Revista de la Facultad de Ciencias Universidad Nacional de Colombia, Sede Medellín
V 14 N°2 julio - diciembre 2025 • ISSN-e 2357-5749 • Artículo Investigación • Páginas 61 a 80
DOI: https://10.15446/rev.fac.cienc.v14n2.114063

PREDICCIÓN DE CASOS DE MALARIA EN LA REGIÓN DE ASHANTI GHANA UTILIZANDO LOS MODELOS NAIVE BAYES Y K VECINOS MÁS CERCANOS[a]

PREDICTION OF MALARIA CASES IN THE ASHANTI REGION GHANA USING THE NAIVE BAYES AND K NEAREST NEIGHBORS MODELS

JAVIER MOSQUERA RENTERIA[b*] JUAN CARLOS SALAZAR-URIBE[c] ELLIS KOBINA PAINTSIL[d]

Recibido 29-08-2024, aceptado 26-03-2025, versión final 03-06-2025.
Artículo Investigación

RESUMEN: La malaria es una enfermedad causada por parásitos del género Plasmodium, transmitidos por la picadura de mosquitos hembra del género Anopheles. En Ghana, esta enfermedad representa un problema de salud pública significativo, siendo la principal causa de morbilidad y responsable de entre el 40 % y el 60 % de las hospitalizaciones en el país. El objetivo de este estudio fue comparar el desempeño de los modelos de aprendizaje estadístico K vecinos más cercanos y Naive Bayes en la predicción del número de casos probables de malaria. Se utilizó variables hematológicas, edad y género de pacientes, cuyos datos fueron extraídos del laboratorio del Hospital St. Patrick's de Ashanti, en Ghana desde enero de 2018, a junio del mismo año, a partir del libro de registros hematológicos. Para este estudio se utilizaron datos de 2076 pacientes. En la evaluación de los modelos, Naive Bayes obtuvo un AUC del 80.6 % y una precisión del 77.9 %, destacando variables hematológicas como Hemoglobina (Hb), Plaquetas (Plt) y Linfocitos (Lymph), además de la Edad (Age), como las más relevantes dentro del conjunto analizado. Su sensibilidad fue del 70.3 %, lo que indica una buena capacidad para identificar casos positivos, aunque con un valor predictivo positivo del 38.1 %, reflejando una alta tasa de falsos positivos. Sin embargo, su valor predictivo negativo alcanzó el 93.6 %, lo que sugiere una alta confiabilidad en la identificación de casos negativos. En contraste, KNN presentó un AUC del 70.3 % y una precisión del 78.4 %, pero con una menor sensibilidad del 39 %. Su especificidad fue del 94 %, permitiéndole identificar correctamente la mayoría de los casos negativos, con un valor predictivo positivo del 71.9 %, indicando mayor precisión en la clasificación de casos positivos. En conclusión, Naive Bayes mostró mejor desempeño en términos de AUC y sensibilidad, aunque su capacidad para predecir correctamente casos positivos fue inferior a la de KNN.

PALABRAS CLAVE: K vecinos más cercanos; machine learning; naive bayes; aprendizaje estadístico; malaria.

---

[a]Mosquera-Rentería, J., Salazar-Uribe, J. C. & Kobina-Paintsil, E. (2025). Predicción de casos de malaria en la región de Ashanti Ghana utilizando los modelos naive Bayes y k vecinos más cercanos. *Rev. Fac. Cienc.*, 14 (**2**), 61–80. DOI: https://10.15446/rev.fac.cienc.v14n2.114063

[b]M. Sc. en Ciencias - Estadística. Facultad de Ciencias. Universidad Nacional de Colombia, Medellín, Colombia.
[*]Autor para correspondencia: jamosquerar@unal.edu.co
[c]Ph.D. Profesor Titular. Facultad de Ciencias. Universidad Nacional de Colombia, Medellín, Colombia.
[d]Centro Kumasi de Investigación Colaborativa en Medicina Tropical, KNUST, Kumasi, Ghana.

ABSTRACT: Malaria is a disease caused by parasites of the genus Plasmodium, transmitted by the bite of female Anopheles mosquitoes. In Ghana, this disease represents a significant public health problem, being the leading cause of morbidity and responsible for between 40 % and 60 % of hospitalizations in the country. The objective of this study was to compare the performance of K-nearest neighbors and Naive Bayes statistical learning models in predicting the number of probable malaria cases. Hematological variables, age, and gender of patients were used. Data were extracted from the hematology record book of St. Patrick's Hospital, Ashanti, Ghana, from January 2018 to June of the same year. Data from 2,076 patients were used for this study. In the evaluation of the models, Naive Bayes obtained an AUC of 80.6 % and an accuracy of 77.9 %, highlighting hematological variables such as Hemoglobin (Hb), Platelets (Plt), and Lymphocytes (Lymph), in addition to Age (Age), as the most relevant within the analyzed set. Its sensitivity was 70.3 %, indicating a good ability to identify positive cases, although with a positive predictive value of 38.1 %, reflecting a high rate of false positives. However, its negative predictive value reached 93.6 %, suggesting high reliability in identifying negative cases. In contrast, KNN presented an AUC of 70.3 % and an accuracy of 78.4 %, but with a lower sensitivity of 39 %. Its specificity was 94 %, allowing it to correctly identify most negative cases, with a positive predictive value of 71.9 %, indicating greater accuracy in classifying positive cases. In conclusion, Naive Bayes showed better performance in terms of AUC and sensitivity, although its ability to correctly predict positive cases was inferior to that of KNN.

KEYWORDS: K nearest neighbors; machine Learning; naive bayes; statistical learning; malaria.

1. INTRODUCCIÓN

La malaria es una enfermedad causada por el mosquito hembra Anopheles del género Plasmodium, un género de parásitos protozoarios (Deress & Girma, 2019). Entre los cinco tipos de parásitos que infectan a los seres humanos, el Plasmodium falciparum y el Plasmodium vivax son los más relevantes. La malaria es la enfermedad de mayor frecuencia y mortalidad en todo el mundo (Deress & Girma, 2019). La propagación de está enfermedad, está relacionada con situaciones ambientales como la topografía, las precipitaciones, el clima y las condiciones socioeconómicas de la población (Deress & Girma, 2019).

