6 Método CRISP - DM

Esta metodología (cuyas siglas en inglés provienen de Cross Industry Standart Process for Data Mining) es utilizada en proyectos que detalla el ciclo de vida de un proyecto cuya finalidad es la extracción de valor de los datos. Este contiene 6 fases que pueden ser adaptadas al proyecto que se desee realizar.

Figura 20

Fases del ciclo CRISP - DM



Nota. Adaptado de La metodología CRISP-DM en Ciencia de Datos. Haya, P. (2021)

Presenta las siguientes fases:

1. Entendimiento del negocio: En esta fase se intenta establecer los objetivos del proyecto partiendo de los objetivos del negocio.
2. Comprensión de los datos: Esta etapa se centra en el conocimiento de los datos por medio de la exploración.
3. Preparación de los datos: Aquí se realiza la tarea crucial del proyecto, que es el de limpieza y análisis.

4. Modelado: El principal objetivo de esta etapa es la búsqueda del modelo más conveniente para los datos.
5. Evaluación: Luego de la elección del modelo (o modelos), se deberá evaluar este por medio de métricas que indiquen qué tan robusto es.
6. Despliegue: En esta última fase, se realiza un plan de implementación de lo obtenido a lo largo de todo el proceso descrito a lo largo del proyecto.

3.7 Descripción teórica y descriptiva de la variable

Una variable es la representación de una característica que puede tomar cualquier valor. Según Grau et al. (2004):

> El concepto de variable siempre está relacionado a las hipótesis de investigación. Una variable puede tomar distintos valores en un conjunto determinado y cuya variación es susceptible de ser medida. Una investigación, cualitativa o cuantitativa, exige la operacionalización de sus conceptos centrales en variables, de esta definición operativa depende el nivel de medición y potencia de las pruebas realizadas.

Es importante la definición de las variables, ya que nos ayuda a identificar qué se va a medir, cómo se medirá y cuáles son las acciones que se deben llevar a cabo para su contrastación.

3.8 Marco teórico

3.8.1 Antecedentes internacionales

En una investigación realizada por Cóndor y Correa (2022), titulado "Comparación entre regresión logística binaria y redes neuronales en la predicción de las causas de mortalidad materna en el Ecuador del año 2020" tuvo como objetivo comparar la regresión logística y las redes neuronales artificiales para identificar cuál de ellas proporciona mejores resultados para la predicción de la mortalidad materna. Para este estudio se emplearon datos obtenidos del registro de defunciones generales del Instituto Nacional de Estadística y Censos del Ecuador. Esta base contenía las siguientes variables: etnia, edad al fallecer, provincia, lugar de ocurrencia del deceso y nivel de instrucción. Se analizaron los resultados y se determinó que la regresión logística proporciona un modelo predictivo más robusto que las redes neuronales.

La investigación realizada por Támara et al. (2016), titulada "Regresión logística y redes neuronales como herramientas para realizar un modelo scoring" tuvo como finalidad analizar el riesgo crediticio de una entidad financiera no regulada por la Superintendencia Financiera de Colombia mediante la comparación de dos modelos de score obtenidos mediante regresión logística y redes neuronales. La data estuvo conformada por una muestra de 43,086 registros del portafolio de consumo perteneciente al periodo enero 2014 – julio 2016 observada 12 meses después (es decir, en el periodo agosto 2016 – julio 2017) de dicha institución. La base fue dividida en 70% para el entrenamiento y 30% para el testeo. En la comparación de los modelos, se tuvo un porcentaje de precisión de 71.65% para regresión logística y 71.66% para red neuronal. Por otro lado, se tuvo un porcentaje de error en la validación de 27.96% para el modelo de regresión logística y 28.11% para la red neuronal. Además, se obtiene un índice de KS para la

regresión logística de 18.44, menor que el de la red neuronal que fue de 18.47. Finalmente se tiene como conclusión que el mejor modelo es el de redes neuronales.

3.8.2 Antecedentes nacionales

La tesis realizada por Mendoza (2023), titulada "Método alternativo para la detección de anemia a través de sus factores asociados en mujeres en edad reproductiva: una aplicación de redes neuronales artificiales", tuvo por objetivo la aplicación de las redes neuronales para el diagnóstico de la anemia en mujeres en edad reproductiva por medio de sus factores asociados. Para este estudio se utilizaron los datos de la Encuesta Demográfica y de Salud Familiar (ENDES) del año 2020, y se tomó una muestra de 9,000 mujeres de entre 15 a 49 años. Se utilizó la red neuronal artificial perceptrón multicapa que presentó un del 81% de resultados correctos (AUC=66.3%), Además, se encontró que los factores asociados a la anemia en mujeres en edad reproductiva fueron: la edad, el IMC, el nivel educativo, el nivel de pobreza y si se realizó una prueba de descarte de anemia.

La investigación realizada por Navarro y Oliva (2020), denominado "Modelo de regresión logística para identificar factores de riesgo y pronóstico de anemia en menores de cinco años en el Perú en el año 2018" tuvo por objetivo identificar los factores asociados a la presencia de anemia en menores de cinco años en el Perú en el año 2018. La información fue tomada de la base de datos de la ENDES 2018 y la muestra estuvo conformada por 16499 menores de cinco años. Los factores de riesgo encontrados fueron: lengua materna de la madre, visitas prenatales de la madre durante el embarazo, grado de instrucción de la madre, edad de la madre, índice de riqueza del hogar, presencia de anemia durante el embarazo, área de residencia, sexo del menor y

edad del menor. Se empleó el modelo de regresión logística binario y este logró clasificar correctamente al 69.5% de los casos.

La tesis de Abanto, A. (2022), denominada “Modelo logístico y redes neuronales para pronóstico de anemia en menores de 3 años”, tuvo como objetivo investigar cuál de las técnicas mencionadas era la mejor para la predicción de anemia en niños menores de 3 años que fueron atendidos en el Hospital Víctor Lazarte Echegaray durante el año 2019. En este estudio se analizó a 214 niños. Los resultados mostraron que, para el modelo de regresión logística, los factores que influyen en la presencia de anemia fueron el lugar de procedencia, el tipo de seguro y la edad (% clasificación correcta: 60%, sensibilidad: 44%, especificidad: 76%). Por otro lado, para el entrenamiento de la red neuronal se utilizaron las variables sexo, edad, tipo de seguro y lugar de procedencia (tasa de clasificación correcta del 72%, una sensibilidad de 80% y una especificidad del 62.50%). Finalmente, se concluyó que el mejor modelo fue el de redes neuronales.

