FICHA TÉCNICA DE ANÁLITICA DE NEGOCIOS

1. Número y Nombre del Equipo

Grupo 6 - Los Galácticos del Norte

2. Miembros del Equipo

- Castillo Oliva, Johann
- Chavez Ferrel, Marvin Alberto
- Espinosa Luna, Bruno Adrián
- García Gutiérrez, Willy Francisco
- Marreros Urquiza, Julio Enrique
- Montañez Díaz, Bruno Hiroshi
- Varas Zurita, Piero Lenin

3. Nombre del caso

Modelo de Machine Learning para predecir riesgos de anemia en niños de La Libertad, 2024

ÍNDICE

COMPRENSIÓN DEL NEGOCIO

Realidad problemática

La anemia es un desafío de salud pública relevante en el contexto global, afectando de manera significativa a la población infantil. Según la Organización Mundial de la Salud (OMS), aproximadamente el 40% de los niños entre 6 y 59 meses padecen esta condición, siendo este grupo especialmente vulnerable (OMS, 2023). La persistencia y gravedad de la anemia en la infancia pueden tener repercusiones adversas en el neurodesarrollo y el rendimiento cognitivo, especialmente en entornos socioeconómicos desfavorecidos en países en desarrollo (Leung et al., 2024).

En Perú, casi un tercio de los niños menores de 6 a 59 meses sufren de anemia, siendo factores determinantes la desigualdad socioeconómica, el nivel educativo de las madres y la residencia en zonas rurales y amazónicas, donde el acceso limitado a una variedad adecuada de alimentos es común (Al-Kassab-Córdova et al., 2023). A pesar de los esfuerzos del país mediante programas sociales como Qali Warma, Cuna Más y Juntos, no se consiguen materializar avances significativos para reducir la prevalencia de la anemia (Vázques et al., 2024).

Datos del Instituto Nacional de Estadística e Informática (INEI) evidencian esta situación en la Encuesta Demográfica y de Salud Familiar (ENDES) del año 2023, donde el 43,1% de los niños de 6 a 35 meses en Perú se vieron afectados por la anemia. Esta prevalencia fue más pronunciada en áreas rurales (50,3%) en comparación con áreas urbanas (40,2%), y los departamentos más impactados fueron Puno (70,4%), Ucayali (59,4%) y Madre de Dios (58,3%).

De la misma manera, la situación es particularmente grave en la zona norte del país. Entre los años 2018 y 2022, en la región de La Libertad, la prevalencia de anemia en infantes de 6 a 35 meses experimentó una reducción de apenas 1.8% (ENDES 2018 - 2022). Este rango de variación es bajo en comparación con la región de Junín, que logró una reducción significativa del 14.5% en el mismo periodo. A pesar de los

esfuerzos desplegados por las autoridades sanitarias y diversos programas del gobierno regional, la situación en La Libertad no ha mostrado mejoras sustanciales.

Para abordar esta persistente problemática, es crucial desarrollar estrategias innovadoras que brinden una visión más completa de la situación y permitan implementar intervenciones efectivas y adaptadas a las necesidades específicas de cada comunidad.

Objetivos del negocio

- Reducir tiempo de diagnóstico de anemia infantil en La Libertad
- Determinar niveles de anemia para próximos años en La Libertad
- Mejorar alimentación a niños de La Libertad para evitar anemia

Metas de la minería de datos

- Crear un modelo para diagnosticar la presencia y severidad de anemia infantil en niños de La Libertad
- Crear un modelo para predecir los niveles de anemia para próximos años en La Libertad
- Crear modelo para categorizar a los niños con alimentos que deberían consumir para reducir el nivel de anemia en niños de La Libertad.

Datos a trabajar

- Atributos personales (edad, sexo, ubigeo,)
- Atributos de la sangre (nivel de hierro, hemoglobina)
- Severidad de anemia
- Programas sociales
- Centros de salud
- Nivel de anemia (leve, moderado, severo)

Objetivo general

- Desarrollar un modelo de Machine Learning para predecir riesgo de anemia en niños de La Libertad, 2024

ALGORITMO A APLICAR

Objetivo de Minería 1: Crear un modelo para diagnosticar la presencia y severidad de anemia infantil en niños de La Libertad

Para el diagnóstico de la presencia y severidad de anemia se emplea el modelo **Random Forest**. La elección de este algoritmo se fundamenta en los estudios de Abdul-Jabbar et al. (2023) y Tesfaye et al. (2024), los cuales demuestran que alcanza valores sobresalientes de precisión en estudios comparativos para el diagnóstico de anemia.