Después de la pérdida de vidas, la malaria es la causa de mayor morbilidad en Ghana y representa entre el 40 % y el 60 % de las hospitalizaciones. En 2002, controlar la malaria le costó al gobierno de Ghana 50.05 millones de dólares; mientras que las empresas privadas tuvieron pérdidas alrededor de 6.58 millones de dólares en 2014 debido a esta enfermedad (Kobina *et al*., 2019).

La malaria se considera endémica en más de 100 países en todo el mundo, y muchos casos de muerte debido a la malaria están en África subsahariana. La enfermedad es un problema de salud pública debido a su impacto en personas de todos los grupos de edad, en particular en personas embarazadas y en niños. Debido a la falta de tratamiento rápido, la malaria sigue siendo una enfermedad peligrosa y aproximadamente 3200 millones de personas, casi la mitad de la población mundial, están en riesgo de verse afectadas por esta enfermedad. La malaria se ha mostrado como un problema de salud arraigado en todo el mundo, aunque en

menor proporción, Asia, América Latina, Oriente Medio y una parte de Europa se ven afectadas por esta enfermedad (Ankamah *et al*., 2018).

Debido a que la malaria ha sido erradicada con éxito en las regiones templadas del mundo, su incidencia está aumentando en África, donde el parásito causante de la enfermedad sigue propagándose. En Ghana, la malaria es un problema de salud de gran relevancia que afecta a todos los grupos de la sociedad. Como resultado, el gobierno implementó un sistema nacional para el control de la malaria, con el objetivo de reducir la mortalidad y morbilidad de la enfermedad. Desde los años 1998 hasta 2006, se llevó a cabo un control del vector de la malaria en el país, mediante el uso de mosquiteros tratados con insecticidas de larga duración y fumigación en interiores (Lave *et al*., 2017).

La malaria en la placenta es una de las causas fundamentales de recién nacidos con bajo peso al nacer, principalmente en mujeres embarazadas por primera vez (Gaw *et al*., 2019). Cada año, cerca de 85 millones de mujeres embarazadas en el mundo tienen riesgo de contraer el parásito de la malaria. La malaria en las mujeres embarazadas lleva a malos resultados perinatales como bajo peso al nacer en los niños, problemas de crecimiento del feto y parto prematuro (Gaw *et al*., 2019). La malaria en la placenta de las mujeres se caracteriza por la concentración de glóbulos rojos infectados por Plasmodium falciparum en áreas intervellosas de la placenta (Gaw *et al*., 2019).

Las lesiones renales agudas son una de las complicaciones más temibles de la enfermedad de la malaria. La resistencia al tratamiento y el aumento de la virulencia aumentan las probabilidades de que se desarrollen estas y otras complicaciones generadas por el parásito de la malaria (Muhamedhussein *et al*., 2019). Se plantean algunas hipótesis relacionadas con la patogénesis de la lesión renal relacionada con la malaria, incluida la obstrucción por glóbulos rojos infectados y una respuesta inmune excesiva (Muhamedhussein *et al*., 2019).

En las zonas endémicas de malaria, la inmunidad clínica se desarrolla gradualmente, pero las estrategias preventivas pueden ralentizar su adquisición. Las vacunas eficaces contra la malaria influyen en el riesgo de infección de dos maneras: directamente, al inducir una respuesta inmune, e indirectamente, al reducir la exposición a la enfermedad, lo que retrasa la inmunidad natural. En los ensayos clínicos, este retraso podría afectar la efectividad de las dosis de refuerzo de la vacuna y, una vez que la inmunidad inducida disminuya, la incidencia de malaria en los vacunados podría superar a la del grupo de control debido a diferencias en la inmunidad adquirida (Bun *et al*., 2020).

En África subsahariana, la asistencia a una clínica ocurre cuando el tratamiento en casa no es efectivo. Los centros de salud de nivel primario carecen de herramientas para detectar el parásito de la malaria, por lo que el diagnóstico de la malaria generalmente se basa en síntomas como la fiebre. A pesar de los métodos de algoritmos para mejorar la especificidad del diagnóstico por encima del 30 % al 70 %, dependiendo del

comportamiento de los patrones de transmisión de la malaria (Hume *et al*., 2008).

La prueba de diagnóstico rápido (PDR) para la detección del parásito de la malaria supera muchos obstáculos que enfrenta el uso del microscopio: es más fácil y rápido capacitar a personas no calificadas para el uso de las PDR que capacitar a un técnico en microscopía. Estudios de revisiones recientes han llevado a la conclusión de que la interpretación y el desempeño de las PDR son satisfactorios cuando son utilizadas por personas capacitadas de la comunidad (Ladier *et al*., 2016). La mitad de las PDR para la detección del parásito de la malaria sigue siendo positiva 15 días después del tratamiento de la enfermedad (Dalrymple *et al*., 2018).

Este estudio tiene como objetivo evaluar la capacidad predictiva de los modelos K vecinos más cercanos y Naive Bayes en la predicción del número de casos de malaria. Se utilizarán los parámetros hematológicos, la edad y el género de pacientes de la región de Ashanti, Ghana. Además de evaluar los modelos con fines predictivos, se seleccionará el modelo con el mejor desempeño entre los dos considerados.

Los modelos K vecinos más cercanos (KNN) y Naive Bayes fueron seleccionados en este estudio debido a su capacidad para predecir la probabilidad de que un paciente tenga malaria, lo que resulta crucial para la detección temprana, el tratamiento oportuno y la prevención de complicaciones graves. Su efectividad en tareas de clasificación ha sido ampliamente demostrada, especialmente en contextos donde los recursos médicos son limitados y es fundamental optimizar los diagnósticos.

Naive Bayes es un modelo probabilístico que asume independencia entre las variables, lo que permite clasificar datos de manera eficiente, facilitando su interpretación y rapidez de cálculo. KNN, por su parte, se basa en la similitud entre observaciones sin necesidad de asumir una distribución específica, lo que le brinda flexibilidad para identificar patrones complejos. La combinación de estas características hace que ambos métodos sean herramientas valiosas en el análisis de datos de malaria y respalda su elección en este estudio.

