Revista de investigación de sistemas e informática
16(2): 15-24 (2023)
DOI: https://doi.org/10.15381/risi.v16i2.25776

ISSN: 1815-0268 (Impreso) / ISSN: 1816-3823 (Electrónico)
Facultad de Ingeniería de Sistemas e Informática - UNMSM

Análisis de patrones de morbilidad por anemia mediante algoritmos no supervisados: un enfoque basado en datos de establecimientos de salud a nivel nacional

Analysis of anemia morbidity patterns using unsupervised algorithms: an approach based on data from national health facilities

Maicol Jainor Ramos Salinas
https://orcid.org/0009-0005-2018-4956
maicol.ramos@unmsm.edu.pe

Fernando Miguel Villegas Pancca
https://orcid.org/0009-0007-0869-5707
fernando.villegas@unmsm.edu.pe

Billy Bruce Cordova Chipa
https://orcid.org/0009-0006-0266-7345
billy.cordova@unmsm.edu.pe

Sebastian Pedro Cano Quito
https://orcid.org/0009-0009-2558-1810
sebastian.cano@unmsm.edu.pe

Pedro Martin Lezama Gonzales
https://orcid.org/0000-0001-9693-0138
plezamag@unmsm.edu.pe

Universidad Nacional Mayor de San Marcos, Lima, Perú

Recibido: 30/07/2023 - Aceptado: 10/09/2023 - Publicado: 30/12/2023

RESUMEN

La anemia es un importante desafío de salud pública en Lima, Perú, especialmente entre las poblaciones vulnerables. La aplicación de algoritmos de minería de datos y análisis de patrones ofrece una nueva perspectiva para abordar este problema. Al aprovechar grandes conjuntos de datos, la minería de datos permite descubrir patrones y correlaciones ocultos. Combinando estos hallazgos con algoritmos de análisis de patrones, es posible desarrollar modelos que identifiquen patrones de morbilidad relacionados con la anemia y factores de riesgo clave. Esto permite a los profesionales de la salud tomar medidas preventivas y brindar intervenciones tempranas a quienes corren mayor riesgo. Al anticipar la morbilidad por anemia se pueden implementar estrategias preventivas más efectivas, logrando una disminución de esta enfermedad y brindando a los peruanos una mayor calidad de salud.

Palabras clave: Anemia, Minería de datos, Clustering.

ABSTRACT

Anemia is a significant public health challenge in Lima, Peru, especially among vulnerable populations. The application of data mining algorithms and pattern analysis offers a new perspective to address this issue. By leveraging large datasets, data mining enables the discovery of hidden patterns and correlations. By combining these findings with pattern analysis algorithms, it is possible to develop models that identify patterns of morbidity related to anemia and key risk factors. This enables healthcare professionals to take preventive measures and provide early interventions to those at higher risk. By anticipating anemia morbidity, more effective preventive strategies can be implemented, achieving a decrease in this disease and giving people of Peru a higher quality of health.

Keywords: Anemia, Data mining, Clustering.

I. INTRODUCCIÓN

La anemia es una condición de salud ampliamente reconocida que perjudica a millones de personas alrededor del mundo, especialmente a aquellas que se encuentran en situaciones de vulnerabilidad y con acceso limitado a una atención médica adecuada. Esta condición se caracteriza por una disminución en la cantidad de hemoglobina en la sangre, lo que da como resultado negativo, el hecho de que los tejidos del cuerpo puedan recibir el oxígeno que regularmente reciben. La anemia puede tener diversas causas, como deficiencias nutricionales, enfermedades crónicas, trastornos hereditarios, embarazo y pérdida de sangre.

El diagnóstico temprano y la intervención adecuada son fundamentales para prevenir y tratar la anemia, especialmente en poblaciones vulnerables. No obstante, la tarea de detectar los elementos que aumentan el riesgo y llevar a cabo estrategias preventivas eficaces puede ser complicada debido a la complejidad y la interacción de diversas variables. En este escenario, el uso de algoritmos de minería de datos puede resultar una herramienta valiosa para realizar un análisis y poder encontrar patrones en la morbilidad de la anemia y como está relacionado a diversos factores como el sexo, edad, ubicación geográfica.

Lo que hacemos llamar minería de datos, es finalmente solo un área de investigación que se enfoca en emplear técnicas y algoritmos que tiene como objetivo identificar similitudes, conexiones entre otros patrones dentro de cierta información que se brinda para el análisis, logrando que su aplicación sea numerosa en muchos ámbitos. Centrándonos en el problema de la anemia, la utilización de la minería de datos puede ser de especial apoyo en el análisis y comprensión de las diversas variables, como características demográficas, historiales médicos, factores de riesgo y resultados de pruebas de laboratorio. Esto puede agilizar la identificación de perfiles de riesgo particulares y prever la aparición de la anemia en poblaciones vulnerables.