IV. METODOLOGIA

4.1 Tipo y diseño de investigación

La presente investigación tiene un enfoque cuantitativo, ya que, según Hernández, et al., (2014a), "Se sigue un conjunto de procesos secuencial, las variables son medidas en un contexto específico, las mediciones obtenidas son analizadas mediante métodos estadísticos y se extraen conclusiones".

El diseño es observacional, dado que, según Hernández et. al. (2014b), "Sólo se estudiaron las variables; retrospectivo, ya que el evento analizado ocurrió en el pasado; no experimental, pues las variables no han sido manipuladas en el transcurso de toda la investigación" y según Liu y Tucker (2004), "es de corte transversal, ya que los datos fueron recolectados en un único momento en el tiempo".

4.2 Población

Según la ficha técnica de la Encuesta Nacional Demográfica y de Salud Familiar – ENDES, la población está compuesta por las mujeres de todo el Perú en edad reproductiva entre 12 y 49 años.

4.3 Muestra

Para el presente estudio, se detallan las características de la muestra:

- El marco muestral está constituido por la información proveniente de los Censos Nacionales XII de Población y VII de Vivienda del año 2017 y el material cartográfico realizado para la ejecución de la ENDES.

- La muestra se caracteriza por ser bietápica, probabilística de tipo equilibrado, estratificada e independiente a nivel departamental por área urbana y rural.
- La distribución de la muestra de la ENDES 2022 fue estimada previa evaluación de los resultados obtenidos de las encuestas ENDES ejecutadas entre los periodos 2012 a 2020.
- El tamaño de la muestra está compuesto por 17,814 mujeres que participaron de la Encuesta Demográfica y de Salud Familiar del año 2022 que cumplieron con los criterios de inclusión y exclusión.
- Las unidades de muestreo están compuestas por el conglomerado y la vivienda particular en el área urbana y el área de empadronamiento rural y la vivienda particular en el área rural.

4.3.1 Criterios de selección de la muestra

Se debe tener en cuenta los siguientes criterios para la elección de los elementos de la muestra:

- **Criterios de inclusión:**

Se incluirá en el estudio a las mujeres de 12 a 49 años que hayan realizado la prueba diagnóstica de anemia en la etapa de gestación.

Se incluirá en el estudio a las mujeres de 12 a 49 años en edad fértil que se encuentren gestando al momento que se les aplicó la encuesta.

- **Criterios de exclusión:**

Serán excluidas del presente estudio a las mujeres de 12 a 49 años en edad fértil a las que no se les haya realizado la prueba de diagnóstico de anemia en la etapa de gestación.

Serán excluidas del presente estudio a las mujeres de 12 a 49 años en edad fértil que no se encuentren gestando al momento que se les aplicó la encuesta.

4.4 Método de recolección de los datos

La información utilizada en la presente investigación proviene de una fuente secundaria: los datos fueron obtenidos de la Encuesta Demográfica y de Salud Familiar – ENDES 2022.

El método de recolección de datos utilizado fue el de entrevista directa, realizado por personal capacitado que visitó las viviendas seleccionadas. Los informantes fueron mujeres de 12 a 49 años.

4.5 Variables

Tabla 4

Descripción de las variables

Variable	Nombre de la variable	Definición	Tipo de Variable	Escala de medición	Categorías
Dependiente	Diagnóstico de anemia	Indica si la gestante fue diagnosticada con anemia	Cualitativa	Nominal	1. Tiene anemia 0. No tiene anemia
Independiente	Edad	Se refiere a los años de la gestante	Cualitativa	Ordinal	1. Adolescente 2. Joven 3. Adulta
	Región geográfica	Lugar de residencia de la entrevistada	Cualitativa	Nominal	1. Lima Metropolitana 2. Resto Costa 3. Sierra 4. Selva
	Estado civil	Estado civil de la entrevistada	Cualitativa	Ordinal	0. Soltera 1. Casada 2. Conviviente 3. Viuda 4. Divorciada 5. Separada

Grado de instrucción	Último nivel educativo aprobado	Cualitativa	Ordinal	0. Inicial 1. Primaria 2. Secundaria 3. Superior No Universitario 4. Superior Universitario 5. Posgrado
Nivel de pobreza	Índice de riqueza de la gestante	Cualitativa	Ordinal	1. Más pobre 2. Pobre 3. Medio 4. Rico 5. Más rico
Número de hijos	Número de hijos de la gestante (incluyendo los no vivos)	Cualitativa	Nominal	
Seguro de salud	Indica si la entrevistada cuenta o no con un seguro médico	Cualitativa	Nominal	0. No 1. Si
Información de alimentación durante el embarazo	Indica si el médico le informó acerca de su alimentación en el embarazo	Cualitativa	Nominal	0. No 1. Si
Consumo de suplementos de hierro durante el embarazo	Indica si a la gestante le recetaron suplementos a base de hierro	Cualitativa	Nominal	0. No 1. Si
Tratamiento con hierro en embarazo	Indica si durante el embarazo la gestante consumió alimentos ricos en hierro de acuerdo a las indicaciones del personal de salud	Cualitativa	Nominal	0. No 1. Si

4.6 Metodología aplicada

Esta investigación se realizará bajo el enfoque de la metodología CRISP – DM cuyos pasos se indican a continuación:

1. Entendimiento del negocio: La información para la presente investigación se obtuvo de la Encuesta Demográfica y de Salud Familiar – ENDES 2022. Se tomaron datos de la página del INEI, en la sección microdatos. Aquí se buscó la base relacionada al ENDES y se descargaron los módulos correspondientes.
2. Comprensión de los datos: Los datos de la encuesta fueron recolectados mediante la técnica de entrevista directa. Luego de su recopilación, estas fueron almacenadas en bases de datos divididas en módulos.
3. Análisis de datos y elección de variables: Se cuenta con una base de datos que será explotada para su posterior análisis. Esto comenzará con la limpieza de los datos, que es de vital importancia antes de realizar cualquier estudio. Para esta fase, se utilizará el software Python, apoyado en la interfaz Anaconda Navigator y en el entorno de trabajo web Jupyter Notebook. Aquí se realizó el análisis exploratorio mediante gráficos y estadísticos resumen.
4. Modelado: Se hará la división de la data en muestra de entrenamiento y validación. Posteriormente, se aplicará el modelo de regresión logística y las redes neuronales. Para las redes neuronales se definirán la estructura de las capas y las funciones de activación a emplear.
5. Evaluación: Para evaluar la calidad de las predicciones, se examinarán los modelos con la matriz de confusión y el área bajo la curva (AUC).