2. MÉTODOS

2.1. Fuente de información

Para este estudio, los datos fueron obtenidos del laboratorio del Hospital St. Patrick's, ubicado en Ashanti, Ghana desde enero de 2018 hasta junio del mismo año, a partir del libro de registros hematológicos (Kobina *et al*., 2019). Se utilizaron datos de hematología y pruebas de malaria de un total de 2076 pacientes con síntomas de malaria, de los cuales hubo 1200 mujeres y 876 hombres con edades comprendidas entre 1 y 102 años. El diagnóstico médico de malaria se realizó aplicando pruebas de diagnóstico rápido (PDR), y un microscopista capacitado confirmó los resultados de la prueba mediante la técnica de tinción llamada Giemsa. Los análisis de recuentos sanguíneos se realizaron aplicando el analizador Mindray BC-300 plus.

Los parámetros hematológicos considerados en este artículo fueron la hemoglobina (Hb), glóbulos blancos (WBC), plaquetas (Plt), linfocitos (Lymph), recuentos de células (MXD) y neutrófilos (Neut). A cada paciente con la enfermedad se le asignó el valor 1, y a los pacientes sin la enfermedad se les asignó el valor 0; para el género masculino se asignó el valor 1 y para el femenino el valor 0. En este estudio, se tuvieron en cuenta los modelos Naive Bayes y K vecinos más cercanos, teniendo en cuenta los datos de malaria; se consideraron estos métodos por ser estrategias de predicción modernas que han mostrado muy buen desempeño en la práctica.

2.2. Análisis de los datos

Se usaron técnicas descriptivas numéricas y gráficas para resumir las características clínicas y demográficas presentes en la base de datos y obtener posibles datos faltantes o extremos. Esto se hizo usando el software R Core Team (2023) y RStudio Team (2023). Luego se ajusta Naive Bayes y K vecinos más cercanos. El desempeño de los modelos de predicción se mide con áreas bajo la curva ROC y AUC, el accuracy, la sensitividad y el RMSE, que son formas estándar aceptadas científicamente para comparar clasificadores (Bishop, 2006).

2.3. Conceptos básicos del modelo Naive Bayes y K vecinos más cercanos

2.3.1. Naive Bayes

Teniendo en cuenta que la estimación puede requerir una gran cantidad de datos, el uso del modelo Naive Bayes (NB) se vuelve una alternativa valiosa en diferentes situaciones. Principalmente, el supuesto de independencia entre las variables en Naive Bayes puede generar un sesgo en la estimación de las probabilidades condicionales, ya que en muchos casos las variables no son completamente independientes. Sin embargo, este sesgo se compensa con una reducción en la varianza, lo que conduce a un clasificador que funciona bien y mantiene un equilibrio entre varianza y sesgo (James *et al*., 2021).

El teorema de Bayes se puede expresar como (Walpole *et al*., 2012):

$$P(Y=1|X) = \frac{P(X|Y=1)P(Y=1)}{P(X)}, \tag{1}$$

donde,

- $P(Y=1|X)$: es la probabilidad de que el paciente tenga malaria, dado el conjunto de datos hematológicos de los pacientes representado por X.
- $P(X|Y=1)$: es la probabilidad de observar el conjunto de datos hematológicos X dado que el paciente tiene malaria.
- $P(Y=1)$: es la probabilidad a priori de que un paciente tenga malaria.

- $P(X)$: es la probabilidad a priori de observar los conjuntos de datos hematológicos X.

El Naive Bayes para los casos de malaria asume que hay independencia entre las componentes del vector de hematología X. Teniendo en cuenta esto, la probabilidad de observar el vector de hematología $X = (X_1, X_2, \ldots, X_P)$, dado que pertenece al grupo de pacientes con malaria es (Wackerly *et al*., 2008):

$$P(X|Y=1) = P(X_1|Y=1) \times P(X_2|Y=1) \times \ldots \times P(X_p|Y=1). \tag{2}$$

$P(X_i|Y=1)$ es la probabilidad de que el grupo de pacientes con malaria genere el i-esimo vector de hematología observado, siendo $i = 1, 2, \ldots, p$.

Para el vector de hematología X, el modelo Naive Bayes se puede aplicar de la siguiente manera:

$$P(Y=1|X) = \frac{P(X|Y=1)P(Y=1)}{P(X)} \tag{3}$$

$$= \frac{P(X_1|Y=1) \times P(X_2|Y=1) \times \cdots \times P(X_p|Y=1) \times P(Y=1)}{P(X)} \tag{4}$$

$$= \frac{\prod_{i=1}^{p} P(X_i|Y=1) \times P(Y=1)}{P(X)} \propto \prod_{i=1}^{p} P(X_i|Y=1) \times P(Y=1).$$

La ecuación anterior se utiliza para clasificar a los pacientes que posiblemente tengan la enfermedad en el grupo de los que tienen malaria o en el grupo de los que no la padecen. Para Naive Bayes, se usa el 80 % como muestra de entrenamiento y 20 % como muestra de prueba.

2.3.2. K vecinos más cercanos

El método de K vecinos más cercanos (KNN) es un método relevante por su flexibilidad (James *et al*., 2021). Para cualquier entero positivo K y una observación de prueba denotada como x_0, el KNN encuentra los K puntos en el conjunto de datos de entrenamiento denotado como $\mathcal{N}_0$. En este caso, el conjunto de datos de entrenamiento de malaria consta de 1660 observaciones, lo que corresponde al 80 % del total de observaciones (2076). Luego, se estima la probabilidad para la clase j teniendo en cuenta el conjunto de datos de prueba (test). En este caso, se tomó un conjunto de prueba que equivale a 416 observaciones, es decir, el 20 % de las 2076 observaciones, mediante el uso de la siguiente ecuación (James *et al*., 2021):

$$P(Y=j|X=x_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(y_i = j),\ j \in \{0,1\}. \tag{5}$$

La ecuación anterior representa la probabilidad de clasificación de una observación en cualquiera de los grupos de malaria, para los K vecinos más cercanos a la observación x_0, donde $I(y_i = j)$ se define como:

$$I(y_i = j) = \begin{cases} 1, & si\ y_i = j \\ 0, & si\ y_i \neq j \end{cases} \tag{6}$$

donde $i = 1,2,\ldots,n$ y $j \in \{0,1\}$.