El propósito de este estudio consiste en examinar el empleo de algoritmos de minería de datos con el fin de poder encontrar patrones que están fuertemente relacionados a diferentes factores de morbilidad de la anemia según datos recogidos en establecimientos de salud a nivel nacional.

Asimismo, la aplicación del concepto de minería de datos para el análisis de información puede proporcionar una comprensión más completa de las causas, aspectos y consecuencias entre otros puntos asociados a la morbilidad de la anemia. Esto a su vez puede contribuir a la formulación de políticas de salud más eficaces y a la implementación de estrategias preventivas y de tratamiento adaptadas a las necesidades específicas de cada grupo poblacional.

II. PROBLEMÁTICA

La anemia es una condición de salud que actualmente afecta a muchas personas especialmente niños, mujeres embarazadas, personas de bajos ingresos y aquellas que han introducido la atención médica y alimentaria adecuada.

La anemia puede causar graves consecuencias para la salud, como la fatiga, la debilidad, la concentración difícil, especialmente el reconocimiento de las dificultades de desarrollo cognitivo y físico de los niños, como una disminución en la llegada del oxígeno a las partes del cuerpo. Tengo. Además, la anemia puede aumentar el riesgo de complicaciones durante el embarazo y el parto y afectar la productividad y la calidad de las personas afectadas.

Este problema es la desigualdad económica social, la falta de acceso a una dieta nutricional bien balanceada, falta de información, conciencia de importancia dietética saludable, restricciones al acceso a la salud, la salud y los servicios médicos.

Corregir este problema es importante para mejorar la salud del grupo vulnerable de Lima y la presencia de pozos. Es necesario identificar factores de peligro específicos y desarrollar estrategias efectivas para tener una respuesta preventiva y una detección que ayude a poder hacer que los efectos de la anemia sean mínimos. En este sentido, la aplicación de algoritmos de minería de datos y análisis predictivo puede desempeñar un papel importante en permitir patrones y factores de riesgo. Además, el desarrollo de modelos predictivos puede mejorar la toma de decisiones de salud y maximizar los recursos disponibles. Gracias a esto, podremos tener una reducción de la cantidad de esta enfermedad y lograr una mejora en el aspecto de la salud en la población.

ÁRBOL DE PROBLEMAS (ver Figura 1)

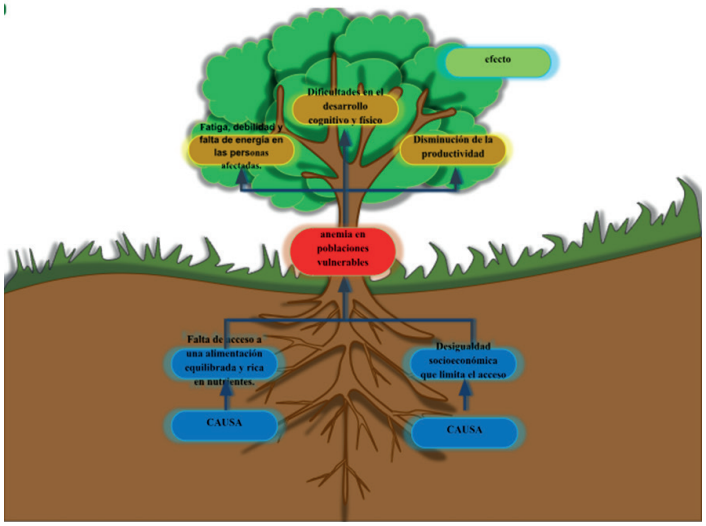
III. ANTECEDENTES

A continuación, se exponen cuáles son los antecedentes de la investigación.

La anemia en niños peruanos representa un importante desafío en términos de salud pública, lo cual

Figura 1

Árbol de problemas.



Fuente: Elaboración propia

demanda enfoques eficaces para su detección y clasificación. En este contexto, existe un estudio [1] donde el autor examina el empleo del algoritmo del bosque aleatorio como herramienta para desarrollar un modelo de clasificación preciso y confiable, capaz de identificar la anemia en niños peruanos. Este estudio presenta una perspectiva promisoria en cuanto a la aplicación de diversos algoritmos de Machine Learning para la detección temprana y la gestión efectiva de la anemia en la población infantil del país. No obstante, se requiere una comprensión más amplia de qué factores pueden llegar a influir, así como indicadores sobre la anemia en niños peruanos, lo cual constituye un incentivo para la presente investigación.