En el caso específico de Agramonte Mayhua et al. (2022), se aplicó el algoritmo de árbol de decisión para predecir si un niño tiene anemia y su nivel de severidad, utilizando los datos proporcionados. Las variables de entrada consideradas fueron: 'Diresa', 'Sexo', 'EdadMeses', 'Peso', 'Talla', 'Hemoglobina', 'HBC', 'ProvinciaREN' y 'DistritoREN', mientras que la variable de salida fue 'DxAnemia'. Los resultados de esta investigación mostraron una precisión global del modelo de 0.9933, lo que valida el algoritmo como una herramienta de alta precisión para el diagnóstico de anemia.

Objetivo de Minería 2: Crear un modelo para predecir el nivel de prevalencia de anemia para próximos años en La Libertad

Para pronosticar la prevalencia de anemia en los próximos años, se utilizará Prophet. Este modelo se destaca por su capacidad para manejar datos de series temporales con cambios abruptos y patrones aleatorios, incluso con observaciones limitadas. La elección de Prophet se basa en la investigación de Satrio et al. (2021), donde se demostró que Prophet supera a otros modelos como ARIMA en la precisión de pronósticos de casos de enfermedades como el COVID-19.

Objetivo de Minería 3: Crear modelo para categorizar a los niños para determinar el nivel de posibilidad de anemia en base a la dieta que consumen los niños de La Libertad.

De acuerdo con Bendezu e Ysla (2020), la aplicación del algoritmo de árboles de decisión permitirá la clasificación de cuáles grupos de niños van a consumir ciertos tipos de alimentos para reducir sus niveles de anemia.

Los mismos autores usaron como variables de entrada: Edad, género, nivel económico, grado de anemia, consumo de proteínas, consumo de carbohidratos, consumo de frutas y consumo de grasas.

COMPRENSIÓN DE LOS DATOS

Objetivo de Minería 1: Crear un modelo para diagnosticar la presencia y severidad de anemia infantil en niños de La Libertad

El conjunto de datos utilizado en este estudio se obtuvo de la Plataforma Nacional de Datos Abiertos del Estado Peruano. Estos datos fueron recopilados por el Sistema de Información del Estado Nutricional de Niños y Gestantes (SIEN), implementado por el Instituto Nacional de Salud (INS). El SIEN reporta información sobre el estado nutricional de niños menores de cinco años y mujeres gestantes que acuden a establecimientos de salud.

El análisis se enfocará en La Libertad entre los años 2019 y 2023, el set de datos proporciona información sobre las siguientes características:

Tabla 1. *Diccionario de conjunto de variables finales de los datos recopilados por el SIEN*

Descripción	**Denominación**
Sexo del niño (Masculino y Femenino)	Sexo
Edad del niño en meses	EdadMeses
Peso del niño	Peso
Talla del niño	Talla
Nivel de hemoglobina	Hemoglobina
Control de crecimiento y desarrollo (No y Si)	Cred
Consumo de suplementos en niños (No y Si)	Suplementación
Provincia donde vive el niño	Provincia REN
Distrito al que pertenece el niño	DistritoREN
Nivel de anemia que posee un niño (Salida u Objetivo a predecir)	**Dx_Anemia**

Fuente: https://polibits.cidetec.ipn.mx/ojs/index.php/CyS/article/view/4315/3590

A partir de este conjunto de datos, se realizó un análisis estadístico básico de las variables cuantificables como el peso, la talla y el nivel de hemoglobina para evaluar la calidad de los datos. La Tabla 1 presenta

los resultados del cálculo de la media, la desviación estándar, los cuartiles, así como los valores máximos y mínimos. Los resultados revelaron un problema de calidad de datos en las variables de peso y talla, donde la media es de 12.78 kg y 82.54 cm, respectivamente, pero los valores máximos alcanzan los 999 kg y 999 cm. Esto se debe a una falta de normalización en las unidades de medida de los registros, lo que resulta en una variación inconsistente en las unidades utilizadas.Según el artículo de Marcos Valdez et al. (2023), los valores de peso y talla fuera del rango de 212 gramos a 50 kg y de 24 cm a 170 cm en niños menores de 5 años son considerados anómalos y deben ser eliminados del conjunto de datos para asegurar la calidad de los mismos.

