

Aplicación de algoritmos de Machine Learning para el diagnóstico de anemia

Carlos Figueroa
Cañedo,
Brian Bolaños Terán,
Ricardo Sáenz
Gardea y
Mariana Olvera
Armenta

Universidad
Autónoma de
Chihuahua, Facultad
de Medicina y
Ciencias Biomédicas.

Recibido:
11 de octubre 2023
Aceptado:
26 de octubre de
2023

RESUMEN

La anemia es la enfermedad hematológica más común y que se diagnostica con mayor frecuencia, por eso es importante tener un diagnóstico oportuno. Con ayuda de inteligencia computacional, nos es posible crear herramientas basadas en diferentes algoritmos que sirvan de soporte para detectar si el paciente tiene o no tiene esta enfermedad.

Se utilizaron tres algoritmos para la clasificación de muestras, los cuales fueron KNN, Bayesiano y una Red Neuronal, que se tuvo que someter a un entrenamiento para lograr el mejor resultado. El presente trabajo está creado con ayuda de una base de datos que cuenta con 5 características que ayudan a determinar el estado del paciente y dos diferentes clases que indican si el paciente tiene o no anemia.

ABSTRACT

Anemia is the most common and most frequently diagnosed hematological disease, so it is important to have a timely diagnosis. With the help of computational intelligence, it is possible to create tools based on different algorithms that serve us to detect whether the patient has or does not have this disease. Three algorithms were used for the classification of samples, which were KNN, Bayesian, and a Neural Network, which had to undergo training to achieve the best result. This paper is created with the help of a database that has 5 characteristics that help us determine the patient's status and two different classes that indicate whether the patient has anemia or not.

Palabras clave: Clasificación, Detección, Inteligencia artificial.

INTRODUCCIÓN

Desde la antigüedad, el ser humano ha desarrollado técnicas que logren facilitar el realizar cada actividad. Con el avance de la tecnología y su aplicación dentro de diferentes áreas, su implementación en temas relacionados con la salud ha mostrado una enorme ventana de oportunidades con la aplicación de diferentes métodos de *machine learning*, esto se define como la variedad de algoritmos que procesan conjuntos de datos por

medio de la habilidad de aprendizaje de los mismos o experiencias anteriores para posteriormente clasificar o predecir la existencia de patrones en conjuntos de datos nuevos que analiza (Rathore & Mannepilli, 2021); por la naturaleza de clasificación de estos algoritmos, los convierte en herramientas útiles en el diagnóstico de enfermedades y patologías.

Los sistemas de diagnóstico de varias enfermedades tenían una evaluación basada en diferentes características, principalmente en parámetros o rangos promedios, en donde los datos existentes facilitan el encontrar características que determinen si alguien tiene, o no, una enfermedad; Según el aprendizaje o acomodo de datos en programas podemos obtener respuestas deseadas, como los diagnósticos claros y precisos.

La anemia es una de las enfermedades más comunes y es el trastorno hematológico que se diagnostica con mayor frecuencia y afecta a la cuarta parte de la población mundial. Se define como una disminución de la cantidad de hemoglobina en sangre o como una disminución de la capacidad de la misma para transportar oxígeno. Es una enfermedad difícil de diagnosticar debido a que los síntomas pueden llegar a ser indefinidos (Shalik & Subashini, 2017). Existen diferentes tipos de anemia, las cuales varían según sus orígenes, por ejemplo, la anemia por deficiencia de vitamina B12 o la anemia hipocrómica. Para la detección de este trastorno se pueden utilizar muchas características que nos ayudarían a tener un diagnóstico correcto, en el caso de este trabajo se utilizó el nivel de hemoglobina y la hemoglobina corpuscular media (MCH). La MCH ayuda clínicamente a evaluar el color de la sangre.