La detección precisa y el diagnóstico adecuado de la anemia en niños peruanos son de vital importancia para abordar de manera efectiva este problema. Teniendo este enfoque, existe un estudio [2] que ha proporcionado valiosa información sobre enfoques específicos de diagnóstico para esta enfermedad. Los autores exploraron el uso de los árboles de decisión como una herramienta para desarrollar un modelo de diagnóstico efectivo para la anemia en niños en la ciudad de Arequipa. Este estudio se centró en un contexto geográfico particular y examinó los factores de riesgo e indicadores relacionados con la anemia en esta población específica. Los hallazgos obtenidos subrayan la relevancia de los árboles de decisión como una técnica prometedora para una identificación preventiva y el respectivo manejo de cuidado sobre la anemia en niños de Perú. Sin embargo, aún existen áreas de investigación adicionales que deben explorarse para comprender más a fondo los factores de riesgo y los indicadores relacionados con la anemia en esta población.

La literatura científica ha explorado el estudio de la anemia empleando enfoques basados en técnicas de clasificación para llevar a cabo un análisis estadístico detallado. En esta línea, se halló un estudio [3] donde los autores exploraron el empleo de diversas técnicas de clasificación para analizar estadísticamente la anemia, presentando un enfoque innovador en cuanto al diagnóstico y la predicción de esta enfermedad. El estudio se fundamentó en la implementación de dichas técnicas para su aplicación y los resultados obtenidos permitieron identificar factores de riesgo e indicadores clave asociados a la anemia. Estos hallazgos proporcionan una perspectiva valiosa para mejorar la comprensión de la anemia y desarrollar enfoques más efectivos en su diagnóstico y manejo. Sin embargo, es necesario realizar investigaciones adicionales para explorar la aplicabilidad y generalización de estas técnicas de clasificación en poblaciones específicas, como los niños peruanos afectados por la anemia.

También tenemos otro estudio [4] donde se centraron en la creación de un enfoque de bajo costo para la detección de la anemia, utilizando una Red Neuronal Artificial (RNA). Los resultados obtenidos respaldan la viabilidad de utilizar una RNA como herramienta de detección de anemia, lo que podría mejorar la accesibilidad y la efectividad de los mé-

todos de diagnóstico, especialmente en áreas con recursos limitados. No obstante, se requieren más investigaciones y validaciones para evaluar completamente la precisión y el rendimiento de este enfoque en diferentes poblaciones y contextos clínicos.

Otro estudio [5] desarrolló un modelo de aprendizaje automático. El objetivo de este estudio fue utilizar técnicas de aprendizaje automático para estimar los niveles de hemoglobina y clasificar la presencia de anemia. El enfoque se centró en la aplicación de este modelo como una herramienta para proporcionar estimaciones precisas de los niveles de hemoglobina y para clasificar la anemia. Los resultados obtenidos demostraron la viabilidad y eficacia de este modelo de aprendizaje automático en la estimación de la hemoglobina y la clasificación de la anemia. Estos hallazgos tienen el potencial de mejorar la detección temprana y el manejo de la anemia, especialmente en entornos con recursos limitados. No obstante, se requieren investigaciones adicionales y validaciones para evaluar y mejorar la precisión y aplicabilidad de este modelo en diferentes poblaciones y configuraciones clínicas.

También se encontró un estudio [6] donde usando el método de análisis de grupos, se investigó la relación entre deficiencias nutricionales y anemia. Se recopilaron datos sobre la ingesta de nutrientes y la hemoglobina de 4,762 estudiantes que asistían a escuelas públicas en Cisjordania. Se dividieron los niveles de hemoglobina en dos grupos mediante el análisis de grupos K-means. Las ingestas de folato, hierro y vitamina B12 se utilizaron como indicadores de la ingesta de nutrientes relacionada con la anemia.

Por último, tenemos un estudio [7], en el cual los autores aplicaron dos modelos híbridos utilizando algoritmos genéticos y algoritmos de aprendizaje profundo para predecir diferentes tipos de anemia, sin embargo, una de las conclusiones a las que llegaron fue de que cada algoritmo no funciona bien para todos los conjuntos de datos, por lo que es necesario desarrollar nuevas técnicas.

IV. MARCO TEÓRICO

A. CLASIFICACIÓN DE LA ANEMIA SEGÚN CIE-10

La CIE-10, se define a sí misma cómo "...una clasificación estadística de enfermedades y otros problemas de salud, para satisfacer una amplia gama de necesidades de recopilación de datos de mortalidad y de asistencia sanitaria…" [8].

Según la clasificación de la CIE-10 hemos podido identificar los siguientes grupos para los diferentes tipos de anemias.