Se debe tener en cuenta que el KNN clasifica la observación de prueba x_0 en el grupo de pacientes con la mayor probabilidad. Para calcular las distancias más cercanas al punto x_0, se utilizan las distancias euclidianas, que están dadas de la siguiente manera: Para dos puntos en $\mathbb{R}^p$, con $X_h = (X_{h1},\ldots,X_{hp})$ y $X_i = (X_{i1},\ldots,X_{ip})$, la distancia euclidiana se expresa de la siguiente manera (Johnson & Wichern, 2007):

$$d_{hi} = \left[\sum_{j=1}^{p}(X_{hj} - X_{ij})^2\right]^{1/2}. \tag{7}$$

La ecuación anterior, donde h e i en X representan dos observaciones distintas del conjunto de datos de malaria, permite hallar las distancias de las observaciones de malaria respecto a una observación específica x_0, que se asigna a uno de los dos grupos de malaria.

3. RESULTADOS

3.1. Análisis descriptivo

Según la Figura 1, de un total de 2076 personas que asistieron al hospital para realizarse la prueba de malaria, 539, lo que representa el 25.95 %, dieron positivo para malaria.

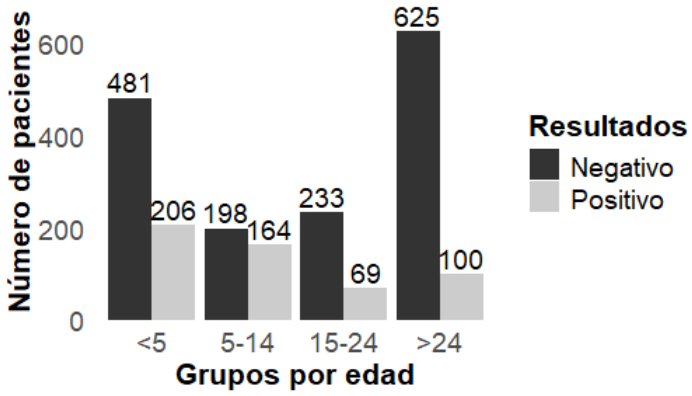

Figura 1: Gráfico de barras por grupos de edad para los casos de malaria. Fuente: Elaboración propia.

Tabla 1: Valores para el Mínimo, Q_1, Q_2, Q_3, el Máximo y prueba t de Student, para comparar la edad de acuerdo al sexo. Fuente: Elaboración propia.

Variable	Mínimo.	1^{er} cuant.	Mediana.	Media.	3^{er} cuant.	Máximo.
Edad (Femenino)	1.00	4.00	20.00	22.65	33.00	102.00
Edad (Masculino)	1.00	2.00	7.00	17.28	28	87
		Prueba t de Student				
		$t = 120.99$ $df = 1885.6$ p-value$< 2.2e^{-16}$				

En cuanto a los diferentes grupos de edad, se observó que los niños menores de 5 años presentaron la mayoría de los casos de malaria, llegando a un 29.98 % (206 de 687), seguidos por los niños de entre 5 y 14 años, con un 45 % (164 de 362). Las edades que registraron la cifra más baja de casos de malaria fueron las comprendidas entre los 15 y 24 años, con un 29.6 % (69 de 302), mientras que los mayores de 24 años tuvieron un porcentaje equivalente al 13.8 % (100 de 725).

Al realizar la prueba t de Student para las edades en los grupos femenino y masculino, y observar que el valor p es menor que 0.05, se concluye que existen diferencias significativas para sus medias (ver Tabla 1). Es decir, existe una diferencia en las edades de las mujeres en comparación con las edades de los hombres, siendo mayores las mujeres.

En la matriz de dispersión (Figura 2), la nube de puntos en la entrada $(2,1)$ muestra una correlación débil entre las variables Age (edad) y Hb (Hemoglobina), con un valor de 0.273. Esto se refleja en la dificultad para ajustar una línea recta a los datos.

Asimismo, al analizar por separado a los pacientes con y sin malaria, se observa que la correlación entre Age (edad) y Hb (Hemoglobina) sigue siendo baja, con valores de 0.195 y 0.390, respectivamente. En la gráfica, los puntos negros representan a los pacientes sin malaria, mientras que los grises corresponden a aquellos que tuvieron la enfermedad.

Estos resultados sugieren una baja colinealidad, lo que indica que no afecta significativamente el proceso de estimación.

En la nube de puntos de la entrada $(4,2)$, se observa una correlación prácticamente inexistente entre la hemoglobina (Hb) y las plaquetas (Plt), con un valor de -0.033. Esto indica que no hay una relación significativa entre ambas variables y que no es sencillo ajustar una línea recta a los datos. Al analizar por separado a los pacientes sin malaria y aquellos que tuvieron la enfermedad, las correlaciones siguen siendo bajas, con valores de -0.220 y 0.118, respectivamente.