V. RESULTADOS

En esta sección se muestran los resultados obtenidos del análisis de los datos, las técnicas aplicadas, los modelos empleados y la evaluación de estos.

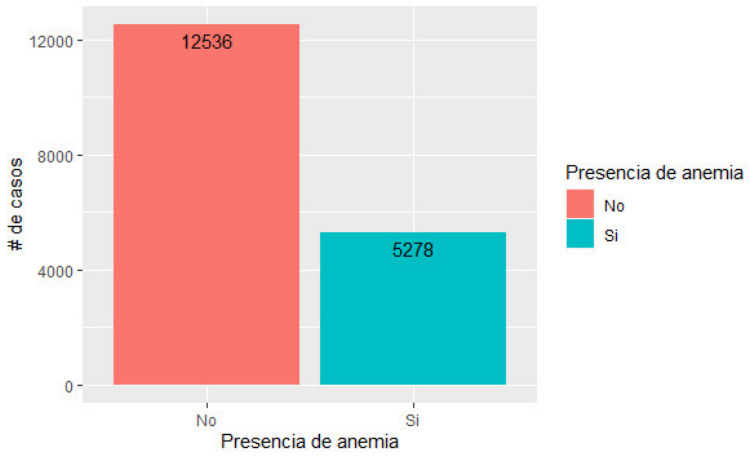
Se tiene una base con 17,814 registros y 10 variables: 9 independientes y 1 dependiente (presencia de anemia: 0. No, 1: Si). Es importante mencionar que la base contaba con valores perdidos y se realizó la limpieza correspondiente. Se decidió retirar estos casos del análisis debido a que no representaban una cantidad significativa de los datos.

5.1 Análisis Exploratorio de Datos

Se muestra un primer análisis con relación a la variable dependiente: se tiene que el 70.4% de las gestantes tiene anemia y, en contraste, el 29.6% no fue diagnosticada con anemia.

Figura 21

Porcentaje de casos de presencia de anemia



Nota. Elaboración propia obtenida a partir del conjunto de datos

Tabla 5

Frecuencia y porcentaje de casos de presencia de anemia

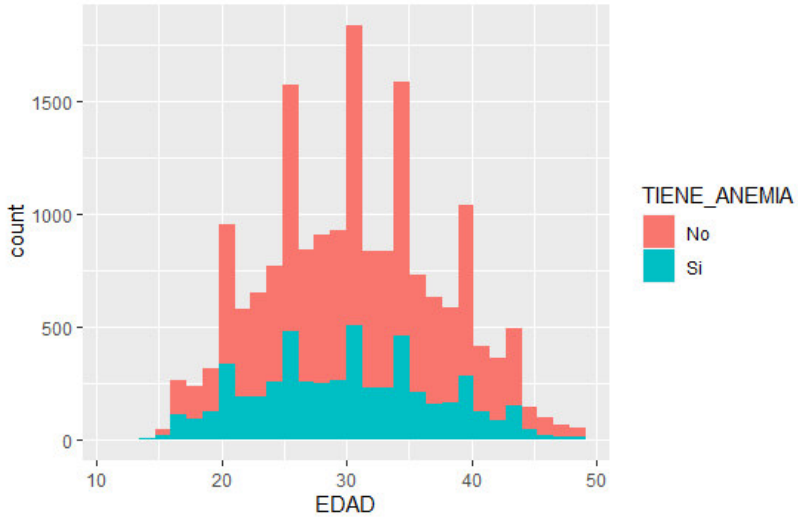
Variable dependiente	Frecuencia	% de casos
Tiene anemia (1)	5,278	29.6%
No tiene anemia (0)	12,536	70.4%
Total	17,814	100.0%

Nota. Elaboración propia a partir del conjunto de datos.

Aquí se puede observar un fenómeno conocido como el desbalanceo de datos y esto podría afectar el desempeño de los modelos. Para solucionarlo, es necesario aplicar técnicas de balanceo de datos. Previamente, continuaremos con el análisis descriptivo de las demás variables.

Figura 22

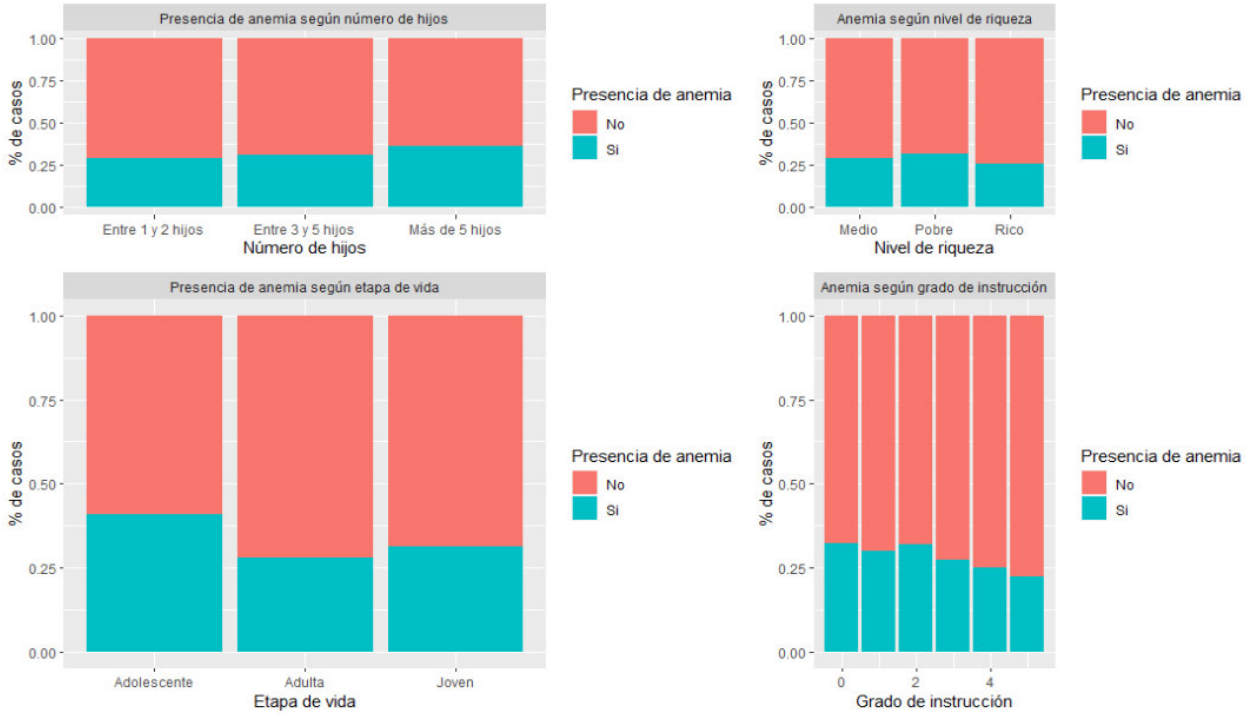
Histograma de edades según presencia de anemia