Tabla #. *Análisis estadístico inicial del conjunto de datos recopilados por el SIEN 2019 - 2023*

	EdadMeses	Peso	Talla	Hemoglobina	Cred	Suplementación
count	228037.00	129305.00	129119.00	228037.00	228037.00	228037.00
mean	25.65	12.78	82.54	12.04	0.60	0.55
std	14.98	15.59	24.10	1.25	0.49	0.50
min	6.00	0.00	0.00	4.20	0.00	0.00
25%	13.00	9.50	73.50	11.20	0.00	0.00
50%	24.00	11.50	82.00	11.90	1.00	1.00
75%	36.00	14.00	92.00	12.90	1.00	1.00
max	60.00	999.00	999.00	18.50	1.00	1.00

Objetivo de Minería 2: Crear un modelo para predecir el nivel de prevalencia de anemia para próximos años en La Libertad

Se utilizaron los mismos datos recopilados por el Sistema de Información del Estado Nutricional de Niños y Gestantes (SIEN), disponibles en la Plataforma Nacional de Datos Abiertos del Estado Peruano, que se emplearon en el primer objetivo. Se analizaron los datos desde el año 2019 hasta el año 2023 para

comprender mejor el comportamiento y la cantidad de casos de anemia en niños en la región de La Libertad.

Se llevó a cabo un análisis de la frecuencia diaria de diagnósticos en niños, basándose en los datos recopilados.

Objetivo de Minería 3: Crear modelo para categorizar a los niños para determinar el nivel de posibilidad de anemia en base a la dieta que consumen los niños de La Libertad.

Se usaron los datos recopilados por la encuesta VIANEV del 2021 sobre hábitos y consumo de alimentos saludables de niños de 5 a 11 años, realizada por el Instituto Nacional de Salud. La base de datos se encuentra en la Plataforma Nacional de Datos Abiertos del Estado Peruano.

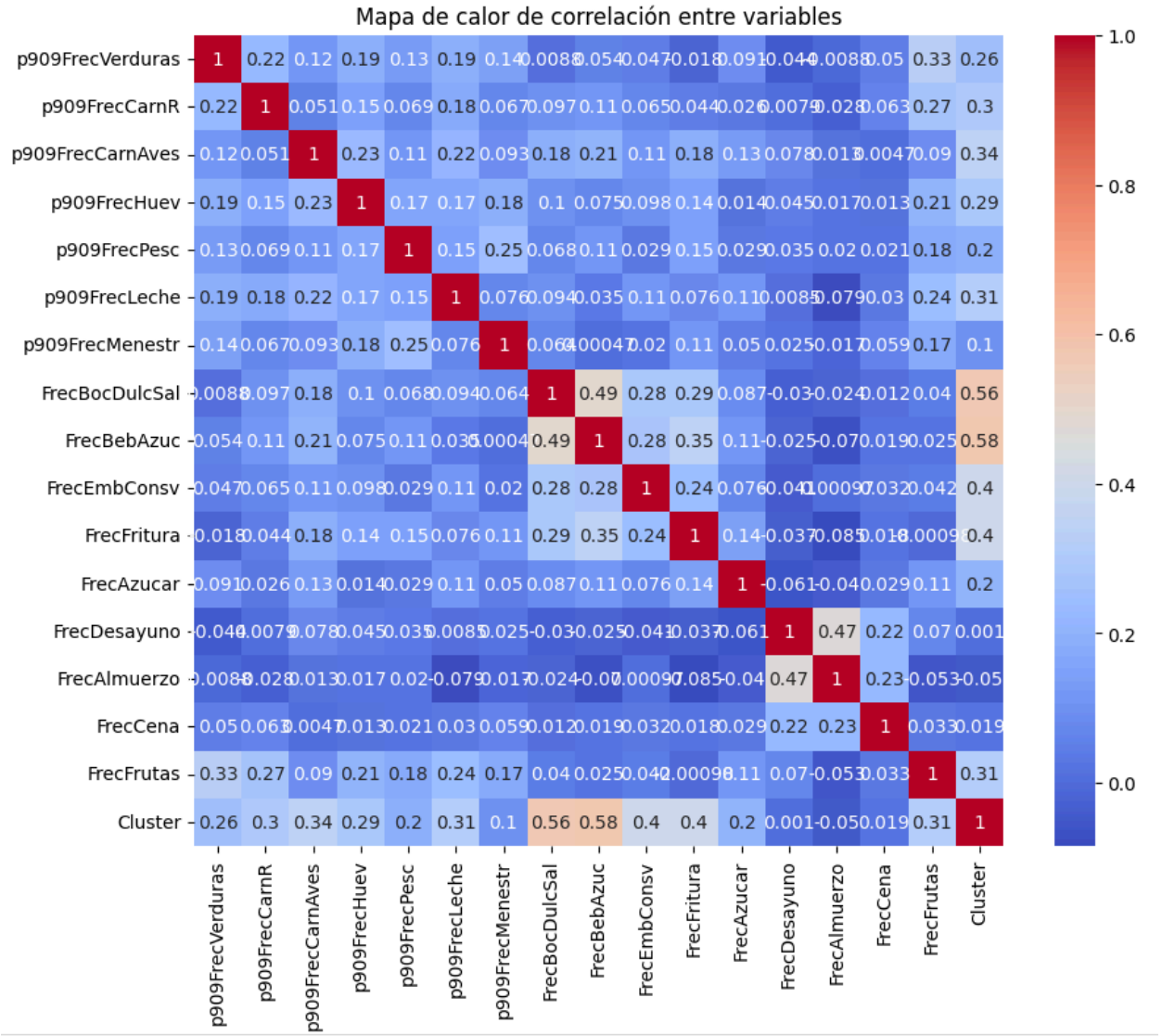
El área de recolección de datos fue Perú, incluyendo zonas de Lima metropolitana, resto urbano y rural. De esta base de datos, se seleccionan las siguientes variables:

