ISSN 2007-9737

Aprendizaje automático para predicción de anemia en niños menores de 5 años mediante el análisis de su estado de nutrición usando minería de datos

Alexander J. Marcos Valdez, Eduardo G. Navarro Ortiz,
Rodrigo E. Quinteros Peralta, Juan J. Tirado Julca,
David F. Valentin Ricaldi, Hugo D. Calderon-Vilca

Universidad Nacional Mayor de San Marcos,
Perú

{alexander.marcos, eduardo.navarro2, rodrigo.quinteros,
juan.tirado, david.valentin, hcalderonv}@unmsm.edu.pe

Resumen. Uno de los principales problemas de salud pública es la desnutrición infantil, ya que afecta negativamente al individuo a lo largo de su vida, limita el desarrollo de la sociedad y dificulta la erradicación de la pobreza. El objetivo de esta investigación es aplicar técnicas de minería de datos para el preprocesamiento, limpieza, reducción y transformación a un lago de datos que ha permitido analizar la anemia en niños menores de 5 años, asimismo, se ha aplicado algoritmos de Machine Learning para obtener el mejor modelo que permita predecir la anemia en niños menores de 5 años. El conjunto de datos fue extraído de la plataforma datos abiertos del gobierno del Perú que corresponde a Lima sur, Lima Norte, Lima Este, Lima Centro y Lima rural que se juntaron en total 138369 instancias y 36 variables, de la cuales 30 son categóricas y 6 numéricas, siendo un conjunto de datos desequilibrado. Para poder obtener las mejores variables predictoras se utilizó los filtros Anova F-test y Chi Cuadrado y se logró reducir a 10 variables, también se realizó casos sin considerar uno de los filtros y ambos filtros. Para buscar el mejor modelo de predicción se ha probado los algorítmicos: árbol de decisión, regresión logística, K vecinos más cercanos, bosque aleatorio y Naive bayes. Como resultado mostramos que el mejor algoritmo que permite predecir la anemia en niños menores de 5 años es el Naive Bayes con mayor recall de 74 %, precisión de 43% y exactitud del 70 %.

Palabras clave: Anemia, modelo predictivo, desnutrición, niños, minería de dato.

Machine Learning for the Prediction of Anemia in Children Under 5 Years of Age by Analyzing their Nutritional Status Using Data Mining

Abstract. One of the main public health problems is child malnutrition, since it negatively affects the individual throughout his life, limits the development of society and makes it difficult to eradicate poverty. The first objective of this research is to apply data mining techniques for preprocessing, cleaning, reduction and transformation to a data lake that has allowed analyzing anemia in children under 5 years of age, the second objective is to apply Machine Learning algorithms to obtain the best model to predict anemia in children under 5 years of age. The data set was extracted from the open data platform of the government of Peru that corresponds to South Lima, North Lima, East Lima, Central Lima and rural Lima, which collected a total of 138,369 instances and 36 variables of which 30 are categorical and 6 numeric, being an unbalanced data set. In order to obtain the best predictor variables, the Anova F-test and Chi Square filters were used, and it was possible to reduce them to 10 variables, cases were also carried out without considering one of the filters and both filters.To find the best prediction model, the algorithms have been tested: decision tree, logistic regression, K nearest neighbors, random forest and naive bayes. As a result, we show that the best algorithm to predict anemia in children under 5 years of age is the Naive Bayes algorithm with the highest recall of 74%, precision of 43% and accuracy of 70%.

Keywords. Anemia, predictive model, malnutrition, children, data mining.

1. Introducción

La desnutrición infantil es uno de los principales problemas de salud pública en el Perú, que afecta al 19,5% de niños menores de cinco años [1]. Este problema se acentúa en la población de más temprana edad y con mayor grado de exclusión, tal es el caso de la población rural, de menor nivel educativo y de menores ingresos económicos.

El INEI precisó que, en el área urbana, la desnutrición crónica afectó al 7,2%, y en el área rural al 24,7% de los niños menores de cinco años según el documento "Perú: Indicadores de Resultados de los Programas Presupuestales 2020″ [2].

En Lima Metropolitana, la relación entre nivel socioeconómico y desnutrición crónica es más estrecha. Los datos analizados muestran que Lima Metropolitana enfrenta un problema de desnutrición crónica muy importante en términos cuantitativos. Pero, Lima no ha estado en el radar de los programas sociales destinados a luchar contra esta condición debido a las bajas tasas de pobreza, que ocultaron la presencia de miles de niños con desnutrición crónica [3].

Un indicador de desnutrición y mala salud es la anemia. La Organización Mundial de la Salud indicó que la anemia repercute en varios problemas de nutrición como retraso del crecimiento, emaciación, peso bajo al nacer o sobrepeso y obesidad en la niñez debido a la falta de energía para hacer ejercicio [4].

A continuación, se describe cómo este problema ha sido afrontado en otras latitudes con la ayuda de la minería de datos, para tener conocimiento de lo que se puede hacer en nuestro país.

En los últimos años, se ha registrado como la desnutrición se está volviendo un problema muy común en los menores de edad, debido a esto, en varios países ya se encuentran registrados varios estudios que ayuden a mitigar la desnutrición.

Se obtuvo que el uso de encuestas simplificadas permite monitorear la nutrición en salud pública de una manera más rentable y así poder reducir el número de errores a medida que se reduzca la carga de participación [5].

En Afganistán se realizó un estudio para la predicción de la desnutrición en niños donde utilizaron clasificadores como Bosques Aleatorios, PART de inducción y Naïve Bayes y tuvieron como resultados que la comparativa entre los clasificadores funcionaron bastante bien con una precisión alta y, en la mayoría de los casos, superior al 90% [6]. Otro país que realizó un estudio similar fue Corea, en donde se utilizó la metodología SEMMA (Sample, Explore, Modify, Model, and Assess) al momento de construir el modelo de datos [7].

En Portugal se enfocaron en el contexto de evaluar nutricionalmente y aplicar algoritmos de clasificación para predecir si un paciente debe tener seguimiento por un especialista en nutrición, realizaron preprocesamiento, transformación y limpieza de datos, aplicación de varios clasificadores y su respectiva evaluación a través de medidas de desempeño que incluyen la matriz de confusión, precisión, tasa de error [8].

Por último, se tiene un estudio en Colombia, en donde se tuvo como objeto de estudio a niños menores de 1 año que se distribuyen según una clasificación nutricional por antropometría compatible con riesgo de retraso en el crecimiento.

Se obtuvo que en los niños clasificados como de riesgo o retraso en el crecimiento al inicio de la intervención mostraron una mayor probabilidad de acercarse o estar en la trayectoria de crecimiento adecuada según el indicador de talla por edad después de la intervención [9].

Así como en otros países este problema ya, en Lima Metropolitana, se encuentran registrados más de 40,000 atenciones a niños que sufren de esta enfermedad y que lo convierte en un problema a resolver.

Otro estudio tuvo como objetivo evaluar la ingesta dietética de los estudiantes del cuarto ciclo de medicina de la Universidad Nacional Autónoma de Honduras. El estudio fue realizado teniendo una muestra de 65 estudiantes, la información fue procesada mediante el software NutrINCAP.

Dando como resultado, la ingesta dietética de los estudiantes es hipercalórica en hombres e hipocalórica en mujeres, además es hiperproteica y baja en fibra en ambos sexos [10].