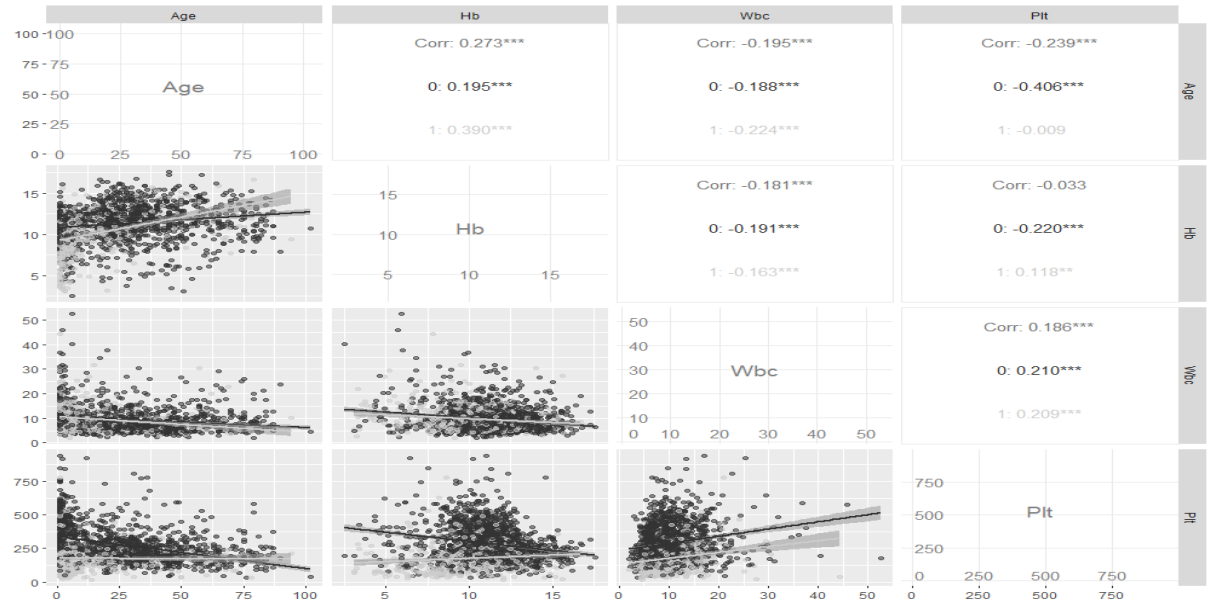

Figura 2: Matriz de dispersión para las variables. Fuente: Elaboración propia.

Al analizar la nube de puntos en la entrada $(3,2)$, se evidencia una correlación prácticamente nula, lo que indica la ausencia de relación entre las variables hemoglobina (Hb) y glóbulo blancos (Wbc), con un valor de -0.181. Además, se observa que no es sencillo ajustar una línea recta a los datos. En cuanto a la nube de puntos que distingue a los pacientes sin malaria de aquellos con diagnóstico positivo, se identificaron correlaciones negativa y moderadamente débiles, con valores de -0.191 y -0.163, respectivamente. De manera similar, las interpretaciones para las nubes de puntos correspondientes a las entradas $(3,1)$, $(4,1)$ y $(4,2)$ siguen el mismo patrón.

3.2. Resultados para el Naive Bayes

Es importante tener en cuenta que la mayor área bajo la curva (AUC) se obtuvo con las variables hemoglobina (Hb), plaquetas (Plt), linfocitos (Lymph) y edad (Age), lo que indica que estas variables son las más relevantes para distinguir entre clases positivas y negativas. Considerando la hemoglobina y la edad, representadas como $X = (Hb, Plt, Lymph, Age)$, para el modelo Naive Bayes, se tiene:

$$P(Y=1|X) = \frac{P(X|Y=1)P(Y=1)}{P(X)} \tag{8}$$
$$\propto \frac{P(Hb|Y=1) \times P(Plt|Y=1) \times P(Lymph|Y=1) \times P(Age|Y=1) \times P(Y=1)}{P(X)},$$

donde

Tabla 2: Matriz de confusión para el Naive Bayes. Fuente: Elaboración propia.

	Predichos		
Observados	Sin malaria (0)	Con malaria (1)	Total
Sin malaria (0)	279	19	298
Con malaria (1)	73	45	118
Total	352	64	416

Tabla 3: Resumen de algunos estadísticos importantes para la matriz de confusión del NB. Fuente: Elaboración propia.

Accuracy	0.779	Pos Pred Value	0.381
95 % CI	(0.736, 0.818)	Neg Pred Value	0.936
Sensitivity	0.703	Specificity	0.793

- $P(Y = 1|X)$: es la probabilidad de pertenecer a la clase 1, dado el vector de características $X = (Hb, Plt, Lymph, Age)$.
- $P(X|Y = 1)$: es la probabilidad de observar el vector $X = (Hb, Plt, Lymph, Age)$ dado que está en el grupo de malaria.
- $P(Y = 1)$: es la probabilidad para el grupo de malaria dado el vector X.
- $P(X)$: es la probabilidad de obtener $X = (Hb, Plt, Lymph, Age)$.

La ecuación anterior permitió clasificar a pacientes con malaria.

3.2.1. Interpretación de los resultados

Para la matriz de confusión presentada en la Tabla 2, se observa que para el modelo NB, de los 118 pacientes que tenían el parásito, el modelo clasificó 45 con malaria, lo que equivale al 38.1 %, y 73 pacientes sin la enfermedad, representando el 61.9 % restante. De un total de 298 pacientes sin el parásito, el modelo predijo 19 de ellos con la enfermedad y 279 sin malaria, lo que equivale al 6.38 % y 93.62 %, respectivamente. El modelo tiene una capacidad de clasificar correctamente el 77.9 % de las observaciones con el parásito de la malaria. En la Figura 3, se observa que el área bajo la curva es del 80.6 %. Se puede concluir que la capacidad de discriminación del NB fue buena.

En la Tabla 3, se presentan algunos estadísticos para la matriz de confusión, Tabla 2. Para la exactitud o precisión (Accuracy), que muestra la proporción de predicciones correctas respecto al total de observaciones, se obtiene un valor del 77.9 % con un intervalo de confianza del 95 % entre el 73.6 % y el 81.8 %.

La sensibilidad (Sensitivity), que muestra la proporción de pacientes con la enfermedad correctamente identificados (conocida como tasa de verdaderos positivos), alcanza el 70.3 %. Para el valor predictivo positivo,

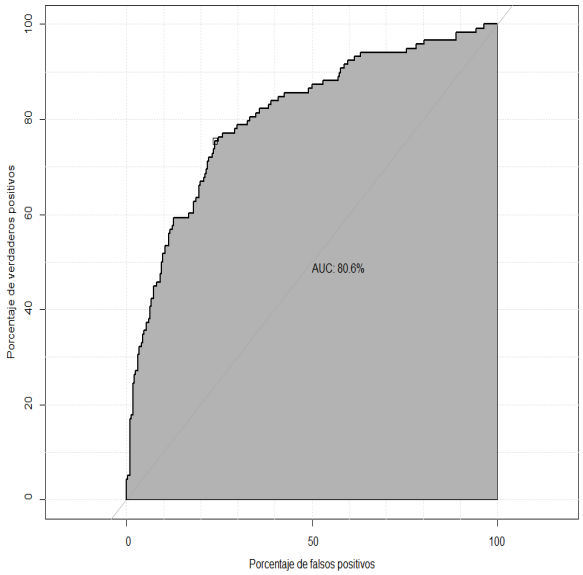

Figura 3: Curva ROC y AUC para el Naive Bayes. Fuente: Elaboración propia.

que es la proporción de pacientes clasificados con malaria que son realmente positivos, se obtuvo un valor del 38.1 %, mostrando una confiabilidad bastante baja en los pacientes con malaria que fueron clasificados. El valor predictivo para pacientes sin la enfermedad, que mide la proporción de pacientes clasificados para malaria que son en realidad negativos, alcanza el 93.6 %.