Nota. Elaboración propia a partir del conjunto de datos.

Figura 23

Gráfico de barras para variables categóricas



Nota. Elaboración propia a partir del conjunto de datos.

En el gráfico anterior se analiza la presencia de anemia versus las variables número de hijos, nivel de riqueza, etapa de vida y grado de instrucción.

En primera instancia, se observa la presencia de anemia por etapa de vida de la gestante. Aquí se tiene que, se presentan más casos de anemia en las mujeres más jóvenes (adolescentes), en el caso de las mujeres jóvenes se observan menos casos de anemia en comparación con las adolescentes y en las mujeres adultas estos casos se van reduciendo.

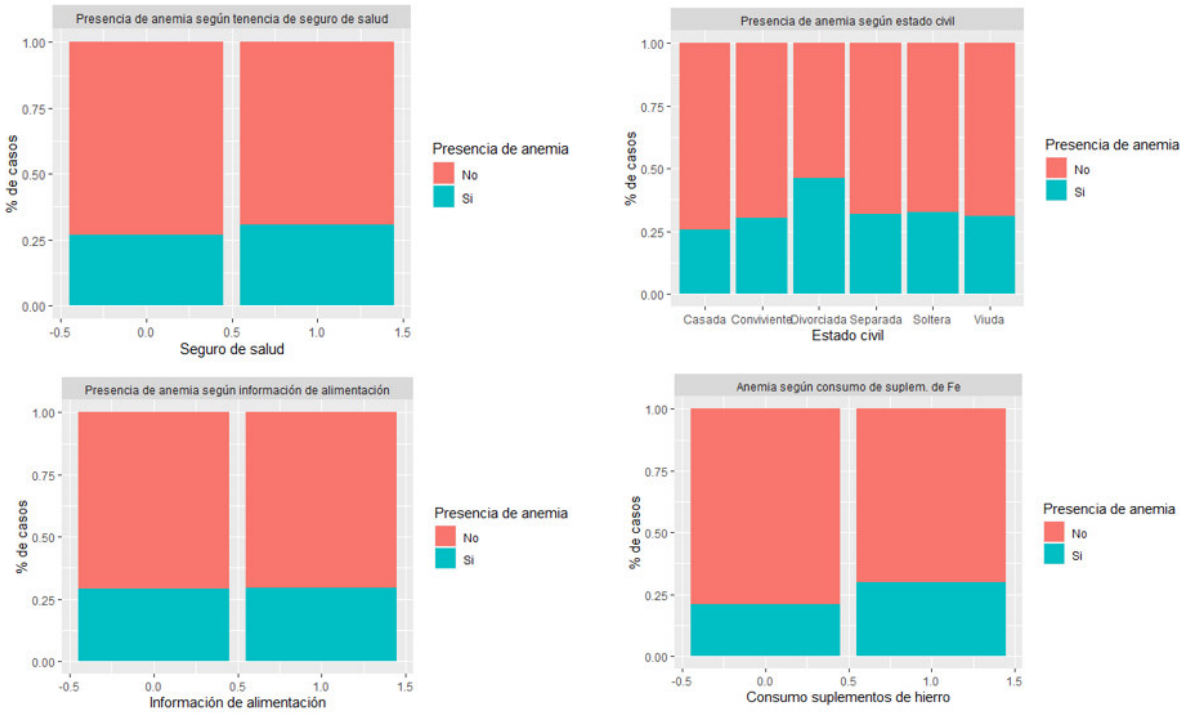
Por otro lado, según el nivel de riqueza, se observan más casos de anemia en las mujeres pobres.

De igual forma, se analiza la variable grado de instrucción. Se puede apreciar que las mujeres que tienen menor grado de instrucción son más propensas a presentar anemia en el embarazo, y los casos van disminuyendo en mujeres que tienen un mayor grado de instrucción.

Según el número de hijos, se puede apreciar que, se presentan más casos de anemia a medida que el número de hijos de estas va incrementando. Las mujeres que tienen más de 5 hijos son más propensas a tener anemia.

Figura 24

Gráfico de barras para variables categóricas



Nota. Elaboración propia a partir del conjunto de datos.

Se analiza también la presencia de anemia para variables como tenencia de seguro de salud, estado civil, información de alimentación en el embarazo y consumo de suplementos de

hierro. Sin embargo, luego del análisis de estas variables, no se observa alguna relación con la presencia de anemia.

Finalmente, se optará por elegir para el modelamiento las siguientes variables: número de hijos, grado de instrucción, etapa de vida y nivel de riqueza.

Antes de iniciar con la etapa de modelamiento, se define el porcentaje de datos que se empleará para el entrenamiento y validación de los modelos.