Tabla 1
Identificadores para los grupos de anemia según CIE-10

Grupos (en intervalos)	Descripción
D50-D53	Anemias nutricionales
D55-D59	Anemias hemolíticas
D60-D64	Anemias aplásicas y otras anemias

B. ALGORITMOS NO SUPERVISADOS

Estos algoritmos [9] se utilizan principalmente para mejorar los resultados de métodos de agrupamiento y no requieren información de etiqueta de clases, estos algoritmos descubren patrones ocultos o agrupaciones de datos que pueden ser útiles para la categorización, se pueden [10] distinguir dos grupos principales para estos algoritmos: métodos jerárquicos y particionales.

C. ANÁLISIS CLÚSTER O CLUSTERING

Según nos muestra [11], una de las técnicas que son muy utilizadas para un contexto de múltiples variables, es el de clusters, esto ya que permite agrupar grandes conjuntos de datos que pueden llegar a ser dificultosos de tratar, y poder tornarse en subconjuntos más pequeños o los llamados clusters, todos los que componen cada cluster, tienen aspectos en común.

Existe un trabajo [12] donde se identifican tres categorías principales para las adaptaciones/variantes:

Clustering de particionamiento: Estos algoritmos requieren que el usuario indique de antemano el número de grupos a utilizar.

Clustering jerárquico: Estos algoritmos prescinden de que el usuario defina previamente el número de grupos (como el clustering aglomerativo, el clustering divisivo). También existen métodos que combinan o modifican los enfoques anteriores.

D. NÚMERO ÓPTIMO DE CLUSTERS

Determinar el número óptimo de clusters es uno de los desafíos más complejos al tener que hacer uso de los métodos de agrupamiento. No existe un enfoque único para determinar el número adecuado de grupos, ya que es un proceso altamente subjetivo

que depende del algoritmo utilizado y de la información previa disponible sobre los datos en estudio. No obstante, según [12], existen tres técnicas que pueden ayudarnos a seleccionar el número óptimo de grupos. En el presente trabajo de investigación, utilizaremos las siguientes técnicas:

1. **Método Elbow**

   Este método [13], lo que hace es analizar la diferencia en la suma de errores cuadrados (SSE) de cada cluster, de esta forma compara la consistencia del mejor número de clusters de manera visual, gracias a esto, podemos decir un número óptimos de grupos basado en el punto donde el ángulo del gráfico de codo sufre un cambio brusco.

2. **Método Average Silhouette**

   Este método [14], es una forma de determinar la validez de los clusters sin utilizar información externa. Dos de los principales datos utilizados para determinar el coeficiente de este método son la *cohesion measure* y el *cluster separation*. La *cohesion measure* mide qué tan cercanos están los datos entre sí dentro del mismo cluster, mientras que el *cluster separation* mide qué tan separados están cada cluster de los demás clusters.

E. K-MEANS

De acuerdo con [15], se trata de un algoritmo de clustering restringido que requiere el número de clusters como parámetro y está diseñado para datos continuos, lo que implica que sólo puede operar con objetos descritos por un conjunto de atributos numéricos. Este algoritmo minimiza una función objetivo mientras calcula iterativamente los centros de agrupamiento. Entre las principales ventajas y desventajas de aplicar este algoritmo [12], [13] tenemos las siguientes:

Principales desventajas:

- Al aplicar este algoritmo es necesario especificar previamente el valor K (cantidad de clusters a utilizar), lo que, si es que no se cuenta con información adicional sobre los datos, puede representar un gran desafío
- Posee problemas con la robustez frente a las excepciones, siendo la única solución eliminarlos o utilizar métodos de clasificación más robustos como K-medoids.

Principales ventajas:

- Fácil de programar, entender e implementar.
- Posee una gran escalabilidad a grandes conjuntos de datos, con la capacidad de adaptarse con facilidad a los nuevos ejemplos de datos.

F. GMM (GAUSSIAN MIXTURE MODELS)

El modelo de mezcla gaussiana (GMM), conocido como Gaussian Mixture Models en inglés, ha sido estudiado por [12]. El GMM puede considerarse como una extensión del algoritmo K-means, donde se asigna una probabilidad de pertenencia a cada miembro de los clusters, siguiendo una distribución de probabilidades. Esta característica permite una mayor flexibilidad en la asignación de observaciones a grupos.

En relación con las ventajas y desventajas de aplicar este algoritmo, los estudios de [12], [16], [17] han identificado algunos aspectos significativos. Entre las ventajas, se destaca la capacidad del GMM para modelar distribuciones de datos complejas y la posibilidad de identificar agrupamientos no lineales. Además, el GMM ofrece una mayor flexibilidad al permitir que una observación pertenezca a varios clusters con diferentes probabilidades.