METODOLOGÍA

A. Descripción de bases de datos.

En el desarrollo de este proyecto se basó la información de una base de datos conformada con información de diferentes niveles sanguíneos (que definimos como características), utilizados en el diagnóstico de anemia, de pacientes anémicos y no anémicos: la información de cada paciente que conforma parte de la base de datos, le denominaremos como muestra. La base de datos (BD), usada para este trabajo, contaba con seis columnas: cinco de características informativas que conforman cada muestra de la BD y una última columna que indica el grupo al que pertenece el paciente, Con Anemia o Sin Anemia. De esta BD se utilizaron 1240 muestras en total y dos de las 5 características: Hemoglobina y MCH, para formar la base de datos de trabajo (BDT); para discriminar características se graficaron los datos con diagramas de dispersión para apreciar la separa-

ción entre las características de los dos grupos, siendo en estas donde hay mayor diferenciación.

De las 1240 muestras de la BDT, la mitad (620) conformaron las muestras para una base de datos de pacientes anémicos (BDPA), y la otra mitad (620) las muestras para la base de datos de pacientes no anémicas (BDPNA).

Para el entrenamiento de los algoritmos implementados, BDPA y BDPNA fueron subdivididas de forma aleatoria en cuatro bases de datos; el 80% del total de muestras (496) de cada una formaron: la base de datos de entrenamiento de pacientes anémicos (BDTPA) y la base de datos de entrenamiento de pacientes no anémicos (BDTPNA); por otro lado, para verificar el funcionamiento de cada algoritmo se utilizaron el 20% restante de las muestras de las BDA y BDNA (124) los cuales conformaron: la base de datos de validación de pacientes anémicos (BDVPA) y la base de datos de validación de pacientes no anémicos (BDVPNA).

B. Algoritmo bayesiano

Esta técnica se utiliza para llevar a cabo la clasificación booleana para el conjunto de datos de entrenamiento. Este clasificador supone que ni la presencia ni la ausencia de un atributo particular esté relacionado con la presencia o ausencia de otros atributos (Dalvi & Vernekar, 2016) (García Chávez & Cañedo Figueroa, 2021).

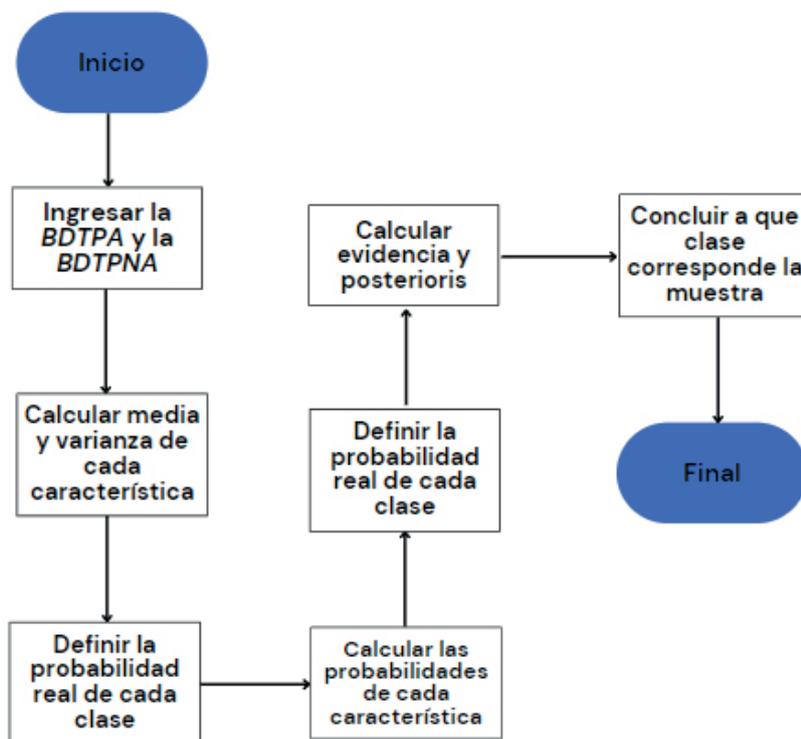


Figura 1. Diagrama de flujo algoritmo Bayesiano.

C. ANN

De las siglas en inglés *Artificial Neural Networks* (ANN), es un tipo de algoritmo o modelo computacional formado por simulaciones informáticas que asemejan la organización neuronal del cerebro humano; una ANN adquiere conocimiento aprendiendo de un conjunto de datos de entrenamiento y aplica dicho conocimiento para clasificar datos de prueba (Dalvi & Vernekar, 2016) (Drew et al., n.d.).