Tabla 6

Frecuencia y porcentaje de casos de la muestra de train y test

Muestra	Tiene anemia	No tiene anemia	Casos	Porcentaje
Entrenamiento	3,945	9,415	13,360	75%
Validación	1,333	3,121	4,454	25%
Total	5,278	12,536	17,814	100%

Nota. Elaboración propia a partir del conjunto de datos.

5.2 Técnica SMOTE para datos desbalanceados

Otro aspecto importante a tener en cuenta en la presente investigación es la distribución de las categorías de la variable dependiente, el cual representa el 29.6% para la clase de interés, lo que evidencia el desequilibro de datos.

Esto podría traer serias consecuencias en el modelado de los datos, ya que podrían producirse resultados que conlleven a conclusiones erróneas. Para evitar esto, emplearemos la técnica SMOTE.

Para esto, se analizará cuál de las técnicas será aplicada. En primer lugar, tenemos la técnica Undersampling, que consiste en equiparar la cantidad de categorías de la clase con menor

frecuencia con la clase de mayor frecuencia. De aplicarse esta técnica, tendríamos 3,945 observaciones de mujeres con anemia y 3,945 observaciones de mujeres sin anemia.

Por otro lado, se tiene la técnica Oversampling, cuya finalidad es equilibrar la clase minoritaria creando observaciones ficticias de esta para que contenga la misma cantidad de observaciones que la clase mayoritaria. En este caso, se tendrían 9,415 observaciones para los casos con anemia y 9,415 observaciones para los casos sin anemia.

5.3 Ajuste y entrenamiento de los modelos

Luego de aplicar la técnica anterior, se procedió a aplicar los modelos.

Para la regresión logística, se observa la matriz de confusión (figura 25). Se observa que 2061 observaciones fueron clasificadas correctamente y 1692 fueron clasificados erróneamente. Además, se tiene un AUC de 0.578.

Figura 25

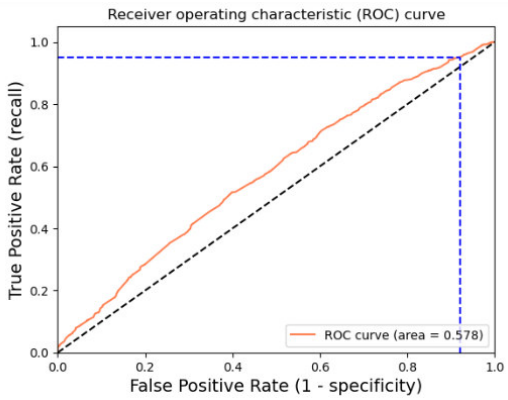
Matriz de confusión bajo el modelo de regresión logística

		Real	
		Anemia	No Anemia
Predicción	Anemia	1084	922
	No Anemia	770	977

Nota. Elaboración propia a partir del conjunto de datos.

Figura 26

Curva ROC para el modelo de regresión logística



Nota. Elaboración propia a partir del conjunto de datos.

Por otro lado, se observan los resultados de las redes neuronales. En la matriz de confusión (figura 27) se tienen 2631 observaciones clasificadas correctamente y 1823 valores clasificados de forma errónea. Además, se tiene un AUC de 0.57.

Figura 27

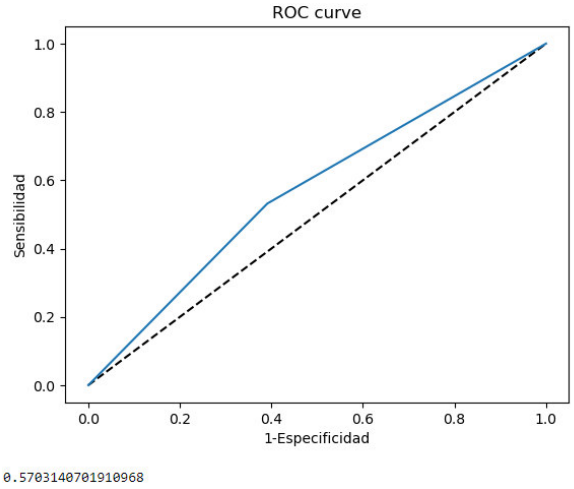
Matriz de confusión para el modelo de redes neuronales

		Real	
		Anemia	No Anemia
Predicción	Anemia	1850	552
	No Anemia	1271	781

Nota. Elaboración propia a partir del conjunto de datos.

Figura 28

Curva ROC para el modelo de redes neuronales



Nota. Elaboración propia a partir del conjunto de datos.

5.4 Evaluación y comparación de los modelos

Luego del ajuste de los modelos, es necesario observar otras medidas de desempeño de los modelos.

Tabla 7

Métricas de desempeño de los modelos

Métrica	Regresión Logística	Redes Neuronales
Precisión	55.00%	59.00%
Sensibilidad	58.47%	59.28%
Especificidad	51.45%	58.59%

Nota. Elaboración propia a partir del conjunto de datos.

Es importante mencionar que los datos utilizados en la presente investigación presentaban desequilibrio, ya que la clase de interés se presentó en un porcentaje menor en comparación a la otra clase. En la tabla 7 se observan los resultados de los modelos aplicados. Para el modelo de regresión logística se observa una precisión de 0.55 y para las redes neuronales se tiene 0.59. En cuanto a la sensibilidad, las redes neuronales muestran un mayor valor que la regresión logística (0.59), es decir, que las redes neuronales logran detectar mejor los casos que dan positivo a la anemia. Por otra parte, la especificidad para el modelo de redes neuronales fue mayor (0.58), lo que quiere decir que este modelo determina mejor los casos negativos para anemia.

VI. CONCLUSIONES

En la presente investigación se compararon dos técnicas para la predicción de anemia en gestantes y se tiene por propósito llamar la atención de las autoridades para que tomen medidas en el sector salud. Dentro de las conclusiones, se encuentran:

- Las variables que tienen influencia en la anemia son: número de hijos, grado de instrucción, nivel de riqueza y edad.
- En cuanto a la frontera de decisión, se tiene que, para el modelo de regresión logística la frontera de decisión es una función lineal, en cambio, para las redes neuronales es una función no lineal llamada función sigmoide.
- De los dos modelos empleados, se tiene que el de redes neuronales comete el menor error. Además, se encontró que dicho modelo presenta métricas más altas que el modelo de regresión logística.