Tabla 1. Diccionario de datos del estado nutricional de niños menores de 5 años

ID	Variable	Descripción
V001	Diresa	Zonas del departamento de lima donde vive el niño
V002	Red	Red a la que pertenece el centro de salud
V003	Microred	Microred a la que pertenece el centro de salud
V004	EESS	Calificación del establecimiento de salud
V005	Dpto_EESS	Departamento a la que pertenece el establecimiento de salud con la calificación EESS
V006	Prov_EESS	Provincia a la que pertenece el establecimiento de salud
V007	Dist_EESS	Distrito a la que pertenece el establecimiento de salud
V008	Renipress	Registro Nacional de Instituciones Prestadoras de Servicios de Salud
V009	Fecha Atencion	Fecha en que fue atendido
V010	Sexo	Sexo del niño
V011	FechaNacimiento	Fecha de nacimiento del niño
V012	EdadMeses	Edad en meses del niño
V013	UbigeoPN	Ubigeo del establecimiento de salud
V014	DepartamentoPN	Departamento a la que pertenece el establecimiento de salud
V015	ProvinciaPN	Provincia a la que pertenece el establecimiento de salud
V016	DistritoPN	Distrito a la que pertenece el establecimiento de salud
V017	Centro PobladoPN	Centro Poblado del establecimiento de salud
V018	Juntos	Programa nacional de apoyo, si el niño fue parte del programa
V019	SIS	Seguro Integral de Salud
V020	Pin	Programa Integral de nutrición
V021	Qaliwarma	Programa nacional de alimento, si el niño fue parte del programa

V022	Peso	Peso del niño
V023	Talla	Talla del niño
V024	Hemoglo bina	Nivel de Hemoglobina en niños
V025	FechaHemoglobina	Fecha donde se registra la hemoglobina del niño
V026	Cred	Control de crecimiento y desarrollo
V027	Suplementacion	Consumo de suplementos en niños
V028	Consejeria	Si el niño pasó por consejería de su estado nutricional
V029	Sesion	Si el niño tuvo una sesión con personal del área de nutrición
V030	UbigeoREN	Ubigeo del niño
V031	DepartamentoREN	Departamento donde vive el niño
V032	ProvinciaREN	Provincia donde vive el niño
V033	DistritoREN	Distrito al que pertenece el niño
V034	AlturaREN	Altura de la zona donde vive el niño
V035	HBC	Nivel de hemoglobina C en los niños
V036	Dx_Anemia	Nivel de anemia en los niños

La muestra tomada por la investigación no cuenta con suficientes datos, lo cual puede sesgar el desarrollo del modelo e influir en los resultados obtenidos. Otro punto por considerar es que la herramienta utilizada fue desarrollada con anterioridad en el Instituto de Nutrición de Centro América y Panamá, y puede estar estrechamente relacionada con la problemática del país siendo poco aplicable a otros contextos.

La evaluación nutricional de los pacientes requiere del manejo de una extensa información, ya que se analizan tantos aspectos relacionados con el proceso de la nutrición como la situación y evolución clínica del paciente, este estudio tuvo como objetivo desarrollar un programa informático que sirva como instrumento para la evaluación del estado nutricional del paciente.

La aplicación ofrece pronósticos nutricionales basados en parámetros antropométricos y bioquímicos, imágenes de estados de desnutrición, cuestionarios de caracterización de enfermedades, etc. [11].

Una de las debilidades que presentó el software desarrollado del estudio anteriormente mencionado es la cantidad de datos que necesita para analizar al paciente produciendo que no se completen todos y como consecuencia se tienen parámetros en blanco para el análisis.

El primer objetivo de esta investigación es aplicar técnicas de minería de datos para el preprocesamiento, limpieza, reducción y transformación a un lago de datos que ha permitido analizar la anemia en niños menores de 5 años, el segundo objetivo es aplicar algoritmos de Machine Learning para obtener el mejor modelo que permita predecir la anemia en niños menores de 5 años.

2. Trabajos relacionados

Los artículos analizados utilizaron diferentes modelos de machine learning, donde las variables que intervienen durante la elaboración que pueden

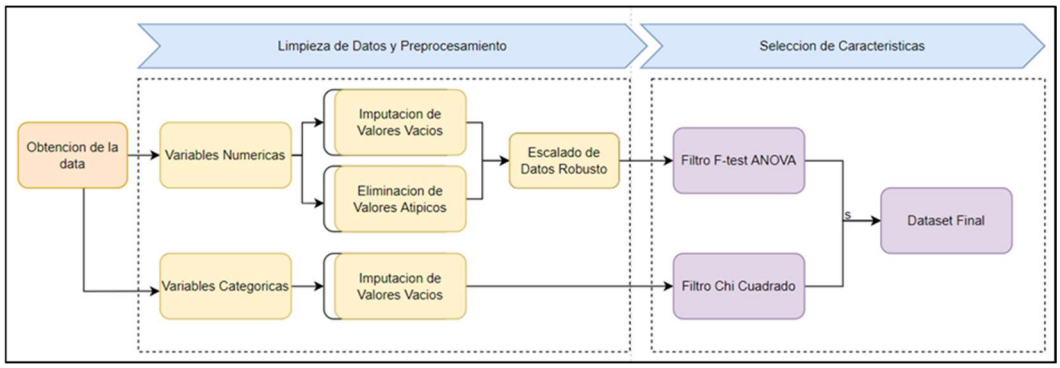

Fig. 1. Metodología propuesta para el modelamiento

variar significativamente de acuerdo con la problemática de la anemia de cada cultura, frente a esta variedad de modelos se decidió agrupar en tres:

2.1. Modelos de predicción de anemia infantil

Los modelos de clasificación utilizados para estudiar la problemática de anemia infantil son Árbol de Decisión, Bosques Aleatorios, K Vecinos más Cercanos, Redes Neuronales y Naive Bayes, siendo un total de 17 artículos que han utilizado uno o algunos de estos modelos.

Se identificó un variado uso de modelos, siendo el de uso más frecuente el árbol de decisión, en algunos artículos fue el único modelo que se empleó en toda la investigación, como en [13] que intenta predecir la anemia entre los niños y establecer una relación entre la salud y la dieta de la madre durante el embarazo y sus efectos sobre el estado anémico de su hijo, en la investigación [18] utilizando un modelo de árbol de decisiones que se combinó con el método antropométrico para monitorear el estado nutricional de los niños menores de cinco años, en [21] veía los hábitos dietéticos relacionados con el estado de obesidad de los niños y en [25] investigó los factores que más aportaron a la evaluación nutricional de niños de 6 a 11 años.

En otras investigaciones utilizaron diferentes modelos, como en [6] que propone un enfoque para predecir el estado de desnutrición, haciendo uso de cuatro modelos, en [8] también utilizaron varios modelos de clasificación, con el objetivo de conocer si un paciente necesita un seguimiento por un especialista de nutrición.

En [9] tiene un enfoque clasificación nutricional por antropometría compatible con riesgo de desnutrición crónica. Otras investigaciones como [17] que diseña un modelo que prediga el estado nutricional de niños menores de cinco años utilizando técnicas de minería de datos, u otros estudios [19, 20, 22, 23, 24] haciendo comparaciones con diferentes modelos de clasificación relacionados al problema de la anemia.

El uso de diferentes modelos se debe principalmente a la propuesta que se desea lograr, ya que en las investigaciones donde solo se usó el modelo de árbol de decisión está enfocado a realizar análisis o clasificaciones, a diferencia de los que usaron más de un modelo, están enfocados a realizar comparaciones para hallar los mejores resultados que cumplieran con su objetivo.

En los artículos recopilados se identificó que de la cantidad de muestras utilizadas, los que han usado una mayor cantidad son, el perteneciente al país de Corea del Sur que fue donde utilizaron la mayor cantidad de muestras las cuales pertenecen a un total de 45553 participantes y fueron obtenidas gracias a las organizaciones llamadas, Centro de Corea para el Control y Prevención de enfermedades, los cuales recuperan las muestras realizando periódicamente encuestas nacionales de Examen de salud y nutrición de su población [7].

La segunda estudio con mayor cantidad de muestras fue realizado en Perú, con una cantidad de muestras las cuales pertenecen a un total de 8901 niños y fueron obtenidos gracias al programa de desarrollo infantil temprano (ECD) de la

Tabla 2. Resumen global del conjunto de datos CENAN-2020

Nombre	Valor
Cantidad de variables	36
Cantidad de instancias	138369
Variables categóricas	30
Variables numéricas	6
Filas duplicadas	0
Tamaño total en memoria	43.9 KB

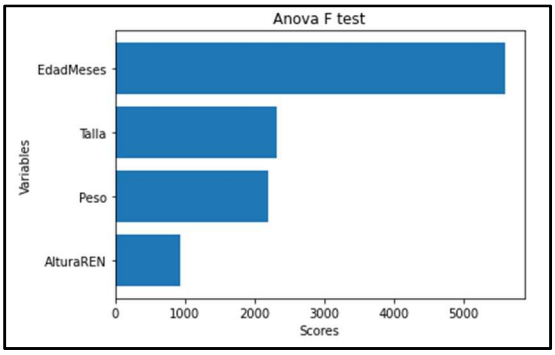

Fig. 2. Comparación de puntuaciones de las características numéricas en el Filtro Anova F-test

Encuesta Demográfica y de Salud Familiar (ENDES) del 2019 [15].