Por otro lado, entre las desventajas se encuentra la mayor complejidad computacional del GMM en comparación con otros algoritmos de clustering. Además, la elección adecuada del número óptimo de clusters sigue siendo un desafío, y la interpretación de los resultados puede ser más compleja debido a la naturaleza probabilística del modelo.

Principales desventajas:

- Necesita especificar previamente el número de grupos que se crearán, al igual que en K-means.
- El entrenamiento de GMM puede ser computacionalmente costoso, especialmente cuando se trabaja con grandes conjuntos de datos o con muchas características.

Principales ventajas:

- A diferencia de otros algoritmos de agrupamiento, GMM tiene la capacidad de capturar agrupamientos de formas más complejas y adaptables.

- En comparación con otros algoritmos de agrupamiento basados en centroides, como K-means, GMM es menos sensible a valores atípicos, debido a ello las excepciones suelen tener un efecto menor en los resultados finales porque GMM utiliza una estimación de densidad.

V. METODOLOGÍA DE TRABAJO

Para el presente trabajo se tiene pensado aplicar la siguiente metodología para trabajar con la data a analizar:

Recopilación de datos: Inicialmente, se llevará a cabo la extracción y recolección de la data relevantes para iniciar con la evaluación de la morbilidad de la anemia en el Perú, en este caso se tomarán los datos recopilados por el Ministerio de Salud del Perú [18] con respecto a la morbilidad de la anemia en el desde el año 2021. Estos datos incluyen información de los fallecidos tales como: edad, tipo de edad, género, fecha de atención, diagnóstico (Código CIE-10), tipo de diagnóstico, identificador de ubigeo, e identificador del establecimiento de salud donde fue diagnosticado.

Preprocesamiento de datos: Una vez recopilados los datos, se realizará un proceso de preprocesamiento para limpiar y organizar la información. Esto implica la identificación y manejo de valores faltantes, la eliminación de datos duplicados o inconsistentes, y la normalización de las variables para asegurar la calidad y coherencia de nuestra información, además se realizará una evaluación a detalle de los atributos presentes en el dataset, evaluando su relevancia y, en caso necesario, realizando transformaciones o extracciones adicionales.

Calcular el valor óptimo de K: En esta etapa usaremos dos métodos que nos ayudarán a poder elegir el valor de K (número de clusters) que usaremos más adelante como parámetro en los algoritmos. Estos dos métodos serán el método Elbow y el método Average Silhouette.

Aplicación de los algoritmos: En esta etapa, sabiendo ya el valor que usaremos para K, lo que se hará es aplicar dos de los algoritmos de minería de datos para la formación de clusters no etiquetados, en este caso usaremos K-Means y GMM.

Validación de los algoritmos: Algunas métricas comunes para la evaluación de los modelos aplicados incluyen el coeficiente de Silhouette, que mide la coherencia interna de los clústeres y la separación entre clústeres. Además, como método adicional de validación, se podría utilizar la validación cruzada y los criterios de información de Akaike (AIC) y bayesiano (BIC), tal como menciona [19] (ver Figura 2).

VI. DESARROLLO

A. RECOPILACIÓN DE DATOS

Como se indicó anteriormente, el dataset utilizado pertenece al Ministerio de Salud del Perú, para un mejor entendimiento del dataset y su posterior uso se deja en evidencia el diccionario de datos.

Figura 2

Metodología de trabajo a aplicar.