3.3. Resultados para el K vecinos más cercanos

Para calcular las K distancias más cercanas a un punto que se clasificó en uno de los dos grupos de malaria, se seleccionaron dos puntos en $\mathbb{R}^4$, con $X_h = (Hb_h, Plt_h, Lymph_h, Age_h)$ y $X_i = (Hb_i, Plt_i, Lymph_i, Age_i)$, y se calcularon las distancias euclidianas.

3.3.1. Interpretación de los resultados

En el método de los K vecinos más cercanos, la matriz de confusión (Tabla 4) muestra que, de un total de 64 pacientes que tuvieron malaria, el modelo clasifica correctamente a 58, lo que representa el 71.9 %. El 28.1 % restante de los 64 pacientes no fue diagnosticado con el parásito de la malaria.Para los pacientes que tuvieron malaria (352 pacientes en total), el modelo predijo correctamente el parásito en 72 pacientes, lo que equivale al 20.5 % de probabilidad de acierto.

En cuanto a los 280 pacientes restantes, el modelo no predijo la enfermedad, lo que corresponde al 79.5 %. El modelo tiene la capacidad de clasificar correctamente el 78.4 % de las observaciones, tanto las relaciona-

Tabla 4: Matriz de confusión para el K vecinos más cercanos. Fuente: Elaboración propia.

	Predichos		
Observados	Sin malaria (0)	Con malaria (1)	Total
Sin malaria (0)	280	72	352
Con malaria (1)	18	46	64
Total	298	118	416

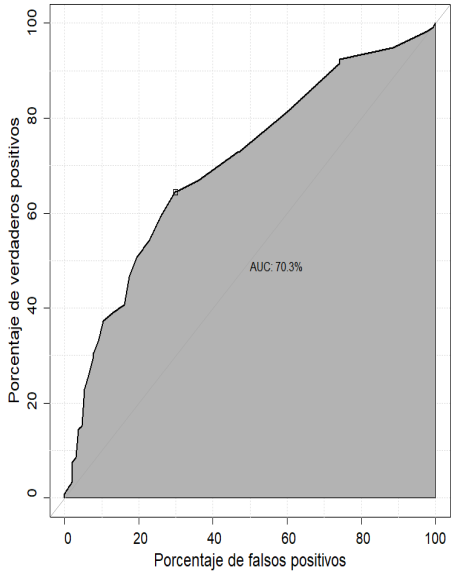

Figura 4: Curva ROC y AUC para K vecinos más cercanos. Fuente: Elaboración propia.

das con la enfermedad como las que no lo están.

Observando la Figura 4, para el caso de los K vecinos más cercanos, el área bajo la curva ROC es del 70.3 %, lo que indica una discriminación moderada, inferior al AUC del NB.

En la Tabla 5 se presenta un resumen de los estadísticos más importantes para la matriz de confusión del método de los K vecinos más cercanos, como se muestra en la Tabla 4. Para el método de los K vecinos más cercanos, el modelo tiene una exactitud del 78.4 % en la clasificación de casos de malaria, lo que indica que aproximadamente el 78.4 % de las predicciones son clasificadas correctamente. Por otro lado, la exactitud del modelo está entre el 74.1 % y el 82.2 % con un nivel de confianza del 95 %.

Por otra parte, la especificidad muestra que el modelo identifica correctamente aproximadamente el 94.0 % de los pacientes que en realidad no tienen la enfermedad.

Tabla 5: Resumen de los estadísticos para la matriz de confusión del KNN. Fuente: Elaboración propia.

Accuracy	0.784	Pos Pred Value	0.719
95 % CI	(0.741, 0.822)	Neg Pred Value	0.795
Sensitivity	0.390	Specificity	0.940

Tabla 6: Áreas bajo la curva ROC para el NB y KNN. Fuente: Elaboración propia.

	Área bajo la curva ROC (AUC)
Naive Bayes (NB)	80.6 %
K vecinos más cercanos (KNN)	70.3 %

La sensibilidad, o tasa de verdaderos positivos, muestra la proporción de pacientes positivos que el modelo identifica correctamente. En este caso, el modelo identifica alrededor del 39 % de los pacientes con el parásito. El valor predictivo positivo (Pos Pred Value), que fue de alrededor del 71.9 % de los pacientes clasificados con la enfermedad, refleja la proporción de pacientes que en realidad tienen malaria. Para el valor predictivo negativo (Neg Pred Value), este alcanzó el 79.5 %, indicando la proporción de pacientes sin la enfermedad que son verdaderamente negativos.

En la Tabla 6 se presenta un resumen de la capacidad de predicción de los dos modelos evaluados para los casos de malaria en Ashanti, donde se observan las áreas bajo la curva para cada modelo. Se puede ver que el modelo con mayor capacidad de predicción para el parásito de la malaria es el Naive Bayes, con un área bajo la curva del 80.6 %. Sin embargo, aunque el área bajo la curva del K vecinos más cercanos es del 70.3 %, tiene una baja capacidad de predicción en comparación con el Naive Bayes.

En la Tabla 7 se muestran los errores cuadráticos medios para los dos modelos propuestos. En resumen, al evaluar los errores cuadráticos medios, se puede observar que K vecinos más cercanos tiene el error cuadrático medio más bajo que el Naive Bayes; esto muestra que en términos de error cuadrático medio el modelo K vecinos más cercanos se podría considerar el de mayor poder predictivo aunque la diferencia con el NB es muy pequeña. Sin embargo, también se pueden considerar otras métricas, como la sensibilidad y el AUC, para evaluar el rendimiento de los modelos.

Tabla 7: Error cuadrático medio para el NB y KNN. Fuente: Elaboración propia.