En cambio, donde la cantidad de muestras fue menor, en Cuba, donde se obtuvo 278 niños de 6 a 11 años como participantes [25] a través de un diseño observacional descriptivo.

Se tiene una gran diferencia de muestras usadas, se debe a que en los 2 primeros se usaron datos que ya estaban recopilados por instituciones del gobierno y organizaciones a través de encuestas nacionales, mientras que en el último la muestra fueron recopiladas específicamente para usarlas en esta investigación.

En nuestro trabajo de investigación seguiremos lo hecho por Corea del Sur, ya que se usarán una gran cantidad de muestras proporcionadas por el Instituto Nacional de Salud - Centro Nacional de Alimentación y Nutrición.

En la comparación de las herramientas utilizadas en los artículos que pertenecen a este grupo, se encuentra que varias investigaciones utilizaron el software llamado Weka (Waikato Environment for Knowledge Analysis) en sus diferentes versiones, en [8] con el objetivo de conocer si un paciente necesita un seguimiento por un especialista de nutrición, en [17] diseña un modelo que prediga el estado nutricional de niños menores de cinco años utilizando técnicas de minería de datos, en [19] explora la cantidad de alimentos sobre los que se requería información sobre la ingesta para predecir con precisión el cumplimiento, o no, de las recomendaciones dietéticas clave, en [21] estudia los hábitos dietéticos relacionados con el estado de obesidad de los niños, en [22] demostrar el análisis de la desnutrición en función de la ingesta de alimentos, el índice de riqueza, el grupo de edad, el nivel educativo, la ocupación, etc. y en [23] explora la cantidad de alimentos sobre los que se requería información sobre la ingesta para predecir con precisión el cumplimiento, o no, de las recomendaciones dietéticas clave.

Otra herramienta que se utilizó fue Anthro v1.0.4 para realizar el análisis de estado nutricional y JMP v11.0.0 para el análisis univariable y multivariable [15], así como también SAS Enterprise Miner [7].

En algunos casos se utilizaron estándares como el SMART o procesos como KDD utilizados en [6] que propone un enfoque para predecir el estado de desnutrición, haciendo uso de cuatro modelos.

Como resultado de esta comparación se podría concluir que Weka es la herramienta más utilizada para la minería de datos en las investigaciones, por otro lado, algunas investigaciones solo se limitan a usar los modelos, en vez de apoyarse en una herramienta.

En la medida final, se obtuvieron datos de 686 niños, identificando que el 17% de los niños progresaron de retraso en el crecimiento a un riesgo de retraso en el crecimiento y que el 4,5% recuperó su trayectoria de crecimiento, logrando una longitud adecuada para su edad.

Para apoyar su alimentación se realizó un control hecho de forma de cuchara; sin embargo,

los niños no lo tomaron muy en cuenta que, a diferencia de la realidad virtual, si lo hicieron [17].

Las investigaciones que utilizaron el algoritmo de árbol de decisión tuvieron en general una precisión en su predicción por encima del 71% esto es debido a que los investigadores escogieron acertadamente dicho algoritmo para el tipo de problema que buscaban resolver, por ejemplo, en el llevado a cabo por Giabbanelli y Adams obtuvieron resultados que van desde el 72% (Reino Unido) de precisión hasta 92.6% (Etiopía) [19].

En general cada investigación supo escoger el algoritmo adecuado para poder llegar al resultado esperado, siendo los problemas que podían resolverse con árboles de decisión los que más tendencia tienen a resolverse con soluciones relacionadas a la minería de datos.

2.2. Modelos de regresión para predicción de anemia infantil

Los modelos de regresión utilizados son, Regresión Lineal y Regresión Logística, siendo un total de 7 investigaciones que han utilizado uno o algunos de estos modelos.

En la investigación [28] desarrollados en el país de los Estado Unidos es aquel que posee un mayor número de algoritmos usados, entre ellos tenemos al de regresión lineal; cuyo objetivo fue probar si la eliminación cuidadosa de los elementos de dos encuestas de nutrición comunitaria guiadas por una técnica de minería de datos llamada selección de características, puede identificar un conjunto de datos reducido, sin dañar la señal dentro de esos datos.

Mientras que en otras investigaciones fue utilizado en conjunto con otros algoritmos como, bosques aleatorios, perceptrón multicapa y modelos de regresión [7] tiene un enfoque diferente donde desea identificar a los paciente con riesgo de enfermedades periodontales debido a una pobre nutrición, finalmente también se identificó investigaciones donde se usaron principalmente algoritmos de regresión como en [26] para examinar los factores de riesgo de retraso en el crecimiento entre los niños en edad preescolar y en [27] diseña un modelo de predicción para la desnutrición. En comparación con los modelos de clasificación, estos no son muy

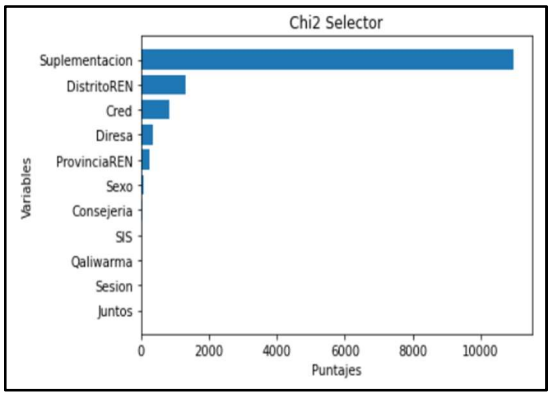

Fig. 3. Comparación de puntuaciones de las características categóricas en la Prueba chi-cuadrado

utilizados y tampoco son un uso principal para la problemática del estudio.

En donde mayor cantidad de muestra fue en un estudio donde se utilizó regresión lineal y otros algoritmos, siendo 9004 datos, los cuales se recopilaron utilizando el analizador de hematología automático Mindray BC-5300[23].

En otro estudio la muestra fue de 731 participantes [26] y en donde menor información se llegó a utilizar es el realizado en Australia con 295 datos y sin indicar la procedencia de la información en [20], teniendo como propuesta demostrar y comparar la utilidad de los métodos de minería de datos en la clasificación de un resultado categórico derivados de una intervención relacionada con la nutrición.

Aunque estos estudios no hagan uso principal de modelos de regresión, se aprecia que la cantidad de información y la procedencia de estos es variada.

En la comparación de las herramientas utilizadas en los artículos que pertenecen a este grupo, se presencia el uso de Weka en sus diferentes versiones [27, 28], pero no es la única, otras investigaciones usaron lenguaje de programación R, SAS Enterprise Miner o la biblioteca Scikit-Learn [24], además de RStudio [20]. Sin embargo, entre los estudios hay dos que no que no especifican la herramienta utilizada [9] [26]. Siendo en esta categoría de modelos una mayor variedad de herramientas.

Finalmente, en cuanto los resultados se pueden afirmar que las investigaciones que tuvieron, el bajo peso al nacer, la duración de la

Tabla 3. Conjunto de variables finales

Descripción	Denominación
Zonas del departamento de lima donde vive el niño. (Lima Rural, Lima Este, Lima Sur, Lima Norte y Lima Centro)	**Diresa**
Sexo del niño (Masculino y Femenino)	**Sexo**
Edad del niño en meses	**EdadMeses**
Peso del niño	**Peso**
Talla del niño	**Talla**
Control de crecimiento y desarrollo (No y Si)	**Cred**
Consumo de suplementos en niños (No y Si)	**Suplementacion**
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**

lactancia materna, la edad de la madre al nacer, la educación de la madre y la ocupación son los factores de riesgo asociados al retraso del crecimiento [26], en otro estudio muestran que las características relacionadas con la asistencia médica y paramédica y la creación de conciencia favorecen enormemente el abordaje de los problemas de desnutrición [27].