Tabla 2. *Diccionario de conjunto de variables finales de los datos recopilados por el INS*

Descripción	**Denominación**
Sexo	Sexo
Edad	Edad
Frecuencia de consumo de verduras	p909FrecVerduras
Frecuencia de consumo de carnes rojas o vísceras	p909FrecCarnR
Frecuencia de consumo de carnes de aves	p909FrecAves
Frecuencia de consumo de huevos	p909FrecHuev
Frecuencia de consumo de pescado	p909FrecPesc
Frecuencia de consumo de leche	p909FrecLeche
Frecuencia de consumo de menestra	p909FrecMenestr
Frecuencia de consumo de bebidas azucaradas	FrecBebAzuc
Frecuencia de consumo de embutidos	FrecEmbConsv
Frecuencia de consumo de frituras	FrecFritura

Frecuencia de consumo de azúcar	FrecAzucar
Frecuencia de consumo de desayuno	FrecDesayuno
Frecuencia de consumo de almuerzo	FrecAlmuerzo
Frecuencia de consumo de cena	FrecCena
Frecuencia de consumo de frutas	FrecFrutas
Nivel de posibilidad de anemia que posee un niño (Salida u Objetivo a predecir)	**Alta, Baja, Media**

Adicionalmente, respecto a los datos de frecuencia de alimentación, es necesario conocer la correlación entre las variables para tener entradas de calidad. Una correlación positiva baja ($0 < x < 0.5$) o media ($0.5 < x < 0.7$) garantiza valores independientes y no redundantes para el análisis, mientras que una correlación fuerte o perfecta indica redundancia, y por lo tanto, variables que deberían ser descartadas al no ser relevantes para la clusterización.



En la figura, se observa una correlación positiva baja en todas las variables, por lo que se considerarán para la clusterización.

MODELADO

Objetivo de Minería 1: Crear un modelo para diagnosticar la presencia y severidad de anemia infantil en niños de La Libertad

Para el entrenamiento del modelo se utilizaron los datos de Anemia en Niños en La Libertad en el año 2023

Dataset

Tomado del Sistema de información del Estado Nutricional de niños y gestantes Perú - INS/CENAN (Instituto Nacional de Salud - Centro Nacional de Alimentación y Nutrición)

Enlace de descarga directo:

https://drive.google.com/file/d/1N5PquvQMd1rW3AdWdcHarFVnZixkgbCM/view?usp=sharing

Enlace a la fuente del dataset:

https://www.datosabiertos.gob.pe/dataset/sien-sistema-de-informaci%C3%B3n-del-estado-nutricional-de-ni%C3%B1os-y-gestantes-per%C3%BA-inscenan

Código

https://colab.research.google.com/drive/1gndVmAhLRsQNaS4oxibKZsRwtCJGHP1X?usp=sharing

Objetivo de Minería 2: Crear un modelo para predecir el nivel de prevalencia de anemia para próximos años en La Libertad

Para el entrenamiento del modelo se utilizaron los datos de Anemia en Niños en La Libertad desde el año 2019 hasta el 2023.