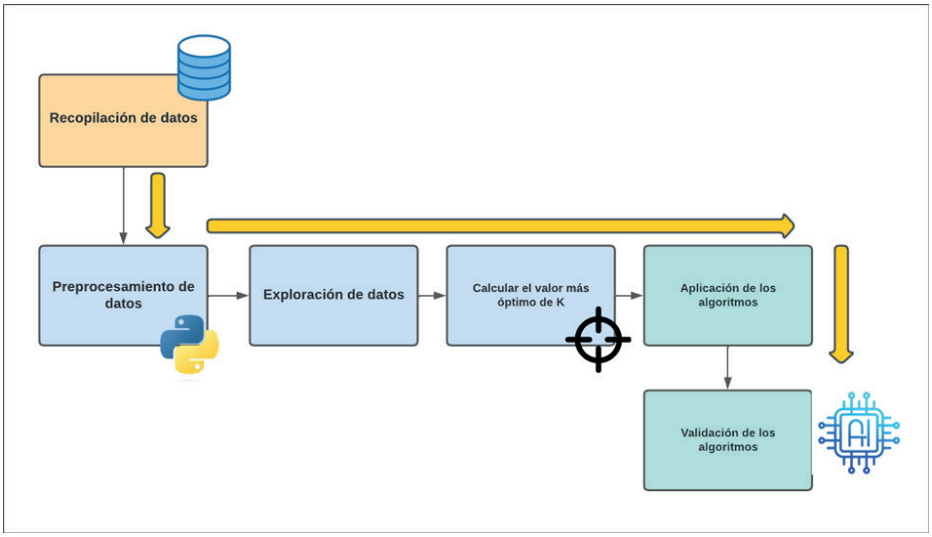


Tabla 2

Diccionario de datos para el dataset utilizado

Campo	Descripción	Tipo de dato
id_persona	Identificador de persona	int
Edad	Edad del paciente	int
Tipo_edad	Tipo de edad: AÑO, MES, DIA, HORA, MIN, SEG, SEM	char
Sexo	Sexo	varchar
id_ubigeo	Identificador ubigeo	int
Fecha_atencion	Fecha de atención	date
Diagnóstico	Código CIE-10	varchar
Tipo_Dx	Tipo de diagnóstico: Definitivo, Presuntivo, Repetido	varchar
id_eess	Identificador establecimiento	int

B. PREPROCESAMIENTO DE DATOS

El objetivo principal del preprocesamiento en nuestro dataset de anemia es transformar y limpiar la información existente, esto implica realizar una serie de pasos y técnicas que nos permitan abordar problemas comunes, como datos faltantes, valores atípicos o inconsistencias en la estructura de los datos, además se realizará un análisis exhaustivo de los atributos presentes en el dataset, evaluando su relevancia y, en caso necesario, realizando transformaciones o extracciones adicionales.

Imputación de valores nulos: Luego de realizar técnicas para hallar valores nulos dentro de algunas columnas, pudimos visualizar que la columna para el identificador de ubigeo contenía gran cantidad de valores nulos, es por ello que decidimos imputar aquellos valores con la moda.

Codificación de variables categóricas: Para este paso lo que se realizo fue codificar la columna “Sexo”, “Tipo_edad” y “Diagnostico”, realizando las transformaciones necesarias.

Eliminación de las columnas constantes y de no interés: Para este paso lo que se realizará será la eliminación de columnas constantes, es decir, aquellas que tienen un mismo valor para todas sus filas y también se eliminan las columnas que no aporten a nuestro trabajo de investigación, siendo para estos casos las columnas “Tipo_Dx”, “Fecha_atención”, “id_persona” y “id_eess”.

Exportación de la data preprocesada: Este es el paso final del preprocesamiento, lo que se realizó fue simplemente exportar en un archivo .csv la nueva data para su posterior análisis y uso (ver Figura 3).

C. CALCULAR EL VALOR ÓPTIMO DE K

A continuación, se realizarán los dos métodos anteriormente expuestos para hallar el mejor valor para K para la aplicación de los algoritmos K-Means y GMM.

- Método Elbow (ver Figura 4).
- Método Average Silhouette (ver Figura 5).

Siguiendo la información proporcionada por Ricardo Moya [20] se puede observar que el valor óptimo para K en ambos métodos, siendo K el número de agrupamientos que se darán, es el valor de 5.

D. APLICACIÓN DE LOS ALGORITMOS

Para la aplicación de estos algoritmos se seleccionaron las siguientes columnas: “Sexo”, “Diagnostico” y “Edad_en_anios” para luego normalizar los datos.

- K-MEANS

  Se aplicó el algoritmo K-means con 5 clusters a los datos normalizados. Las etiquetas de cluster resultantes se asignaron a cada muestra y se agregaron al conjunto de

Figura 3

Flujo Pre Procesamiento.

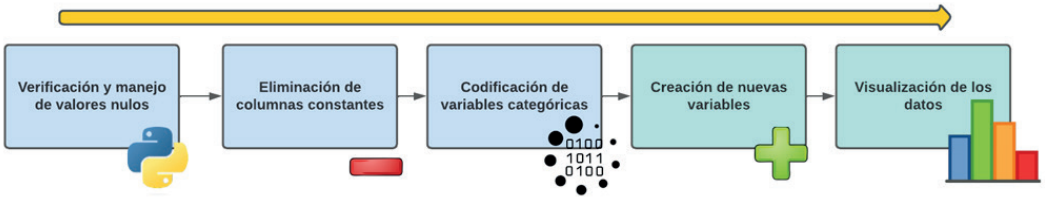


Figura 4

Análisis de K con el Método Elbow.