Los modelos de regresión lineal construidos a partir de los conjuntos de características reducidos tenían valores de 92% y 94% para los datos de restaurantes y tiendas de comestibles, respectivamente [28].

En otro estudio se da como resultado que los bosques aleatorios fue el que dio el error más bajo (RMSE = 0.0123) en comparación con el perceptrón multicapa (0.0954) y la regresión lineal (0.0137) [24]. Como conclusión, si bien los algoritmos de regresión muestran buenos resultados, no son siempre los mejores para la problemática.

2.3. Investigaciones relacionadas con predicción de anemia

Otros de los modelos utilizados, son K vecinos, Text mining, Análisis de sentimientos, Text clustering, siendo un total de 3 artículos, que no solo son de Machine Learning, sino también de Procesamiento del Lenguaje Natural.

Las técnicas de minería de texto, como el análisis a nivel de palabra (p. ej., análisis de frecuencia), el análisis de asociación de palabras (p. ej., análisis de red) y técnicas avanzadas (p. ej., clasificación de texto, agrupación de texto, modelado de temas, recuperación de información y análisis de sentimiento) [12], mientras que en otro estudio aplicaron técnicas similares, utilizando un enfoque de minería de texto utilizando el método de la bolsa de palabras a una muestra aleatoria de artículos obtenidos de todas las revistas en la categoría temática "Nutrición y Dietética" dentro del portal SCImago Journal and Country Rank y publicados en 2018 [14], y el último estudio, realizaron un análisis univariable y multivariable por clúster con el método K vecinos para establecer tipologías nutricionales[16].

Las muestras que se llegaron a utilizar son diferentes debido a los modelos o técnica utilizadas, por ejemplo, en el estudio de que utilizo K vecinos tuvo una 2.955 niños y 3.085 niñas, entre 0 a 60 meses, esta muestra fue obtenida en las unidades operativas de la Dirección Provincial de Salud de Chimborazo durante el año 2013 [16].

Con lo que respecta a las herramientas utilizaron Anthro v1.0.4 y JMP v11, análisis univariable y multivariable [16], MS Excel para la limpieza de los datos y R para utilizar la técnica de minería de texto [14] y en el último estudio no se especifica la herramienta [12].

Los resultados obtenidos son que los departamentos de salud pública, en el futuro, pueden identificar los peligros para la seguridad y la salud antes, para mejorar el desempeño de la administración de alimentos [12]. La estadística

Tabla 4. Comparación de las cuartiles, mínimos y máximos de los modelos analizados

Árbol de decisión (M1)					
	Q1	**Q2**	**Q3**	**Min**	**Max**
Sensibilidad	0.374	0.442	0.492	0.336	0.527
Especificidad	0.847	0.894	0.939	0.796	0.989
VPP	0.511	0.577	0.621	0.437	0.661
VPN	0.778	0.840	0.87	0.883	0.926
Regresión Logística (M2)					
	Q1	**Q2**	**Q3**	**Min**	**Max**
Sensibilidad	0.212	0.259	0.324	0.182	0.379
Especificidad	0.872	0.919	0.968	0.829	0.998
VPP	0.519	0.558	0.603	0.456	0.655
VPN	0.741	0.798	0.861	0.699	0.895
K Vecinos más cercanos (M3)					
	Q1	**Q2**	**Q3**	**Min**	**Max**
Sensibilidad	0.272	0.327	0.381	0.226	0.423
Especificidad	0.865	0.910	0.961	0.823	0.999
VPP	0.528	0.566	0.634	0.478	0.670
VPN	0.748	0.791	0.840	0.709	0.905
Bosques Aleatorios (M4)					
	Q1	**Q2**	**Q3**	**Min**	**Max**
Sensibilidad	0.316	0.363	0.419	0.274	0.473
Especificidad	0.875	0.909	0.957	0.823	0.998
VPP	0.556	0.609	0.651	0.505	0.699
VPN	0.718	0.774	0.818	0.866	0.913

Naive Bayes (M5)					
	Q1	**Q2**	**Q3**	**Min**	**Max**
Sensibilidad	0.687	0.739	0.795	0.647	0.844
Especificidad	0.619	0.671	0.730	0.584	0.783
VPP	0.379	0.437	0.499	0.340	0.535
VPN	0.834	0.885	0.950	0.793	0.987

Tabla 5. Comparación de las métricas de exactitud, precisión y recall

Modelo	**Exactitud**	**Precisión**	**Recall**
Árbol de decisión	0.776	0.562	0.430
Regresión logística	0.766	0.555	0.280
K vecinos más cercanos	0.773	0.574	0.324
Bosques Aleatorios	0.784	0.604	0.374
Naive Bayes	0.699	0.437	0.746

descriptiva numérica fue el grupo de método estadístico más común, apareciendo en el 83,2% de los artículos.

Las estadísticas de IBM SPSS fueron el paquete de software estadístico más común, reportado en el 41,7% de los artículos incluidos [14]. y en el último estudio se tuvo que el conglomerado 3, el de mayor relevancia nutricional, presenta las siguientes características; menor T//E, mayor IMC//E, menor edad, menor tiempo de lactancia exclusiva, menor edad de destete, mediana prescripción de hierro y vitamina A [16].

Estas investigaciones que tienen la misma problemática de desnutrición toman un enfoque diferente, mientras unos analizan variables otros hacen un análisis de diferentes artículos para sacar conclusiones con ayuda de minería de texto.

3. Materiales y métodos

La metodología empleada se muestra en la Fig.1. que existe dos procesos que parten de la obtención de la data: Limpieza y preprocesamiento de datos, en el cual se realizó la imputación de valores vacíos, la eliminación de valores atípicos y el escalado de datos de las variables numéricas y categóricas para pasar al siguiente proceso: Selección de Características donde se usaron los filtros de F-test y Chi cuadrado para poder obtener el dataset final

3.1. Limpieza y preprocesamiento de datos

El conjunto de datos utilizado para el presente estudio fue obtenido de las Historias Clínicas de los Establecimientos de Salud del Ministerio de Salud Perú, publicado por Instituto Nacional de Salud - Centro Nacional de Alimentación y Nutrición (CENAN) con información de Lima actualizada del 2020 en la Plataforma de Datos Abiertos del Perú [29].

La dimensión de este conjunto de datos es de 138369 instancias, compuestos por 36 variables las cuales están relacionadas a las características físicas, niveles de hemoglobina y anemia, tiempo de nacimiento, lugar de estadía, centros de salud

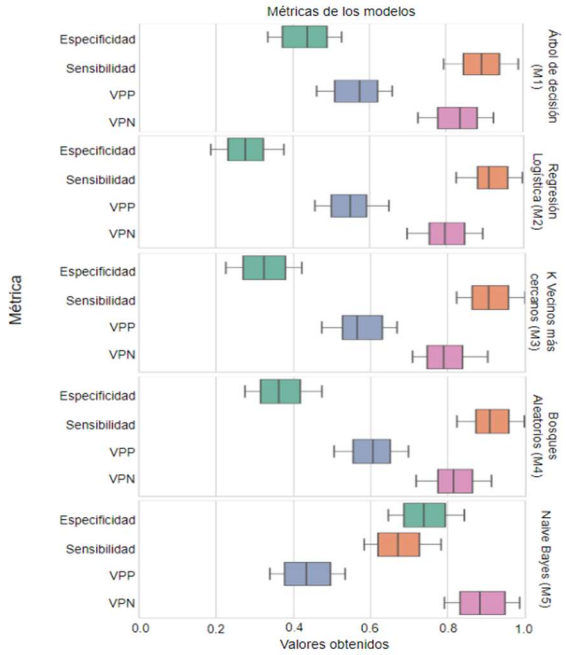

Fig. 4. Diagrama de cajas y bigotes de las métricas

y programas de apoyo con respecto a niños menores de 5 años, mostrándose todas las variables en detalle en la “Tabla 1”.