En este trabajo se llevó a cabo la implementación de una ANN con la BDTPA y la BDTPNA, junto al programa de Matlab, se usaron los hiperparámetros: Número de Épocas Máximas (NEM), Verificación de Mínimos Locales (VML), Error Máximo Permitido (EMP), Factor de Aprendizaje (FA).

La ANN implementada en este trabajo contó con las siguientes características: Capas de la ANN (2), Número de neuronas capa 1(9), Número de neuronas capa 2 (19), NEM (900), VML (1000), EMP (1e-30), FA (0.01)

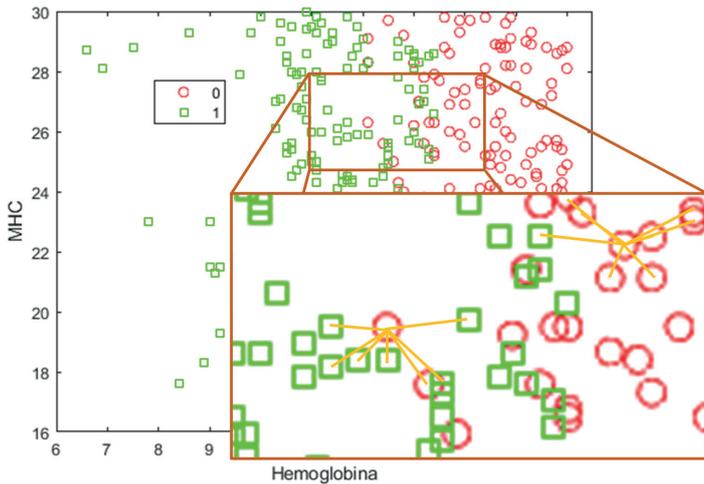
Con los hiperparámetros mencionados, se entrenó la ANN, de la cual obtuvimos una red con 98.8% de validación.

Figura 2. Matriz de confusión de la ANN.



D. KNN

El algoritmo KNN, por sus siglas en inglés *K-Nearest Neighbors*, pertenece a los algoritmos de aprendizaje supervisado. Clasifica las muestras según los votos principales del vecino más cercano. ‘K’ es el número de vecinos que votan en la clasificación (Chy & Rahaman, 2019).



En este caso se utilizaron un total de 7 vecinos, ya que con valores más grandes se requiere más gasto computacional.

Figura 3. Funcionamiento básico KNN.

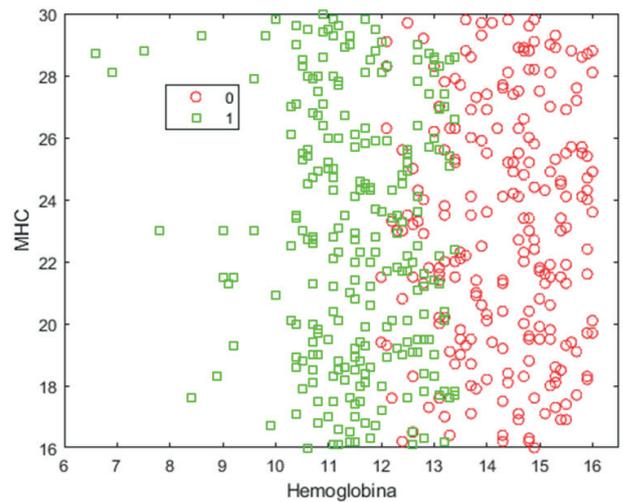


Figura 4. Representación de los datos graficados.

RESULTADOS

A. hardware y software.

Los métodos de evaluación fueron analizados en una PC con las siguientes características: Intel Core i5-4200 CPU 1.60Ghz con 4GB de RAM, con un sistema operativo Windows 10 64 bits.