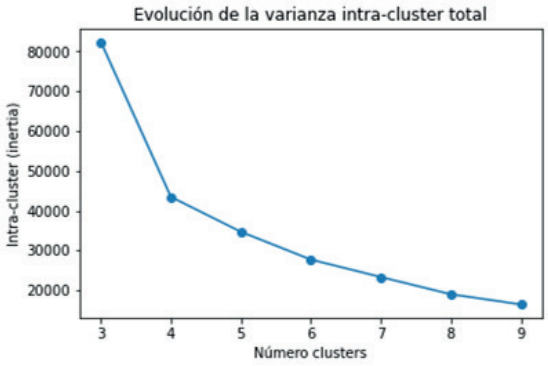
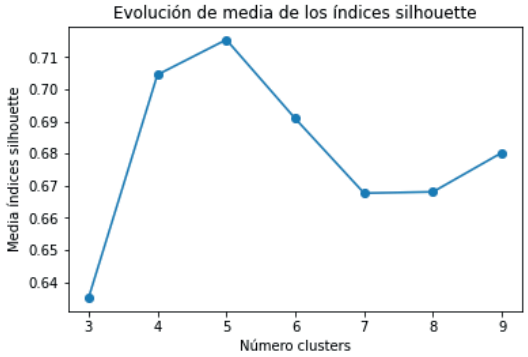


Figura 5

Análisis de K con el Método Average Silhouette.



datos, permitiendo clasificar las muestras en clusters basados en aspectos en común.

Tabla 3

Conteo de muestras por cluster generado en K-Means

Cluster	Cantidad de muestras
0	31782
1	33677
2	3225
3	4094
4	5895

- GMM (GAUSSIAN MIXTURE MODELS)

  Se ajustó la instancia del modelo GMM a los datos para luego calcular las probabilidades de pertenencia de cada grupo para todas las muestras, según esta probabilidad se asigna cada muestra al grupo con la mayor probabilidad.

Tabla 4

Conteo de muestras por cluster generado en GMM

Cluster	Cantidad de muestras
0	35335
1	39096
2	859
3	1946
4	1437

E. VALIDACIÓN DE LOS ALGORITMOS

Una forma de poder validar los algoritmos aplicados es mediante el coeficiente de Silhouette. Se debe hacer el cálculo de esta métrica al nuevo dataset generado, el cual tendrá una nueva columna llamada “Cluster”, el cual su valor será el número de cluster donde pertenece esa muestra.

Al hacer el cálculo pertinente se pudo corroborar que el valor corresponde al que se tenía predicho

antes de la implementación del algoritmo mediante el Método Average Silhouette, con el valor de 0.715. Un valor más cercano a 1 indica que la muestra está bien asignada a su cluster y está alejada de los otros clusters.

VII. RESULTADOS Y DISCUSIONES

Se puede observar que para cada cluster generado por cada algoritmo con el valor 5 para K se obtuvo una cantidad muy distinta de muestras, luego de realizar las interpretaciones sobre los clusters generados se pudo llegar a la conclusión de que existe una tendencia muy marcada para cada agrupamiento. Además, se halló que la característica más distintiva para cada cluster es la edad media.

Al momento de visualizar las medias de las características para cada cluster tanto en el caso con K-Means como con GMM, se pudo observar una alta similitud en el valor de la media la mayoría de las características de cada cluster. Además, se pudo encontrar una relación entre el diagnóstico (Código CIE-10) con las demás características para el conjunto de datos.

VIII. CONCLUSIONES

Se obtuvieron resultados interesantes al analizar los patrones de morbilidad por anemia utilizando algoritmos no supervisados con datos de establecimientos de salud a nivel nacional. A pesar de utilizar el mismo valor de k (5), cada algoritmo generó grupos con una cantidad de muestras bastante diferente. Esto indica que cada algoritmo tenía su propia gama de habilidades para agrupar muestras de manera efectiva.

Al realizar las interpretaciones de los grupos generados, se llegó a la conclusión de que cada grupo presenta una tendencia evidente. Además, se descubrió que en cada grupo, la edad media fue la característica más distintiva. Esto demuestra que la edad puede ser un factor significativo en la prevalencia de anemia y puede distinguir claramente entre los grupos identificados.

Tanto en K-Means como en GMM, se observó una alta similitud en el valor de media para la mayoría de las características de cada cluster al ver las medias de las características para cada cluster. Esto indica que las características analizadas pueden no ser lo suficientemente discriminativas para diferenciar claramente los grupos generados. Sin embargo, se notó que se encontró una correlación entre las características del conjunto de datos adicionales y el diagnóstico de anemia basado en el Código CIE-10. Esto indica que otras características pueden afectar el diagnóstico, lo que puede ser relevante para comprender y abordar la morbilidad por anemia.