Las variables están divididas en 30 de tipo categórico y 6 del tipo numérico, lo que indica que el conjunto de datos es desequilibrado, es decir hay una desigualdad en la cantidad de los tipos de variable, tal como se observa en el resumen global en la “Tabla 2”.

Este conjunto de datos después de su extracción fue recopilado en un repositorio propio [33] para un mejor acceso.

De las 36 columnas, se eliminó a las columnas o variables cuyos valores son únicos, nulos, vacíos o repitentes [30]. Por ejemplo, la columna Departamento REN tiene un valor único, también la columna Fecha Hemoglobina que tiene 126094 valores vacíos siendo el 91.1% de los de los datos de esa columna y las columnas que tienen los mismos datos como Prov_EESS y ProvinciaREN.

Por lo tanto, solo se utilizarán las columnas que tengan a lo más un 20% de datos vacíos o nulos, quedando con 16 columnas en el conjunto de datos.

Se procedió a separar el conjunto de datos en numéricas con 4 columnas (EdadMeses, Peso, Talla y AlturaREN) y categóricas 12 columnas (Diresa, Sexo, Juntos, SIS, Qaliwarma, Cred, Suplementacion, Consejeria, Sesion, ProvinciaREN, DistritoREN, Dx_Anemia), para la imputación de datos faltantes se procedió a identificar las columnas que presentaban este inconveniente resultando las variables categóricas: Juntos, SIS y Qaliwarma; y las variables numéricas: Peso y Talla.

Para las variables categóricas se reemplazaron las celdas vacías de los programas sociales por 0 pues se considera que si dicha celda está vacía es porque dicho registro no cuenta con dichos servicios sociales.

El resto de las variables categóricas de tipo Object también fueron reemplazados por su moda. Por otro lado, en el caso de las variables numéricas se reemplazaron las celdas vacías de Peso y Talla con sus respectivas medias.

Para la eliminación de valores atípicos se tuvo que crear una variable numérica: Índice de Masa Corporal que sirvió como referencia a la hora de identificar los valores atípicos.

Es así como para el umbral mínimo de peso y talla de acuerdo a [31] el peso mínimo a considerar será 212 gramos y la talla mínima 24 cm. Los valores fuera de estos rangos son considerados anómalos y son eliminados del dataset.

Para los límites superiores de acuerdo con la información de [32] el peso máximo será de 50 kg y la talla máxima será de 170 cm.

Los valores fuera de estos rangos son considerados anómalos y son eliminados del dataset.

3.2. Selección de características

Para seleccionar las mejores variables o características se escoge aquellas que tengan las puntuaciones más altas (score), para determinar estas puntuaciones en variables numéricas se ha usado estadístico F de ANOVA y en variables categóricas se ha usado el estadístico chi- cuadrado.

Para la variable Dx_Anemia, la cual es la variable salida u objetivo a clasificar, se establecieron 2 categorías, el número “0” que es para la categoría normal y el número “1” para la categoría de anemia (leve, moderada y alta).

Tabla 6. Comparación de las métricas de los diferentes experimentos

Árbol de decisión (M1)			
	Exactitud	**Precisión**	**Recall**
Aplicando Anova y chi-cuadrado	0.776	0.562	0.430
Aplicando Anova pero sin chi-cuadrado	0.792	0.643	0.358
Aplicando chi-cuadrado pero sin Anova	0.775	0.796	0.266
Sin aplicar ambos filtros	0.792	0.623	0.406
Regresión Logística (M2)			
	Exactitud	**Precisión**	**Recall**
Aplicando Anova y chi-cuadrado	0.766	0.555	0.280
Aplicando Anova pero sin chi-cuadrado	0.783	0.617	0.326
Aplicando chi-cuadrado pero sin Anova	0.769	0.564	0.307
Sin aplicar ambos filtros	0.774	0.584	0.307
K Vecinos más cercanos (M3)			
	Exactitud	**Precisión**	**Recall**
Aplicando Anova y chi-cuadrado	0.773	0.574	0.324
Aplicando Anova pero sin chi-cuadrado	0.785	0.608	0.378
Aplicando chi-cuadrado pero sin Anova	0.775	0.579	0.339
Sin aplicar ambos filtros	0.781	0.595	0.364
Bosques Aleatorios (M4)			
	Exactitud	**Precisión**	**Recall**
Aplicando Anova y chi-cuadrado	0.784	0.604	0.374
Aplicando Anova pero sin chi-cuadrado	0.795	0.652	0.369
Aplicando chi-cuadrado pero sin Anova	0.785	0.605	0.385
Sin aplicar ambos filtros	0.795	0.652	0.375

Naive Bayes (M5)			
	Exactitud	**Precisión**	**Recall**
Aplicando Anova y chi-cuadrado	0.699	0.437	0.746
Aplicando Anova pero sin chi-cuadrado	0.687	0.427	0.774
Aplicando chi-cuadrado pero sin Anova	0.717	0.455	0.713
Sin aplicar ambos filtros	0.714	0.452	0.722

3.2.1. Aplicando el filtro Anova F-test

Para utilizar este filtro se empleó la librería sklearn, esta nos permitió realizar el filtro de Anova F-test, la cual nos facilitara escoger las variables numéricas más significativas para nuestra investigación.

En esta investigación el filtro de Anova F-test se aplicó a las 4 variables numéricas con las que cuenta nuestra data. Las pruebas se hicieron con respecto a la variable numéricas del grado de anemia (Dx_Anemia), esto con el fin de asociar las variables numéricas con la nutrición.

Realizado el filtro de Anova F-test se obtuvo una gráfica de barras horizontales, que permitió observar las variables numéricas más significativas y las menos significativas.

Después de la comparación de puntuaciones de las características numéricas Fig.2. se decidió eliminar la variable, AlturaREN, ya que fue la que menos puntuación obtuvo y nos quedamos con las siguientes 3 variables numéricas: Edad Meses, Talla y Peso.

Finalmente, en el preprocesamiento, se ha validado 16 variables influyentes del total de 36 variables. En la minería de datos hay técnicas de reducción que permiten validar variables influyentes de acuerdo al objetivo.

Por lo que en esta investigación se ha utilizado el Filtro Chi-Cuadrado para reducir variables categóricas y se ha utilizado el Filtro Anova F-Test para reducir variables numéricas.

3.2.2. Aplicando el filtro Chi-cuadrado

Para utilizar este filtro se empleó la librería scipy y sklearn, esta nos permitió hacer la prueba de chi-cuadrado, la cual nos facilitara escoger las variables categóricas más significativas para nuestra investigación.

En esta investigación el filtro de chi-cuadrado se aplicó a las 12 variables categóricas con las que cuenta nuestra data. Las pruebas se hicieron con respecto a la variable categórica del grado de anemia (Dx_anemia), esto con el fin de asociar las variables categóricas con la nutrición.

Realizado el filtro de Chi-cuadrado se obtuvo una gráfica de barras horizontales, que permitió observar las variables categóricas más significativas y las menos significativas.

Después de la Comparación de puntuaciones de las características categóricas Fig.3 se decidió eliminar las 5 siguientes: Juntos, Sesion, Qaliwarma, SIS y Consejeria, debido a que fueron las que obtuvieron menor puntaje en la prueba de Chi-cuadrado, quedándonos con 6 de las 11 variables categóricas iniciales.

Posterior de realizar el Anova F-Test y el chi-cuadrado, la cantidad de variables se redujo a 10 para nuestra data final como se puede visualizar en la “Tabla 3” y se recopilaron en un repositorio propio [33].

3.3. Modelado predictivo

3.3.1. Conjunto de datos de entrenamiento y validación

En esta tarea se realizó la partición del conjunto de datos en 70% para el entrenamiento y 30% para la validación de los modelos.

Los resultados de esta partición nos dan 95963 instancias para entrenar y 41127 instancias para validar. Tras la partición del conjunto de datos, se procede a realizar el entrenamiento de los

modelos y su respectiva evaluación según las métricas establecidas.

3.3.2. Buscando el mejor algoritmo para predecir anemia

En esta tarea se empleó los modelos de Árbol de decisión, Regresión logística, K vecinos más cercanos (KNN), Bosques Aleatorios y Naive Bayes, se llegó a escoger estos modelos debido a que en la mayoría de las investigaciones fueron los más empleados y con mejores resultados proporcionados, además sus diversas aplicaciones e información disponible.