B. Métricas de los algoritmos

Para esta sección del trabajo, se clasificaron los resultados de las muestras entre positivos (los que se harían referencia a muestras de pacientes con anemia) y negativos (los que se harían referencia a muestras de pacientes sin anemia). Por medio del programa de Matlab, se llevó a cabo la revisión de la funcionalidad de cada algoritmo; dependiendo la clasificación de las muestras se obtuvieron lo que describiremos como: Verdaderos Positivos (VP, cuando el algoritmo clasifique como positivo una muestra positiva), Verdaderos Negativos (VN, cuando el algoritmo clasifique como negativo una muestra negativa), Falsos Positivos (FP, cuando el algoritmo clasifique como positivo una muestra negativa) y Falsos Negativos (FN, cuando el algoritmo clasifique como negativa una muestra positiva) de cada algoritmo.

C. Operaciones.

En este trabajo, se tomaron en cuenta cuatro criterios para determinar la efectividad de cada algoritmo: precisión, exactitud, exhaustividad y el F1 score por medio de los VP, VN, FP y FN. Para la obtención de la precisión se llevó a cabo la ecuación 1:

$$\text{Precisión} = \frac{VP}{(VP + FP)} \quad (1)$$

Para la obtención de la exactitud se llevó a cabo la ecuación 2:

$$\text{Exactitud} = \frac{(VP + VN)}{(VP + FP + FN + VN)} \quad (2)$$

Por otro lado, para la obtención de la exhaustividad se llevó a cabo la ecuación 3:

$$\text{Exhaustividad} = \frac{VP}{(FN + VP)} \quad (3)$$

Finalmente, para la obtención del F1 score se llevó a cabo la ecuación 4:

$$\text{F1 score} = \frac{(2 * \text{Precisión} \cdot \text{Exhaustividad})}{(\text{Precisión} + \text{Exhaustividad})} \quad (4)$$

RESULTADOS

Ya formado los sistemas de evaluación se logró analizar la BD-VPA y BDVPNA permitiendo ver la efectividad de cada uno de los algoritmos, empezando con la ANN, donde se obtuvieron 119 VP, 5 FP, 10 FN y 114 VN; por otro lado, con el algoritmo KNN se obtuvieron los mismos resultados de VP y FP, pero con 18 FN y 106 VN; finalmente, en el caso del Bayesiano, se obtuvieron 112 VP, 12 FV, 17 FN y 107 VN.

En base a estos 3 diferentes métodos de análisis nos encontramos con diferentes resultados utilizando las ecuaciones de la sección anterior; en la Tabla 1 se visualiza que los mejores valores obtenidos son por la ANN, siguiendo con el KNN y finalmente el Bayesiano.

Con base en trabajos de otros autores, como se puede visualizar en la Tabla 1, Comparando nuestros resultados con los obtenidos en el artículo “A Comparative Analysis by KNN, SVM & ELM Classification to Detect Sickle Cell Anemia” nos pudimos dar cuenta que en cuestión al algoritmo KNN, es posible decir que este trabajo presenta mejores resultados, ya que como se muestra en la siguiente tabla los autores tienen una precisión de 90 mientras que aquí se tiene una de 95.9, la exactitud del mencionado artículo es de 73.33 y la de este trabajo es de 90. Por último, podemos comparar el F1-Score, en este artículo se obtuvo un 91.18 y los otros autores consiguieron un 81.81 (Chy & Rahaman, 2019).

Siguiendo con las comparaciones realizadas, trabajando en este caso con los datos de la Tabla 1 en la sección de “Anemia detection using ensemble learning techniques and statistical models”, a partir de la cual es posible darse una idea sobre los tres algoritmos utilizados. Podemos observar que, de manera general, el presente trabajo presenta mejores resultados con cierto desliz en la parte de la exactitud de la ANN y del bayesiano. Fuera de estos resultados mencionados, es posible decir que los obtenidos aquí son más satisfactorios. Prueba de ello son los F1 score, los cuales muestran un mejor porcentaje (Dalvi & Vernekar, 2016).