REFERENCIAS BIBLIOGRÁFICAS

[1] B. C. Panduro, “Aplicación del algoritmo del bosque aleatorio a un modelo de clasificación de la anemia en niños peruanos”, Mediciego, vol. 28, núm. 1, p. 3471, 2022.

[2] 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.

[3] Meena, K., Tayal, D. K., Gupta, V., & Fatima, A. (2019). Using classification techniques for statistical analysis of Anemia. Artificial Intelligence in Medicine, 94, 138–152.

[4] A. Ghosh, J. Mukherjee, y N. Chakravorty, “A low-cost test for anemia using an Artificial Neural Network”, Comput. Methods Programs Biomed., vol. 229, núm. 107251, p. 107251, 2023.

[5] El-Kenawy, E. M. T. (2019). A Machine Learning Model for Hemoglobin Estimation and Anemia Classification. International Journal of Computer Science and Information Security (IJCSIS), 17(2).

[6] R. Qasrawi y D. Abu Al-Halawa, “Cluster analysis and classification model of nutritional anemia associated risk factors among Palestinian schoolchildren, 2014”, Front. Nutr., vol. 9, p. 838937, 2022.

[7] S. Kilicarslan, M. Celik, y Ş. Sahin, “Hybrid models based on genetic algorithm and deep learning algorithms for nutritional Anemia disease classification”, Biomed. Signal Process. Control, vol. 63, núm. 102231, p. 102231, 2021.

[8] World Health Organization. (2004). ICD-10: International Statistical Classification of Diseases and related health problems : tenth revision, 2nd ed. World Health Organization. https://apps.who.int/iris/bitstream/handle/10665/42980/9241546530_eng.pdf?sequence=1&isAllowed=y

[9] Pérez Verona, I. C., & Arco García, L. (2016). Una revisión sobre aprendizaje no supervisado de métricas de distancia. Revista Cubana de Ciencias Informáticas, 10(4), 43-67.

[10] J. F. V. Rueda, “Aprendizaje supervisado y no supervisado”, healthdataminer.com, 04-ago-2019. [En línea]. Disponible en: https://healthdataminer.com/data-mining/aprendizaje-supervisado-y-no-supervisado/ [Consultado: 07-jul-2023].

[11] Castro Heredia, L. M., Carvajal Escobar, Y., & Ávila Díaz, Á. J. (2012). ANÁLISIS CLÚSTER COMO TÉCNICA DE ANÁLISIS EXPLORATORIO DE REGISTROS MÚLTIPLES EN DATOS METEOROLÓGICOS. Ingeniería de Recursos Naturales y del Ambiente, (11), 11-20.

[12] J. Amat. (2020, Diciembre). Clustering con Python. Cienciadedatos.net. [En línea]. Disponible en : https://www.cienciadedatos.net/documentos/py20-clustering-con-python

[13] E. Umargono, J. E. Suseno, y S. K. Vincensius Gunawan, “K-means clustering optimization using the elbow method and early centroid determination based on mean and median formula”, en Proceedings of the 2nd International Seminar on Science and Technology (ISSTEC 2019), 2020, pp. 121–129.

[14] D. M. Saputra, D. Saputra, y L. D. Oswari, “Effect of distance metrics in determining K-value in K-means clustering using elbow and silhouette method”, en Proceedings of the Sriwijaya International Conference on Information Technology and Its Applications (SICONIAN 2019), 2020, pp. 341–346.

[15] S. López, Algoritmos de Agrupamiento Global para Datos Mezclados, Instituto Nacional de Astrofísica, Óptica y Electrónica, Puebla, 2007.

[16] R. Sridharan, “Gaussian mixture models and the EM algorithm”, Mit.edu. [En línea]. Disponible en: https://people.csail.mit.edu/rameshvs/content/gmm-em.pdf. [Consultado: 08-jul-2023].

[17] M. de Salud, “Morbilidad: Anemia”, Gob.pe. [En línea]. Disponible en: https://www.datosabiertos.gob.pe/dataset/morbilidad-anemia. [Consultado: 07-jul-2023].

[18] J. VanderPlas, “In depth: Gaussian mixture models”, Github.io. [En línea]. Disponible en: https://jakevdp.github.io/PythonDataScienceHandbook/05.12-gaussian-mixtures.html. [Consultado: 08-jul-2023].

[19] R.Moya, “Selección del número óptimo de Clusters”, Jarroba, 12-sep-2016. [En línea]. Disponible en: https://jarroba.com/seleccion-del-numero-optimo-clusters/. [Consultado: 08-jul-2023].

Financiamiento:

Propio.

Conflictos de interés:

Los autores declaran no tener conflictos de interés.

Contribuciones de autoría:

Todos los autores participaron en las diferentes actividades para la elaboración del articulo.