Para el entrenamiento de los modelos se consideraron los valores de hiperparámetros por defecto que la librería scikit-learn 1.0.2 establece, a excepción de los parámetros max_depth = 7, n_neighbors = 10 y min_samples_leaf = 8, como se muestra a continuación:

- Árbol de decisión: {max_depth: 7}.
- Regresión logística: Valores por defecto.
- K vecinos más cercanos (KNN): {n_neighbors: 10}.
- Bosques Aleatorios: {min_samples_leaf: 8}.
- Naive Bayes: Valores por defecto.

3.3.3. Evaluación y comparación de algoritmos

En este punto se tuvo la necesidad de calcular las métricas de exactitud, sensibilidad, especificidad, área bajo la curva ROC (AUC), valor predictivo positivo, valor predictivo negativo, sin embargo, debido a que la distribución de las clases en la variable objetivo no se encuentra equilibrada se selecciona el mejor modelo en función a la sensibilidad, área bajo la curva ROC (AUC), menor número de falsos negativos y mayor número de verdaderos positivos que se obtienen la matriz de confusión.

4. Resultados

En este estudio se evaluó el desempeño de cinco algoritmos de aprendizaje automático en la predicción del estado nutricional de niños menores de 5 años de Lima en el 2020.

Estos algoritmos son: Árbol de Decisión (M1), Regresión logística (M2), K vecinos más cercanos (M3), Bosques Aleatorios (M4) y Naive Bayes (M5). Primero realizamos la comparación de las métricas de exactitud, precisión y Recall obtenidas, para los casos de aplicación de los filtros de Anova y chi-cuadra, aplicación de solo uno de estos filtros y cuando no se aplicó ninguno de los filtros.

4.1. Diagrama de cajas y bigotes

En la Fig. 4, se observa el diagrama de cajas y bigotes de las métricas obtenidas en el conjunto de datos de entrenamiento. Donde se compara los 5 modelos y por cada modelo se utilizan las siguientes métricas: Especificidad, Sensibilidad, Valor Predictivo Positivo (VPP), Valor Predictivo Negativo (VPN).

Analizando las medianas (Q2) podemos afirmar que: el modelo con el peor valor de Especificidad es el de Regresión Logística (M2) y el de mejor valor el modelo de Naive Bayes (M5); el modelo con los peores valores de Sensibilidad y VPP fue el de Naive Bayes (M5) y el de mejores valores fue el modelo de Bosque Aleatorio(M4); y finalmente el modelo con el peor valor de VPN fue el de K Vecinos más cercanos (M3) y el mejor valor fue el modelo de Naive Bayes (M5).

Siendo los modelos que destacan el modelo de Naive Bayes y el de Bosques Aleatorios. Donde el modelo de Naive Bayes diagnóstica más niños con anemia adecuadamente, además, el modelo de Bosques Aleatorios diagnóstica, más niños sanos adecuadamente.

Por otro lado, partiendo de los valores de cuartiles Q1 y Q3 se aprecia que en el modelo de Bosques Aleatorios la mayoría de los datos predichos están entre 0.556 y 0.651, lo que significa que entre el 55.60 % y 65.10 % de niños son diagnosticados con anemia, del mismo modo en el modelo de Naive Bayes la mayoría de los datos predichos están entre 0.834 y 0.950, lo que significa que entre el 83.5 % y 95 % de niños son diagnosticados como sanos.

Con respecto a los valores mínimos y máximos, el modelo con el valor de Especificidad más baja fue el de Naive Bayes (0.584) y el modelo con el valor de Especificidad más alto fue el de KNN (0.999); el modelo con el valor de Sensibilidad más baja fue el de Regresión Logística (0.182) y el modelo con el valor de Sensibilidad más alto fue el

de Naive Bayes (0.844); el modelo con el valor de VPP más baja fue el de Naive Bayes (0.340) y el modelo con el valor de VPP más alto fue el de Bosques Aleatorios (0.699); y por último el modelo con el valor de VPN más baja fue el de Regresión Logística (0.699) y el modelo con el valor de VPN más alto fue el de el de Naive Bayes (0.987).

En la “Tabla 4”, se muestra una comparación con base a los diagramas de cajas y bigotes de los 5 modelos, y sus 4 métricas, donde cada métrica se divide en primer cuartil (Q1), segundo cuartil (Q2) y tercer cuartil (Q3), Min y Max, siendo Q1, Q2 y Q3 los segmentos verticales del diagrama de la caja, Min el valor mínimo y Max es el valor máximo que puede tomar cada métrica.

4.2. Matriz de confusión

En la Fig. 5, muestra los resultados de las matrices de confusión que al comparar los diferentes resultados que se obtuvo de los 5 algoritmos (Árbol de Decisión (M1), Regresión logística (M2), K vecinos más cercanos (M3), Bosques Aleatorios (M4) y Naive Bayes (M5)) se obtuvo que el modelo Naive Bayes (M5) en comparación con los demás modelos, obtuvo el menor puntaje en el cuadrante de verdadero negativo con 21178 (Niños que no tienen anemia y están clasificados como tal.), este también obtuvo a su vez el mayor puntaje en el verdadero positivo con 7597 (Niños que tienen anemia y están clasificados como tal) y en el falso positivo posee el mayor valor con 9770 (Niños que no tienen anemia y están clasificados como anémicos).

El modelo Regresión Logística (M2) posee el mayor valor en comparación con los demás modelos, obtuvo el mayor puntaje en el cuadrante de verdadero negativo con 28663 (Niños que no tiene anemia y están clasificados como tal.) y este obtuvo a su vez el menor puntaje en el verdadero positivo con 2851 (Niños que tienen anemia y están clasificados como tal).

Descripción de las Métricas obtenidas luego de la predicción del conjunto de datos. Para la evaluación de los modelos utilizados se comparan las métricas de precisión y recall, debido a que permite evaluar la calidad de las predicciones del modelo para que esta se realice con exactitud.

En este caso se tomó las métricas de los valores positivos, tanto para la evaluación de la métrica de precisión como para recall.

Como se puede observar en la “Tabla 5” el modelo que cuenta con mayor precisión es Bosque Aleatorios, sin embargo, el valor de su recall es el tercero más bajo en comparación con los otros modelos, lo cual indica que el modelo no puede detectar correctamente la clase.

Por otro lado, el modelo de Naive Bayes es el que presenta mejor recall, pero su valor en la métrica de precisión es el más bajo entre todos los modelos, pese a esto el modelo detecta bien la clase, no obstante, también incluye muestras de la otra clase.

4.3. Reporte de clasificación

De los reportes de clasificación que se muestran a continuación se obtuvieron las siguientes conclusiones:

- De las instancias clasificadas como positivas, el que tuvo mayor porcentaje de exactitud fue el modelo de Bosques Aleatorios (0.784), seguido por Árbol de Decisión (0.776), K vecinos más cercanos (0.773), Regresión Logística (0.766) y el menor porcentaje fue de Naive Bayes (0.699).
- Los modelos que tienen mayor porcentaje de precisión fueron los modelos de Bosques Aleatorios (0.604) y K vecinos más cercanos (0.574), seguidos por Árbol de Decisión (0.562), Regresión Logística (0.555) y el de menor porcentaje fue de Naive Bayes (0.428).
- Los modelos con mayor porcentaje de clasificación correcta fueron, Naive Bayes (0.746), y Árboles de Decisión (0.430), seguidos por el Bosque Aleatorios (0.374), K vecinos más cercanos (0.324) y el de menor porcentaje fue Regresión logística (0.280).