Además de los trabajos ya mencionados, se encontró con, “A Machine Learning Model for Hemoglobin Estimation and Anemia Classification”, en donde a su vez se implementaron algoritmos KNN y Bayesiano para medir niveles de hemoglobina y clasificación de anemias. En este trabajo de referencia, se utilizaron un total de once características; a pesar de ello, como se presenta en la Tabla 1, los resultados obtenidos en estos algoritmos a la hora de clasificar anemias obtuvieron en F1 score

Valores obtenidos en este trabajo				
Método	Precisión	Exactitud	Exhaustividad	F1 score
ANN	95.96	93.95	92.24	94.07
KNN	95.96	90.72	86.86	91.18
Bayes	90.32	88.3	86.82	88.53
Valores de "Anemia detection using ensemble learning techniques and statistical models"				
ANN	90.86	94.28	76.8	83.24
KNN	88.016	51.698	51.68	65.12
Bayes	88.658	92.318	71.07	78.896
Valores de "A Machine Learning Model for Hemoglobin Estimation and Anemia Classification"				
ANN	-	-	-	71.56
Bayes	-	-	-	80.16
Valores de "A Comparative Analysis by KNN, SVM & ELM Classification to Detect Sickle Cell Anemia"				
KNN	90	73.33	-	81.81
Valores de "Classifying anemia types using artificial learning methods"				
ANN	78.2	79.6	79.9	79
Bayes	88.658	92.318	71.07	62.5

Tabla 1. Comparación de resultados obtenidos.

71.5% para el KNN y 80.16% para el bayesiano, mientras que en nuestro trabajo de clasificación de pacientes con anemia y no anemia se obtuvo un F1-Score de 91.18% utilizando únicamente dos características (El-Sayed & El-kenawy, 2019).

Finalmente, con el trabajo "Classifying anemia types using artificial learning methods", se utilizaron diferentes algoritmos para compararlos entre sí para la clasificación de tipos de anemias. En este trabajo se utilizaron un total de 25 características obteniendo en los dos algoritmos que compartimos (Bayesiano y ANN) valores inferiores al 92.3%, mientras que en nuestros resultados se presentan valores por encima de esos (Yildiza & Yurtay, 2021).

CONCLUSIONES

El mejor de los métodos probados fue la ANN, mostrando los resultados más satisfactorios al analizar los datos originales y la base de prueba. Puede ser aplicado correctamente para la determinación de la anemia con una evaluación de la muestra

de sangre, teniendo como principal determinación la hemoglobina. Observando los datos comparativos, se puede reconocer una mejora promedio en el rendimiento de 17.715% en el KNN, 16.13% en la ANN y 14.68% en el bayesiano, dejando la oportunidad de implementar, el mejor algoritmo del presente trabajo, en el desarrollo de algún software de diagnóstico de algunos tipos de anemia.

REFERENCIAS

- Chy, T. S., & Rahaman, M. A. (2019). A Comparative Analysis by KNN, SVM & ELM Classification to Detect Sickle Cell Anemia. 455–459. <https://doi.org/10.1109/ICREST.2019.8644410>.
- Dalvi, P. T., & Vernekar, N. (2016). Anemia detection using ensemble learning techniques and statistical models. 1747–1751. <https://doi.org/10.1109/RTEICT.2016.7808133>
- Drew, P. J., John, R. T., & Monson, M. D. (n.d.). Artificial neural networks. <https://doi.org/10.1067/msy.2000.102173>
- El-Sayed, M., & El-kenawy, T. (2019). A Machine Learning Model for Hemoglobin Estimation and Anemia Classification. 17(2).
- García Chávez, H., & Cañedo Figueroa, C. (2021). Diseño de algoritmo compuesto por Machine Learning y un modelo probabilístico para la detección de diabetes. Memorias Del 44 Congreso Nacional de Ingeniería Biomédica.
- Rathore, D. K., & Mannepalli, P. . (2021). A Review of Machine Learning Techniques and Applications for Health Care. 4–8. <https://doi.org/10.1109/ICATME50232.2021.9732761>
- Shaik, M. F., & Subashini, M. M. (2017). Anemia diagnosis by fuzzy logic using LabVIEW. 1–5. <https://doi.org/10.1109/I2C2.2017.8321790>
- Yıldız, T. K., & Yurtay, N. (2021). Classifying anemia types using artificial learning methods. 24(1), 50–70.