Dataset

Tomado del Sistema de información del Estado Nutricional de niños y gestantes Perú - INS/CENAN (Instituto Nacional de Salud - Centro Nacional de Alimentación y Nutrición)

Enlace de descarga directo:

https://drive.google.com/drive/folders/1U-HRgdqQF7skn4UeL3VcjvZio7JdfAX4?usp=sharing

Enlace a la fuente del dataset:

https://www.datosabiertos.gob.pe/dataset/sien-sistema-de-informaci%C3%B3n-del-estado-nutricional-de-ni%C3%B1os-y-gestantes-per%C3%BA-inscenan

Código

https://colab.research.google.com/drive/1eAt7rhW-mlCOVmOSzlqA3cKaqoNWdsT1?usp=sharing

Objetivo de Minería 3: Crear modelo para categorizar a los niños para determinar el nivel de posibilidad de anemia en base a la dieta que consumen los niños de La Libertad.

Dataset

Tomado del Encuesta de Vigilancia Alimentaria Nutricional por Etapas de Vida - INS/CENAN (Instituto Nacional de Salud - Centro Nacional de Alimentación y Nutrición)

Enlace de descarga directo:

https://drive.google.com/file/d/1Y6NzSEu6lyo1xoGQBSDfl_knGi6Tggc1/view?usp=sharing

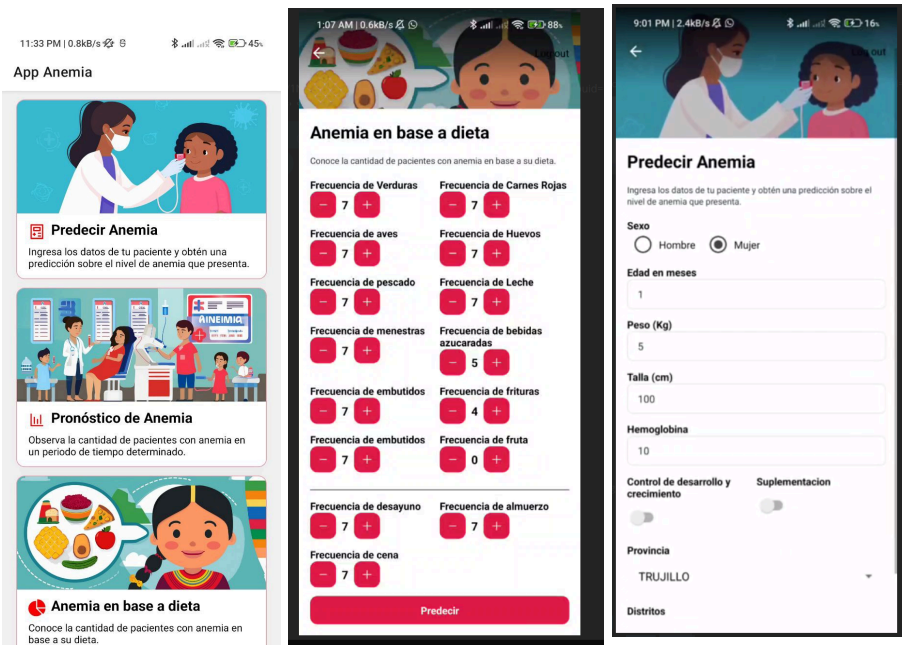
Enlace a la fuente del dataset:

https://www.datosabiertos.gob.pe/dataset/encuesta-vianev-2021-h%C3%A1bitos-y-consumo-de-alimentos-saludables-de-ni%C3%B1os-de-5-11-a%C3%B1os-ins

Código

https://drive.google.com/file/d/119a34SmNisxL1sbR24zwCnmD56F6RS36/view?usp=sharing

APLICACIÓN MÓVIL ANEMIA



Desarrollado con React Native

Enlace de descarga:

https://drive.google.com/drive/folders/1Q18ih-jLeUNHA9L3b7v2p9VrUbr2JuR1?usp=sharing