VII. RECOMENDACIONES

- Se recomienda que las autoridades de salud tomen especial atención a las variables Número de hijos, nivel de riqueza, Edad y Grado de instrucción, ya que en esta investigación se determinó que estas tienen influencia en la presencia de la anemia.
- La frontera de decisión de las redes neuronales es lineal, mientras que la del modelo de regresión logística es no lineal, lo que permite que más datos puedan ser ajustados.
- Se sugiere el uso de las redes neuronales para la predicción de la anemia, ya que es el modelo que comete menos error.

VIII. BIBLIOGRAFÍA

Abanto, A. (2019). *Modelo logístico y de redes neuronales para pronóstico de anemia en niños menores de 3 años.* [Tesis de pregrado, Universidad Nacional de Trujillo]. http://dspace.unitru.edu.pe/bitstream/handle/UNITRU/16145/Geldres%20Del%20Risco.pdf;jsessionid=6A1217FF5B22FCCAFD9EC0E2AD29E8C2?sequence=1

Aggarwal, C. (2018). Neural Networks and Deep Learning. Springer.

Alvarez, D. (2021, 14 de enero). Metodología CRISP – DM. https://www.adictosaltrabajo.com/2021/01/14/metodologia-crisp-dm/

Ayala, F. y Ayala, D. (2019). Implicancias clínicas de la anemia durante la gestación. *Revista Peruana de Ginecología y Obstetricia, 65(4),* 487-488. http://www.scielo.org.pe/scielo.php?script=sci_arttext&pid=S2304-51322019000400012

Berzal, F. (2018). Redes Neuronales y Deep Learning. (pp. 185-203).

Chavez, R. (2019). *Comparación entre regresión logística y redes neuronales para predecir cáncer de piel en perros.* [Trabajo de investigación para obtener el grado de Bachiller, Universidad de Lima] https://repositorio.ulima.edu.pe/handle/20.500.12724/8401

Chawla, N. et al. (2002). *SMOTE: Synthetic Minority Over-Sampling Technique*. Journal of artificial intelligence research. (pp. 321- 357).

Calvo, D. (2018, 7 de diciembre). Funciones de activación – Redes Neuronales. https://www.diegocalvo.es/funcion-de-activacion-redes-neuronales/

Cóndor, E. y Correa, S. (2022). *Comparación entre regresión logística binaria y redes neuronales en la predicción de las causas de mortalidad materna en el Ecuador del año*

2020. [Proyecto de Investigación de Pregrado, Universidad Central del Ecuador]. http://www.dspace.uce.edu.ec/handle/25000/26706

Espitia, F. y Orozco, L. (2013). Anemia en el embarazo, un problema de salud que puede prevenirse. *Medicas UIS, 26*(3), 45-50.

Ficha Técnica ENDES 2022 [Internet]. INEI, Disponible en: https://proyectos.inei.gob.pe/iinei/srienaho/Descarga/FichaTecnica/786-Ficha.pdf

Foro Económico (2019, 17 de agosto). *Anemia: un problema de salud pública.* https://dev.focoeconomico.org/2019/08/17/anemia-un-problema-de-salud-publica/

Gavilán, I. (2020, 25 de mayo). *Catálogo de componentes de redes neuronales (III): funciones de pérdida.* https://ignaciogavilan.com/catalogo-de-componentes-de-redes-neuronales-iii-funciones-de-perdida/

Grau, R. et al. (2004). *Metodología de la investigación.* Ibagué: Universidad de Ibagué. Coruniversitaria.

Gurney, K. (1997). *An introduction to Neural Networks.* UCL Press Limited.

Haya, P. (2021). La metodología CRISP-DM en ciencia de datos. https://www.iic.uam.es/innovacion/metodologia-crisp-dm-ciencia-de-datos/

Hernández Sampieri, R. et al. (2014). *Metodología de la investigación.* (6ta ed.). McGraw Hill.

Hosmer, D. y Lemeshow, S. (2004). *Applied Logistic Regression.* (2nd ed.) A Wiley-Interscience Publication.

Instituto Nacional de Estadística e Informática [INEI], (2021). *Plan Nacional de Capacitación 2021.* https://www.inei.gob.pe/media/ENEI/servicios_participante/Plan_Capacitacion_2021_v2.pdf

Kleinbaum, D. y Klein, M. (2002). *Logistic Regression: A self-learning text.* (2nd ed.) Springer

Kroese, B. y Van der Smagt, P. (1996). *An introduction to Neural Networks.* (8va ed.). Amsterdam University of Applied Science. https://www.researchgate.net/publication/272832321_An_introduction_to_neural_networks

Llinás, H. (2021, 14 de agosto). El caso binario: Regresión Logística (curva ROC). https://rpubs.com/hllinas/R_Logit_Binario_RocCurve

Madsen, H. y Thyregod, P. (2011). *Introduction to General and Generalized Linear Models.* Chapman & Hall.

Matich, D. (2001). *Redes Neuronales: Conceptos Básicos y Aplicaciones*. https://www.frro.utn.edu.ar/repositorio/catedras/quimica/5_anio/orientadora1/monograias/matich-redesneuronales.pdf

Martinelli, J. (2022). *Clasificación de datos desbalanceados.* [Trabajo final de especialización, Universidad Nacional de La Plata]. http://sedici.unlp.edu.ar/handle/10915/147410

Mendoza, K. (2023). Método alternativo para la detección de la anemia a través de sus factores asociados en mujeres en edad reproductiva: una aplicación en redes neuronales artificiales. [Tesis de pregrado, Universidad Nacional Mayor de San Marcos] https://cybertesis.unmsm.edu.pe/handle/20.500.12672/19764

McCullagh, P. y Nelder, J. (1989). *Generalized Linear Models.* (2nd ed.)