4.4. Análisis comparativo de los modelos sin aplicar los filtros en la selección de características

Se realizó un análisis comparativo en la “Tabla 6” para tener la certeza de la obtención de un mejor resultado cuando se aplicaron los filtros de Anova y chi-cuadrado, por lo cual se obtuvieron las

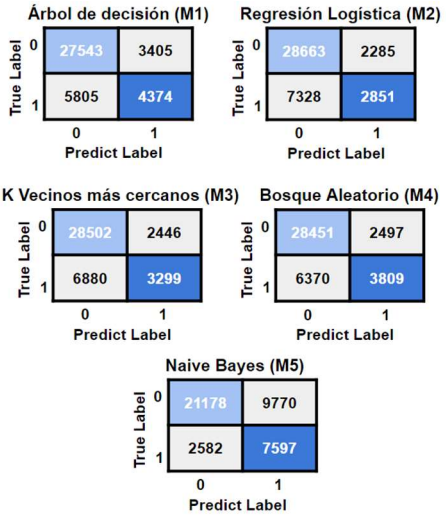

Fig. 5 Matriz de confusión de modelos estudiados

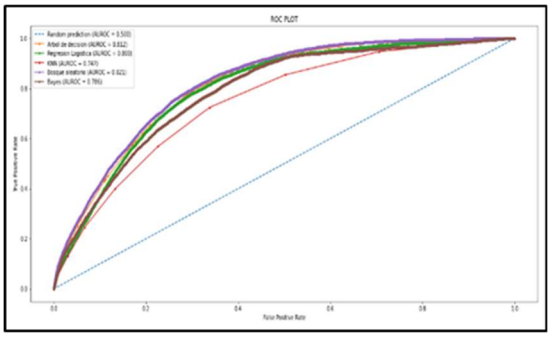

Fig. 6. Gráfica de las curvas ROC de los modelos procesados

métricas de exactitud, precisión y recall del experimento de cuando se aplicó Anova pero sin chi-cuadrado, del experimento de chi-cuadrado pero sin Anova y del experimento donde no se aplicó ninguno de los filtros.

En la “Tabla 6”, se puede analizar que el modelo Naive Bayes (M5) es el más indicado para los distintos tipos de experimentos ya que se adapta a nuestro caso de estudio al poseer un alto recall y una baja precisión, lo cual es ideal para nuestro conjunto de datos desequilibrado.

Los mejores resultados obtenidos fueron cuando se aplicó ambos filtros y cuando se hizo el experimento de aplicar Anova pero sin chi- cuadrado.

Para determinar el mejor entre estos dos, se analizó una tercera métrica la cual es, la exactitud, donde el valor dado por esta métrica debe ser el mayor posible, obteniendo que para el caso donde se aplicaron los 2 filtros la exactitud fue de 0.699 y en el experimento donde se aplicó Anova pero no chi-cuadrado fue de 0.687, lo cual nos permite demostrar que es necesario aplicar los filtros de Anova y chi-cuadrado para obtener mejores resultados.

4.5. Curva ROC

Como resultado se obtuvo la curva ROC de todos los modelos tratados y se pusieron todas en una misma gráfica para la comparación y evaluación de los modelos. Esta gráfica se puede observar en nuestra figura 6. En donde se muestra nuestra predicción random, que es la recta del

Tabla 7. Puntuación AUC obtenida del entrenamiento

Algoritmo	Puntuación AUC
Árbol de Decisión	0.812
Regresión Logística	0.800
KNN	0.749
Bosque Aleatorio	0.822
Naive Bayes	0.785

medio, que nos ayudará a poder dividir nuestros modelos.

Como se puede ver las curvas, el modelo de KNN obtuvo un menor valor y es el peor modelo para poder representar nuestra predicción. Con respecto a los demás, los valores son muy similares. Por ello, se procedió a utilizar una comparación de sus puntuaciones que se representan en la "Tabla 7".

Con estas puntuaciones podemos ver que nuestro modelo de Bosque Aleatorio ha obtenido una puntuación de 0.822, siendo la mejor puntuación obtenida.

Los modelos de Árbol de Decisión, Regresión logística y Bayes obtuvieron una puntuación superior al 0.785 y son considerados como modelos a seguir y tomar en cuenta. El modelo KNN resultó ser nuestro modelo que obtuvo la puntuación más baja ya que 0.749, siendo así el peor modelo.

5. Discusión

En las métricas de precisión y recall se tiene que el modelo Naive Bayes posee los valores de 0.43 y 0.74 respectivamente, siendo el mejor modelo en nuestro caso ya que se tiene un conjunto de datos desequilibrados.

En [23] el resultado obtenido en las métricas de precisión fue de 0.93 y de recall fue de 0.92 para el mismo modelo, sin embargo, no fue elegido como el mejor al ser comparado con el modelo de árbol de decisión J48 quien brinda el mejor rendimiento, recall, tasa de verdaderos positivos, precisiones y el valor más bajo en los falsos positivos.

Los resultados obtenidos por [13] utilizando árboles de decisión fueron superiores al 90%, mientras que en esta investigación se obtuvo el 56% de precisión, esta diferencia puede ocurrir debido a que se empleó la técnica de 25 reglas de decisión para discriminar cada paso que seguía el algoritmo, también se menciona una preclasificación de las clases que se emplearon al momento de generar las reglas de decisión, lo cual en esta investigación no se realizó.

Otro factor importante es considerar variables que no se encuentran relacionadas al problema a simple vista, pero su implementación podría mejorar la visión del problema.

Los modelos analizados no presentan modificaciones para mejorar el entrenamiento de los datos, sin embargo, en algunos estudios como en [17], se menciona que combinar algunas técnicas ayuda a mejorar y reforzar las capacidades evitando que se produzcan sesgos al momento de aplicar el algoritmo.

Así mismo ayuda a minimizar el tiempo que se emplea durante el procesamiento de cada entrada al modelo.

En la Fig. 6 se muestra el área bajo la curva ROC, donde el bosque aleatorio obtuvo un valor de AUC=0.82 en la curva ROC, y en segundo lugar se encuentra al modelo de árbol de decisión con un valor de AUC=0.81, demostrando la capacidad de discriminar si un niño tiene anemia o no. En [7, 17, 20] se obtienen valores similares para la clasificación utilizando los modelos de árboles de decisión, redes neuronales y regresión.

6. Conclusiones

En esta investigación se ha aplicado técnicas de preprocesamiento, además se ha aplicado el filtro ANOVA F-test para reducir las variables numéricas y el filtro Chi-Cuadrado para reducir las variables categóricas. Los resultados demuestran que las variables más importantes son 9 más la variable objetivo.

En cuanto al segundo objetivo, tras la evaluación y comparación de los modelos construidos por los algoritmos de Árbol de decisión, Regresión logística, K vecinos más cercanos (KNN), Bosques Aleatorios y Naive Bayes, el algoritmo que obtuvo los mejores resultados en el conjunto de datos analizados fue el de Naive Bayes al poseer el mayor valor en verdaderos positivos con 7597 y falso positivos con 9770 en la matriz de confusión, además obtuvo una baja precisión de 0.43 y un alto recall de 0.74, ya que se tiene un conjunto de datos desequilibrados, es preferible para nuestro caso que algunos niños sanos sean etiquetados como anémicos en lugar de dejar a unos niños anémicos etiquetados como sanos.