	Error cuadrático medio (MSE)
Naive Bayes (NB)	0.221
K vecinos más cercanos (KNN)	0.216

3.3.2. Comparación de los resultados

Después de realizar una comparación detallada de los modelos Naive Bayes y K vecinos más cercanos en la tarea de predicción de casos de malaria, se obtuvieron las siguientes observaciones clave:

- En cuanto a la exactitud, el modelo Naive Bayes presenta un valor ligeramente superior al del modelo de K vecinos más cercanos, con exactitudes de 0.784 y 0.779, respectivamente. Esto indica que Naive Bayes clasifica correctamente a una proporción levemente mayor de pacientes con malaria.
- En cuanto al valor predictivo positivo, el modelo K vecinos más cercanos tiene un valor (0.719) significativamente más alto que el modelo Naive Bayes (0.381), lo que sugiere que K vecinos más cercanos es mejor para predecir casos positivos de malaria. Esto significa que cuando el modelo K vecinos más cercanos predice que un caso de malaria es positivo, es más probable que sea correcto en comparación con Naive Bayes.
- Para el intervalo de confianza al 95 %, ambos modelos tienen intervalos que no se superponen, lo que sugiere diferencias significativas en la precisión entre los dos modelos.
- En cuanto a sensibilidad y especificidad, el modelo K vecinos más cercanos tiene una sensibilidad más baja (0.390) pero una especificidad mucho más alta (0.940) en comparación con el Naive Bayes, que tiene una sensibilidad de 0.703 y una especificidad de 0.793. Esto indica que el modelo K vecinos más cercanos es mejor para identificar casos de malaria que son verdaderos negativos, pero menos efectivo en identificar casos que son verdaderos positivos en comparación con Naive Bayes.
- En cuanto al área bajo la curva (AUC), Naive Bayes tiene un AUC más alto del 80.6 % que K vecinos más cercanos, que tiene un AUC del 70.3 %. Esto muestra que Naive Bayes tiene un mejor rendimiento en la clasificación para ambos grupos de malaria en general.
- Para el error cuadrático medio (MSE), K vecinos más cercanos tiene un MSE ligeramente menor, en este caso 0.216 en comparación con Naive Bayes con un MSE de 0.221. Esto sugiere que K vecinos más cercanos tiene un mejor rendimiento en precisión de las predicciones de casos de malaria en términos de la diferencia entre los valores predichos y reales. En resumen, mientras que Naive Bayes tiene un mejor desempeño en términos de exactitud, área bajo la curva y valor predictivo negativo, el modelo K vecinos más cercanos lo supera en cuanto al valor predictivo positivo y tiene un MSE ligeramente más bajo.

4. DISCUSIÓN DE RESULTADOS

4.1. Comparación entre modelos

Los resultados obtenidos en la clasificación de casos de malaria utilizando los modelos Naive Bayes (NB) y K vecinos más cercanos (KNN) muestran diferencias significativas en su desempeño. En términos de

área bajo la curva ROC (AUC), el modelo Naive Bayes tuvo un mejor desempeño con un valor de 80.6 %, mientras que el KNN obtuvo un AUC del 70.3 %. Esto indica que Naive Bayes tiene una mejor capacidad discriminativa para diferenciar entre pacientes con y sin malaria.

Al evaluar la matriz de confusión, el modelo NB logró clasificar correctamente el 77.9 % de las observaciones con malaria, mientras que el KNN tuvo una precisión ligeramente superior del 78.4 %. No obstante, la sensibilidad del NB fue del 70.3 %, mayor que la del KNN (39.0 %), lo que sugiere que el NB es más confiable para identificar correctamente a los pacientes con malaria.

Por otro lado, el KNN tuvo un mejor desempeño en términos del error cuadrático medio (MSE), con un valor de 0.216 en comparación con el 0.221 obtenido por NB. Aunque la diferencia es mínima, este resultado indica que KNN tiene un menor error promedio en la clasificación de los casos.

Otro aspecto a destacar es el valor predictivo positivo (VPP), que indica la proporción de pacientes predichos con malaria que realmente tienen la enfermedad. En este caso, el KNN mostró un VPP del 71.9 %, superior al 38.1 % de NB. Sin embargo, en el valor predictivo negativo (VPN), NB tuvo un mejor resultado con 93.6 %, frente al 79.5 % del KNN, lo que sugiere que NB es más fiable para descartar casos de malaria.

4.2. Consideraciones sobre el desempeño de los modelos

El mejor desempeño del modelo Naive Bayes en términos de AUC y sensibilidad puede deberse a su suposición de independencia condicional entre las variables predictoras, lo que facilita una mejor generalización en la clasificación. Sin embargo, su bajo valor predictivo positivo indica que tiende a producir más falsos positivos en comparación con el KNN.

Por otro lado, el método de K vecinos más cercanos presenta una mayor capacidad para identificar correctamente los casos sin malaria, lo que se evidencia en su especificidad del 94.0 %. Sin embargo, su menor sensibilidad sugiere que puede no ser el modelo ideal cuando la prioridad es identificar todos los casos de malaria.

4.3. Comparación con otros estudios

4.3.1. Comparación del K vecinos más cercano con otros estudios

El presente estudio y el de (Suseela *et al*., 2019) coinciden en la aplicación del algoritmo de K Vecinos Más Cercanos (KNN) para la detección de malaria; sin embargo, presentan diferencias sustanciales en cuanto a los datos utilizados, la metodología y los resultados obtenidos. Mientras que (Suseela *et al*., 2019) se enfoca en la clasificación de parásitos de malaria en imágenes de frotis sanguíneo grueso, utilizando redes convolucionales (CNN) para la extracción de características y KNN para la clasificación, este estudio emplea KNN

para clasificar pacientes con o sin malaria en función de características clínicas, como parámetros hematológicos, edad y género.

En términos de rendimiento, el modelo basado en VGG19 + KNN del estudio de Suseela *et al*. (2019) reporta una sensibilidad del 97 %, mientras que el presente estudio obtuvo una exactitud del 78.4 %, una sensibilidad del 39 % y un área bajo la curva (AUC) del 70.3 %. Esta diferencia puede explicarse por el tipo de datos analizados, ya que las imágenes procesadas con redes neuronales profundas permiten extraer patrones más complejos, lo que facilita la detección de casos positivos.