REFERENCIAS BIBLIOGRÁFICAS

Abdul-Jabbar, S. S., Farhan , A. K., & Luchinin, A. S. (2023). A Comparative Study of Anemia Classification Algorithms for International and Newly CBC Datasets. International Journal of Online and Biomedical Engineering (iJOE), 19(06), pp. 141–157. https://doi.org/10.3991/ijoe.v19i06.38157

Agramonte Mayhua, I., Chaco Huamani, A., Valdiviezo Tovar, A., & Ramos Challa, M. (2022). Aplicación de los árboles de decisión en el diagnóstico de Anemia en niños de la ciudad de Arequipa. *Innovación Y Software*, *3*(2), 26-39. https://doi.org/10.48168/innosoft.s9.a69

Bendezu, C., Ysla, R. (2020). App de recomendaciones alimentarias para reducir la mala alimentación en casos de anemia en niños del colegio "Apóstol de Punchauca". https://repositorio.usmp.edu.pe/handle/20.500.12727/6824

Satrio, C. B. A., Darmawan, W., Nadia, B. U., & Hanafiah, N. (2021). Time series analysis and forecasting of coronavirus disease in Indonesia using ARIMA model and PROPHET. *Procedia Computer Science*, *179*, 524-532. https://doi.org/10.1016/j.procs.2021.01.036

Tesfaye, S. H., Seboka, B. T., & Sisay, D. (2024). Application of machine learning methods for predicting childhood anemia: Analysis of Ethiopian Demographic Health Survey of 2016. PLOS ONE, 19(4), 1–14. https://doi.org/10.1371/journal.pone.0300172

ANEXOS

Juicio de expertos 1

Validación de variables por juicio de expertos

1. **Datos generales**
   - Fecha: 17/07/24
   - Validador: Dr. José Manual Burgos Zavaleta
   - Cargo y lugar de trabajo: Pediatría – Hospital Belén
   - Años de experiencia: 27 como pediatra

Título de investigación: Modelo de Machine Learning para predecir riesgos de anemia en niños de La Libertad, 2024

Objetivo general: Desarrollar un modelo de Machine Learning para predecir riesgo de anemia en niños de La Libertad, 2024

Asunto a validar: Variables de entrada para modelo de Machine Learning

Objetivo de validación: Emplear correctamente variables que permitirán el entrenamiento y predicción adecuada sobre diagnóstico de anemia, casos de anemia y categorización de niños según su alimentación

Criterios de evaluación:

- *Relevancia:* La variable es de importancia para el objetivo del estudio. Valores del 1 al 5 (1: Nada relevante, 5: muy relevante)
- *Claridad:* La variable está claramente definida. Valores del 1 al 5 (1: No hay claridad, 5: muy claro)
- *Pertinencia:* La variable es adecuada para predecir diagnóstico de anemia, casos de anemia o categorización de niños según su alimentación. Valores del 1 al 5 (1: Nada pertinente, 5: muy pertinente)
- *Redundancia:* La variable puede ser eliminada por estar contenida o ser similar a otra variable. Valores sí o no.

VARIABLES PARA OBJETIVOS 1 Y 2

Variable	**Definición operativa**	**Relevancia** (1-5)	**Claridad** (1-5)	**Pertinencia** (1-5)	**Redundancia** (Sí o no)	**Observaciones**
Sexo	Sexo del niño	1	5	3	No	
Edad	Edad en meses del niño	4	5	5	No	
Peso	Peso en kilogramos del niño	4	5	5	No	
Talla	Talla en centímetros del niño	3	5	5	No	
Nivel de hemoglobina	Gramos de hemoglobina por decilitro	5	5	5	No	
Control de crecimiento y desarrollo	Asistencia o no al control de crecimiento y desarrollo	3	5	5	No	
Consumo de suplementos	Consumo o no de suplementos de hierro	4	5	5	No	
Provincia	Provincia de La Libertad del niño	2	5	3	No	
Distrito	Distrito de La Libertad del niño	2	5	3	No	
Nivel de anemia	Grado de anemia	5	5	5	No	