Referencias

1. **Sánchez-Abanto, J. (2014).** Evolución de la desnutrición crónica en menores de cinco años en el Perú, Revista Peruana de Medicina Experimental y Salud Pública, Vol. 29, No. 3, pp. 402–407. DOI: 10.17843/RPMESP.2012.293.377.
2. **INEI (2021)**. Indicadores de resultados de los programas presupuestales, 2015-2020. https://proyectos.inei.gob.pe/endes/2020/ppr/Indicadores_de_Resultados_de_los_Programas_Presupuestales_ENDES_2020.pdf.
3. **Aldana**, **Ú. (2013).** La desnutrición crónica en Lima Metropolitana. Argumentos Hacia un diagnóstico Lima Metropolitana, pp. 21–24.
4. **Organización Mundial de la Salud (2021)**. Anemia.https://www.who.int/es/healthtopics/anaemia#tab=tab_1
5. **Rosso, N., Giabbanelli, P. (2018).** Accurately inferring compliance to five major food guidelines through simplified surveys: Applying data mining to the UK national diet and nutrition survey. JMIR Public Health and Surveillance, Vol. 4, No. 2. DOI: 10.2196/publichealth.9536.
6. **Momand, Z., Mongkolnam, P., Kositpanthavong, P., Chan**, **J. (2020).** Data mining-based prediction of malnutrition in Afghan children. Procedings 12th International Conference on Knowledge and Smart Technology, pp. 12–17. DOI: 10.1109/KST48564.2020.9059388.
7. **Lee, J., Jeong, S., Choi, S. (2018).** Predictive data mining for diagnosing periodontal disease: the Korea national health and nutrition examination surveys (KNHANES V and VI) from 2010 to 2015. Journal of Public Health Dentistry, Vol. 79, No. 1, pp. 44–52. DOI: 10.1111/JPHD.12293.
8. **Ferreira, D., Peixoto, H., Machado, J., Abelha, A. (2018).** Predictive data mining in nutrition therapy. Proceedings of 13th APCA International Conference on Automatic Control and Soft Computing, pp. 137–142. DOI:10.1109/CONTROLO.2018.8516413.
9. **Castro-Prieto, P., Trujillo-Ramirez, K. M., Moreno, S., Holguín, J. S., Pineda, D. M., Tomasi, S., Ramirez-Varela, A. (2021).** Reduction of chronic malnutrition for infants in Bogotá, Colombia. BMC Public Health, Vol. 21, No. 1, DOI: 10.1186/s12889-021-10620-3.
10. **Gonzales-Pineda J. O., Mejía-Rodríguez, S. A., Corea-Cruz, C. R., Sánchez-Mendoza, J. G., Majano-Hernández, W. R., Carranza-Linares, R. J., Elvir-Gale, P. M. (2017).** Evaluación de la ingesta dietética en estudiantes de cuarto año de medicina. Revista Científica De La Escuela Universitaria De Las Ciencias De La Salud, Vol. 4, No. 2, pp. 51–57. DOI: 10.5377/rceucs.v4i2.7112.
11. **Garcia de Diego, L., Cuervo, M., Martinez, J. A. (2013).** Programa informático para la realización de una valoración nutricional fenotipica y genotipica integral. Nutrición Hospitalaria, Vol. 28, No. 5, pp. 1622–1623. DOI: 10.3305/nh.2013.28.5.6622.
12. **Tao, D., Yang, P., Feng**, **H. (2020).** Utilization of text mining as a big data analysis tool for food science and nutrition. Comprehensive Reviews in Food Science and Food Safety,

Vol. 19, No. 2, pp. 875–894. DOI: 10.1111/1541-4337.12540.

13. **Meena, K., Tayal, D., Gupta, V., Fatima, A. (2019).** Using classification techniques for statistical analysis of Anemia. Artificial Intelligence in Medicine, Vol. 94, pp. 138–152. DOI: 10.1016/J.ARTMED.2019.02.005.
14. **Coenen, A., Batterham, M., Beck**, **E. (2021).** Statistical methods and software used in nutrition and dietetics research: A review of the published literature using text mining. Nutrition & Dietetics, Vol. 78, No. 3, pp. 333–342. DOI: 10.1111/1747-0080.12678.
15. **Hernández, A., Pisfil, N., Vargas, R., Azañedo, D. (2021).** Nutritional status and effective verbal communication in Peruvian children: A secondary analysis of the 2019 Demographic and Health Survey. PLoS One, Vol. 16, No. 2, DOI: 10.1371/JOURNAL.PONE.0246542.
16. **Ramos-Padilla, P., Delgado, V., Villavicencio, V., Carpio, T. (2018).** Tipologías nutricionales en población infantil menor de 5 años de la provincia de Chimborazo, Ecuador. Revista Española de Nutrición Humana y Dietética, Vol. 22, No. 4, pp. 287–294. DOI: 10.14306/RENHYD.22.4.695.
17. **Markos**, **Z., Doyore, F., Yifiru, M., Haidar, J. (2014).** predicting under nutrition status of under-five children using data mining techniques: The case of 2011 Ethiopian demographic and health survey. Journal of Health & Medical Informatics, Vol. 5, No. 2. DOI: 10.4172/2157-7420.1000152.
18. **Ermatita, Destriatania, S., Yulnelly. (2020).** Nutrition anthtropometric status model by data mining: case study in Palembang South Sumatera. International Journal of Engineering Trends and Technology, Vol. 13, No. 1, pp. 97–103. DOI: 10.14445/22315381/CATI2P215.
19. **Giabbanelli, P., Adams**, **J. (2016).** Identifying small groups of foods that can predict achievement of key dietary recommendations: data mining of the UK National Diet and Nutrition Survey, 2008-12. Published Online by Cambridge University Press, Vol. 19, No. 9, pp. 1543–1551. DOI: 10.1017/S1368980016000185.
20. **Batterham, M., Neale, E., Martin, A., Tapsell, L. (2017).** Data mining: Potential applications in research on nutrition and health. Nutrition and Dietary Assessment Methodology, Vol. 74, No. 1, pp. 3–10. DOI: 10.1111/1747-0080.12337.
21. **Lazarou, C., Karaolis, M., Matalas, A., Panagiotakos, D. (2012).** Dietary patterns analysis using data mining method. An application to data from the CYKIDS study. Computer Methods and Programs in Biomedicine, Vol. 108, No. 2, pp. 706–714. DOI: 10.1016/J.CMPB.2011.12.011.
22. **Thangamani, D., Sudha, P. (2014).** Identification of malnutrition with use of supervised datamining techniques –decision trees and artificial neural networks. International Journal of Engineering and Computer Science, Vol. 3, No. 9, pp. 8236– 8241.
23. **Abdullah, M., Al-Asmari**, **S. (2017).** Anemia types prediction based on data mining classification algorithms. Communication, Management and Information Technology, pp. 615–621.
24. **Sayed, E. (2019)**. A machine learning model for hemoglobin estimation and anemia classification. International Journal of Computer Science and Information Security, Vol. 17, No. 2, pp 100–108.
25. **Martín, R., Ramos, R. S., Grau, R., García, M. M. (2007).** Aplicación de métodos de selección de atributos para determinar factores relevantes en la evaluación nutricional de los niños. Gaceta Médica Espirituana, Vol. 9. No. 1.
26. **Sk, R., Banerjee, A., Rana**, **M. J. (2021).** Nutritional status and concomitant factors of stunting among pre-school children in Malda, India: A micro-level study using a multilevel approach. BMC Public Health, Vol. 21, No. 1, pp. 1–13. DOI: 10.1186/S12889-021-11704-W/TABLES/2.
27. **Khare, S., Kavyashree, S., Gupta, D., Jyotishi**, **A. (2017).** Investigation of nutritional status of children based on machine learning

techniques using Indian demographic and health survey data. Procedia Computer Science, Vol. 115, pp. 338–349. DOI: 10.1016/J.PROCS.2017.09.087.

28. **Partington, S. N., Papakroni, V., Menzies**, **T. (2014).** Optimizing data collection for public health decisions: A data mining approach. BMC Public Health, Vol. 14, No. 593. DOI: 10.1186/1471-2458-14-593.
29. **Instituto Nacional de Salud Centro Nacional de Alimentación y Nutrición (2021).** Sistema de información del Estado Nutricional de niños y gestantes Perú INS/CENAN (Instituto Nacional de Salud Centro Nacional de Alimentación y Nutrición) | Plataforma Nacional de Datos Abiertos. https://www.datosabiertos.gob.pe/dataset/sistema-de-información-del-estado-nutricional-de-niños-y-gestantes-perú-inscenan-instituto.
30. **Brownlee**, **J. (2020).** Data Preparation for Machine Learning: Data Cleaning, Feature Selection, and Data Transforms in Python, 1.2.
31. **Tewari**, **S. (2021).** Smallest baby at birth' home after 13 months in hospital. BBC News. https://www.bbc.com/news/worldasia58141 7 5 6.
32. **Contributors to Wikimedia projects (2010)**. Dzhambulat Khatokhov. Wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/Dzhambulat_Khatokhov
33. **Repositorio datos anemia (2022**). Repositorio de datos de anemia en niños menores de 5 años. GitHub. https://github.com/JuanJoseTJ29/RepositorioDatosAnemia.

Article received on 15/07/2022; accepted on 06/11/2022.
Corresponding author is Hugo D. Calderon-Vilca.