Ministerio de Salud [MINSA], (2022). *Informe Gerencial SIEN HIS. Estado Nutricional de niños menores de 5 años y gestantes que acuden a establecimientos de Salud.*

https://cdn.www.gob.pe/uploads/document/file/4628853/Informe%20Gerencial%20SIEN-HIS%20Gestantes%202022.pdf

Ministerio de Salud [MINSA], (2021). *Informe Gerencial SIEN HIS. Estado Nutricional de niños menores de 5 años y gestantes que acuden a establecimientos de Salud.* https://web.ins.gob.pe/sites/default/files/Archivos/cenan/van/informes/2021/Inf%20Gerencial%20SIEN-HIS%202021.pdf

Ministerio de Salud [MINSA], (2020). *Informe Gerencial SIEN HIS. Estado Nutricional de niños menores de 5 años y gestantes que acuden a establecimientos de Salud.* https://cdn.www.gob.pe/uploads/document/file/4525940/Informe%20Gerencial%20SIEN-HIS%202020%20FINALyeQNp.pdf?v=1683560083

Ministerio de Salud [MINSA], (2019). *Informe Gerencial SIEN HIS. Estado Nutricional de gestantes que acuden a establecimientos de Salud.* *https://cdn.www.gob.pe/uploads/document/file/4525954/informe_gerencial_sien_his_2019LzE1B.pdf?v=1683560085*

Ministerio de Salud [MINSA], (2017). *Plan Nacional para la reducción y Control de la anemia Materno Infantil y la Desnutrición Crónica Infantil en el Perú: 2017-2021.* https://anemia.ins.gob.pe/sites/default/files/2017-08/RM_249-2017-MINSA.PDF

Navarro, Y., Oliva, J. *Modelo de Regresión Logística para identificar factores de riesgo y pronóstico de anemia en menores de cinco años en el Perú – 2018.* [Tesis de pregrado, Universidad Pedro Ruiz Gallo] https://repositorio.unprg.edu.pe/handle/20.500.12893/9193

Narkhede, S. (2018, 26 de Junio). *Understanding AUC – ROC Curve* https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

Niño, V. (2011). *Metodología de la Investigación.* (1ra ed.) Ediciones de la U.

Organización Mundial de la Salud [OMS]. (2023). *Anemia* https://www.who.int/es/health-topics/anaemia#tab=tab_1

Ortiz, Y. et al. (2019). Factores sociodemográficos y prenatales asociados a la anemia em gestantes peruanas. *Enfermería global, 56,* 273-281. https://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S1695-61412019000400010

Ortiz, Z. (2005). *Regresión logística y su aplicación en un caso de epidemiología.* [Monografía de pregrado, Universidad Nacional Mayor de San Marcos] https://cybertesis.unmsm.edu.pe/bitstream/handle/20.500.12672/11791/Ortiz_mz.pdf?sequence=1&isAllowed=y

Sáenz, N., Ballesteros, M. (2002). Redes Neuronales: Concepto, aplicaciones y utilidad en medicina. (pp. 119-120).

Satpahty, S. (2020, 6 de octubre). SMOTE for Imbalanced Classification with Python. https://www.analyticsvidhya.com/blog/2020/10/overcoming-class-imbalance-using-smote-techniques/

Singh, N. (2020, 2 de setiembre). *Métricas De Evaluación De Modelos En El Aprendizaje Automático.* https://www.datasource.ai/es/data-science-articles/metricas-de-evaluacion-de-modelos-en-el-aprendizaje-automatico

Sotaquirá, M. (2018, 6 de agosto). *La neurona artificial y la regresión logística.* https://www.codificandobits.com/blog/regresion-logistica-y-neurona-artificial/

Stanford Medicine Children's Health (2023). *Anemia en el embarazo.* https://www.stanfordchildrens.org/es/topic/default?id=anemiainpregnancy-90-P05537

Suresh, A. (2020, 17 de noviembre). *What is a confusión matrix?* https://medium.com/analytics-vidhya/what-is-a-confusion-matrix-d1c0f8feda5

Támara, A. et al. (2019). Regresión logística y redes neuronales como herramientas para realizar un modelo Scoring. *Revista Lasallista de Investigación, 16*(1) 187-200. http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S1794-44492019000100187

Tomek, I. (1976). Two modifications of cnn. IEEE Trans. Syst. Man Cybern., 6:769- 772.

Tovar, J. (2022, 20 de setiembre). *Aprendizaje automático – Manejo de conjuntos de datos desbalanceados – Python.* https://forum.huawei.com/enterprise/es/aprendizaje-autom%C3%A1tico-manejo-de-conjuntos-de-datos-desbalanceados-python/thread/667232355921838080-667212895009779712

Van Hulse, J. et al. (2007). *Experimental perspectives on learning from imbalanced data. ICML* (pp. 935-942, 2007). https://icml.cc/imls/conferences/2007/proceedings/papers/62.pdf

Vásquez, C. y Gonzales, G. (2019). Situación mundial de la anemia en gestantes. *Nutrición Hospitalaria, 36*(4), 996-997. http://dx.doi.org/10.20960/nh.02712

Velasco, L. (2021). Primer modelo Deep Learning con TensorFlow – Clasificación binaria red neuronal. [Video]. YouTube. https://www.youtube.com/watch?v=NERPPvoj3Go

Zeballos, M. (2022, 8 de julio). *Evaluación de rendimiento de un modelo de machine learning.* https://dev.to/awsmllatam/evaluacion-de-rendimiento-de-un-modelo-de-machine-learning-53pb

IX. ANEXOS

Anexo 1

Balanceo de los datos

```
from imblearn.over_sampling import SMOTE

sm = SMOTE(random_state=42)

X_sm, y_sm = sm.fit_resample(X_train,y_train)

print(f'''Cambio de X antes de SMOTE: {X.shape}
Cambio de X después SMOTE: {X_sm.shape}''')

print('\nBalance positivo y negativo de las clases (%):')
y_sm.value_counts(normalize=True)*100
```

```
Cambio de X antes de SMOTE: (13360, 21)
Cambio de X después SMOTE: (15012, 21)

Balance positivo y negativo de las clases (%):

0    50.0
1    50.0
Name: TIENE_ANEMIA, dtype: float64
```