VARIABLES PARA OBJETIVO 3

Consumo de verduras	Frecuencia diaria por semana de consumo de verduras	3	5	5	No	
Consumo de carnes rojas	Frecuencia diaria por semana de consumo de carnes rojas	4	5	5	No	
Consumo de pescado	Frecuencia diaria por semana de consumo de pescado	5	5	3	No	
Consumo de menestra	Frecuencia diaria por semana de consumo de menestra	3	5	3	No	

Consumo de bebidas azucaradas	Frecuencia diaria por semana de consumo de bebidas azucaradas	2	5	3	No	
Consumo de frituras	Frecuencia diaria por semana de consumo de frituras	2	5	1	Sí	Redunda con carnes rojas
Consumo de azúcar	Frecuencia diaria por semana de consumo de azúcar	3	5	1	Sí	Redunda con bebidas azucaradas y frutas
Consumo de frutas	Frecuencia diaria por semana de consumo de frutas	3	5	5	No	
Consumo de desayuno	Frecuencia diaria por semana de consumo de desayuno	4	5	5	No	

Juicio de expertos 2

Validación de variables por juicio de expertos

1. **Datos generales**
   - Fecha: 24/07/24
   - Validador: Debora Julissa Medina Palma - CMP: 71724
   - Cargo y lugar de trabajo: MÉDICO –Hospital Especialidades Basicas La Noria
   - Años de experiencia: 8 años

Título de investigación: Modelo de Machine Learning para predecir riesgos de anemia en niños de La Libertad, 2024

Objetivo general: Desarrollar un modelo de Machine Learning para predecir riesgo de anemia en niños de La Libertad, 2024

Asunto a validar: Variables de entrada para modelo de Machine Learning

Objetivo de validación: Emplear correctamente variables que permitirán el entrenamiento y predicción adecuada sobre diagnóstico de anemia, casos de anemia y categorización de niños según su alimentación

Criterios de evaluación:

- *Relevancia:* La variable es de importancia para el objetivo del estudio. Valores del 1 al 5 (1: Nada relevante, 5: muy relevante)
- *Claridad:* La variable está claramente definida. Valores del 1 al 5 (1: No hay claridad, 5: muy claro)
- *Pertinencia:* La variable es adecuada para predecir diagnóstico de anemia, casos de anemia o categorización de niños según su alimentación. Valores del 1 al 5 (1: Nada pertinente, 5: muy pertinente)

- *Redundancia:* La variable puede ser eliminada por estar contenida o ser similar a otra variable. Valores sí o no.

VARIABLES PARA OBJETIVOS 1 Y 2

Variable	**Definición operativa**	**Relevancia** (1-5)	**Claridad** (1-5)	**Pertinencia** (1-5)	**Redundancia** (Sí o no)	**Observaciones**
Sexo	Sexo del niño	5	5	5	No	
Edad	Edad en meses del niño	5	5	5	No	
Peso	Peso en kilogramos del niño	5	5	5	No	
Talla	Talla en centímetros del niño	5	5	5	No	
Nivel de hemoglobina	Gramos de hemoglobina por decilitro	5	5	5	No	
Control de crecimiento y desarrollo	Asistencia o no al control de crecimiento y desarrollo	5	5	5	No	
Consumo de suplementos	Consumo o no de suplementos de hierro	3	2	3	No	
Provincia	Provincia de La Libertad del niño	5	4	5	No	
Distrito	Distrito de La Libertad del niño	5	4	5	No	
Nivel de anemia	Grado de anemia	5	4	5	No	

VARIABLES PARA OBJETIVO 3

Consumo de verduras	Frecuencia diaria por semana de consumo de verduras	5	4	5	No	
Consumo de carnes rojas	Frecuencia diaria por semana de consumo de carnes rojas	4	4	4	No	
Consumo de pescado	Frecuencia diaria por semana de consumo de pescado	5	5	5	No	
Consumo de menestra	Frecuencia diaria por semana de	5	5	5	No	

	consumo de menestra					
Consumo de bebidas azucaradas	Frecuencia diaria por semana de consumo de bebidas azucaradas	5	4	5	No	
Consumo de frituras	Frecuencia diaria por semana de consumo de frituras	5	4	5	No	
Consumo de azúcar	Frecuencia diaria por semana de consumo de azúcar	5	5	5	Sí	Redunda con bebidas azucaradas
Consumo de frutas	Frecuencia diaria por semana de consumo de frutas	5	5	5	No	
Consumo de desayuno	Frecuencia diaria por semana de consumo de desayuno	5	5	5	No	