Nota. Elaboración propia

Anexo 2

Código para el modelo de regresión logística

```
# Aplicando regresión logística
X_train, X_test, y_train, y_test = train_test_split(X_sm, y_sm, test_size=0.25, random_state=2)

# Verificando classification scores de la Regresión Logistica
logreg = LogisticRegression()
logreg.fit(X_train, y_train)
y_pred = logreg.predict(X_test)
y_pred_proba = logreg.predict_proba(X_test)[:, 1]
[fpr, tpr, thr] = roc_curve(y_test, y_pred_proba)
print('Train/Test split results:')
print(logreg.__class__.__name__+" accuracy is %2.3f" % accuracy_score(y_test, y_pred))
print(logreg.__class__.__name__+" log_loss is %2.3f" % log_loss(y_test, y_pred_proba))
print(logreg.__class__.__name__+" auc is %2.3f" % auc(fpr, tpr))



idx = np.min(np.where(tpr > 0.95)) # index of the first threshold for which the sensibility > 0.95

plt.figure()
plt.plot(fpr, tpr, color='coral', label='ROC curve (area = %0.3f)' % auc(fpr, tpr))
plt.plot([0, 1], [0, 1], 'k--')
plt.plot([0,fpr[idx]], [tpr[idx],tpr[idx]], 'k--', color='blue')
plt.plot([fpr[idx],fpr[idx]], [0,tpr[idx]], 'k--', color='blue')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate (1 - specificity)', fontsize=14)
plt.ylabel('True Positive Rate (recall)', fontsize=14)
plt.title('Receiver operating characteristic (ROC) curve')
plt.legend(loc="lower right")
plt.show()

print("Using a threshold of %.3f " % thr[idx] + "guarantees a sensitivity of %.3f " % tpr[idx] +
      "and a specificity of %.3f" % (1-fpr[idx]) +
      ", i.e. a false positive rate of %.2f%%." % (np.array(fpr[idx])*100))
```

Nota. Elaboración propia

Anexo 3

Matriz de confusión

Matriz de confusión con SMOTE

```
X_train, X_test, y_train, y_test = train_test_split(X_sm,y_sm,test_size=0.25,random_state=42)

logreg = LogisticRegression()
logreg.fit(X_train, y_train)
y_pred = logreg.predict(X_test)

print(f'Accuracy = {accuracy_score(y_test,y_pred):.2f}\nRecall = {recall_score(y_test,y_pred):.2f}\n')
cm = confusion_matrix(y_test,y_pred)
plt.figure(figsize=(8,6))
plt.title('Matriz de Confusión (Con SMOTE)',size=16)
sns.heatmap(cm,annot=True,cmap='Blues_r');
```

Nota. Elaboración propia

Anexo 4

Código para las Redes Neuronales

```
import tensorflow as tf
```

```
model = tf.keras.Sequential([
    tf.keras.layers.Input((None,21,)),
    tf.keras.layers.Dense(14, activation='relu'),
    tf.keras.layers.Dense(22, activation='leaky_relu'),
    tf.keras.layers.Dense(18, activation='leaky_relu'),
    tf.keras.layers.Dense(12, activation='leaky_relu'),
    tf.keras.layers.Dense(1, activation='sigmoid')
])
```

```
model.compile(optimizer='adam',
              loss=tf.keras.losses.binary_crossentropy,
              metrics=['accuracy'])
```

```
cols22 = ["ind_riqueza_0","ind_riqueza_1","ind_riqueza_2","rango_edad_0",
                      "rango_edad_1","rango_edad_2","num_hijos_0","num_hijos_1","num_hijos_2",
                      "GRADO_INSTRUCCIÓN_0","GRADO_INSTRUCCIÓN_1","GRADO_INSTRUCCIÓN_2",
                      "GRADO_INSTRUCCIÓN_3","GRADO_INSTRUCCIÓN_4","GRADO_INSTRUCCIÓN_5",
                      "SUPLEMENTOS_HIERRO","SEGURO_SALUD",
                      "REGION_NATURAL_1","REGION_NATURAL_2","REGION_NATURAL_3","REGION_NATURAL_4"
                      ]



# create X (features) and y (response)
X22 = train[cols22]
y22 = train['TIENE_ANEMIA']
```

```
columns_to_extract_1 = ["ind_riqueza_0","ind_riqueza_1","ind_riqueza_2","rango_edad_0",
                      "rango_edad_1","rango_edad_2","num_hijos_0","num_hijos_1","num_hijos_2",
                      "GRADO_INSTRUCCIÓN_0","GRADO_INSTRUCCIÓN_1","GRADO_INSTRUCCIÓN_2",
                      "GRADO_INSTRUCCIÓN_3","GRADO_INSTRUCCIÓN_4","GRADO_INSTRUCCIÓN_5",
                      "SUPLEMENTOS_HIERRO","SEGURO_SALUD",
                      "REGION_NATURAL_1","REGION_NATURAL_2","REGION_NATURAL_3","REGION_NATURAL_4"
                       ]
```

```
X_sm22 = X_sm[columns_to_extract_1]
```

```
history = model.fit(X_sm22, y_sm, validation_split=0.25,epochs=400,
                    validation_data=(X22, y22))
```

```
#Redes neuronales para el test
X3 = test[cols22]
y3 = test['TIENE_ANEMIA']
```

```
test_loss, test_acc = model.evaluate(X3,y3,verbose=2)
```

Nota. Elaboración propia

Anexo 5

Precisión del modelo y la función de pérdida

```
plt.plot(history.history['accuracy'])
plt.plot(history.history['val_accuracy'])
plt.title('model accuracy')
plt.ylabel('accuracy')
plt.xlabel('epoch')
plt.legend(['train','test'],loc='upper left')
plt.show()
#Summarize history for loss
plt.plot(history.history['loss'])
plt.plot(history.history['val_loss'])
plt.title('model loss')
plt.ylabel('loss')
plt.xlabel('epoch')
plt.legend(['train','test'],loc='upper left')
plt.show()
```

Nota. Elaboración propia

Anexo 6

Área bajo la curva de las Redes Neuronales

Área bajo la curva

```
#import classification_report

from sklearn.metrics import classification_report
print(classification_report(y22,y_pred))
from sklearn.metrics import roc_curve
y_pred_proba = out
fpr, tpr, thresholds = roc_curve(y22, y_pred_proba)
plt.plot([0,1],[0,1],'k--')
plt.plot(fpr,tpr, label='ANN')
plt.xlabel('1-Especificidad')
plt.ylabel('Sensibilidad')
plt.title('ROC curve')
plt.show()
#Area under ROC curve
from sklearn.metrics import roc_auc_score
roc_auc_score(y22,y_pred_proba)
```

Nota. Elaboración propia