Además, mientras que Suseela *et al*. (2019) emplea redes convolucionales para generar características antes de aplicar KNN, este estudio compara directamente el desempeño de KNN con Naive Bayes en datos clínicos, encontrando que Naive Bayes presentó una mejor AUC (80.6 %), aunque con un error cuadrático medio (MSE) mayor. Estos hallazgos sugieren que, si bien KNN es una herramienta útil para la clasificación de malaria basada en características clínicas, su desempeño podría mejorarse mediante la implementación de técnicas avanzadas de selección de características o el uso de enfoques híbridos que combinen aprendizaje automático con métodos de extracción de características más sofisticados, como en el estudio realizado por Suseela *et al*. (2019).

4.3.2. Comparación del Naive Bayes con otros estudios

Los resultados obtenidos en el modelo de Naive Bayes en el reciente estudio muestran una capacidad de discriminación aceptable, con un área bajo la curva (AUC) del 80.6 % y una exactitud global del 77.9 %. Sin embargo, la sensibilidad obtenida (70.3 %) y el bajo valor predictivo positivo (38.1 %) indican que el modelo presenta limitaciones al clasificar correctamente a los pacientes con malaria. Esta tendencia también fue observada en el estudio de Lamine & Sy (2021), donde se reportaron dificultades en la clasificación precisa de casos positivos, sugiriendo que Naive Bayes tiende a favorecer la clasificación de pacientes sin la enfermedad en detrimento de aquellos con malaria.

Ambos estudios identificaron variables relevantes para la clasificación de la malaria. En el reciente estudio, las variables hemoglobina (Hb), plaquetas (Plt), linfocitos (Lymph) y edad (Age) se destacaron como las más influyentes en la predicción de la enfermedad.

Sin embargo, una diferencia notable entre ambos estudios radica en la interpretación del desempeño del modelo. Mientras que en el estudio realizado por Lamine & Sy (2021) se enfatiza la necesidad de ajustar los modelos para reducir los falsos negativos, en el estudio reciente se hace más hincapié en la baja confiabilidad de las predicciones positivas. Esto pone de manifiesto que, aunque el modelo Naive Bayes es una herramienta valiosa para la clasificación de enfermedades, su aplicación en entornos médicos requiere mejoras en la selección de variables y en la calibración de probabilidades.
En síntesis, ambos estudios destacan el potencial del Naive Bayes para el diagnóstico de malaria, pero

también exponen sus limitaciones. La combinación de Naive Bayes con otros enfoques estadísticos o de aprendizaje automático podría ser una estrategia prometedora para mejorar la capacidad de detección de la enfermedad y reducir los errores de clasificación.

4.4. Implicaciones y futuras investigaciones

Estos resultados sugieren que la selección del modelo depende del objetivo principal del análisis. Si se busca minimizar los falsos negativos y mejorar la sensibilidad en la detección de malaria, Naive Bayes es la mejor opción. Si, en cambio, se desea reducir la tasa de falsos positivos y mejorar la especificidad, KNN podría ser más adecuado.

Futuras investigaciones podrían enfocarse en la combinación de ambos modelos o en el uso de enfoques híbridos para mejorar la precisión general en la clasificación. Además, podría explorarse el impacto de la selección de variables y el ajuste de hiperparámetros en el desempeño de ambos modelos.

4.5. Recomendaciones

4.5.1. Importancia de los resultados en salud pública y vigilancia epidemiológica:

Los resultados de este estudio pueden ser clave para el desarrollo de estrategias más eficientes en la vigilancia epidemiológica en Ghana. El uso de modelos como NB y KNN tiene el potencial de optimizar la detección temprana de la malaria, facilitando una respuesta más ágil y efectiva por parte de las autoridades de salud.

4.5.2. Aciertos y limitaciones del estudio:

- Aciertos: Se emplearon métodos de validación estrictos para evaluar el rendimiento de los modelos y se analizaron diversas métricas para garantizar una comparación detallada y completa.
- Limitaciones: La hipótesis de independencia en NB podría no sostenerse en todos los contextos, y la sensibilidad de KNN al tamaño de la muestra podría impactar su capacidad de generalización. Además, no se consideraron otros enfoques, como la optimización de hiperparámetros o la combinación de modelos.

4.5.3. Facilidad de aplicación en software estadístico:

- Tanto NB como KNN necesitan ajustes particulares para su implementación en paquetes estadísticos convencionales.
- A diferencia de la regresión logística, ampliamente disponible en la mayoría de los software estadísticos tradicionales, NB y KNN pueden no estar integrados de forma predeterminada en algunas plataformas utilizadas por epidemiólogos, lo que podría limitar su aplicación en la práctica.

- Versiones futuras del software podrían incorporar estas pruebas con parámetros ajustados, lo que simplificaría su uso en el ámbito de la salud pública.

5. CONCLUSIONES

Evaluando el desempeño de los dos modelos, se obtienen los siguientes resultados:

- Naive Bayes tiene un desempeño ligeramente superior en términos de exactitud y AUC en comparación con K vecinos más cercanos.
- K vecinos más cercanos, por otro lado, muestra un rendimiento significativamente mejor en valor predictivo positivo, lo que indica una mayor capacidad para predecir correctamente casos positivos de malaria.
- Naive Bayes tiene una mayor especificidad y un valor de predicción negativo más alto, lo que sugiere una mejor capacidad para predecir correctamente casos negativos de malaria.
- K vecinos más cercanos tiene un MSE ligeramente más bajo, lo que indica que, en promedio, comete menos errores en las predicciones de casos de malaria.

Basado en lo anterior, el modelo Naive Bayes es mejor para predecir casos de malaria. Tiene una mayor exactitud, un área bajo la curva más alta y un valor predictivo negativo más alto en comparación con el modelo K vecinos más cercanos. Aunque el modelo K vecinos más cercanos tiene un valor predictivo positivo más alto y un MSE ligeramente menor, en general, el Naive Bayes supera al K vecinos más cercanos en términos de rendimiento en la tarea de predicción de casos de malaria.

Contribución de los autores

Javier Mosquera Rentería: Puesta a punto de los datos de malaria, diseño de programas en R, diseño general del artículo, interpretación y discusión de resultados.
Juan Carlos Salazar: Revisión de resultados y apoyo en el diseño de los programas en R.
Ellis Kobina Paintsil: Facilitó la base de datos con la que se implementaron los modelos y aportó en la discusión de resultados.

Referencias

Ankamah, S., Nokoe, K. & Iddrisu, W. (2018). Modelling Trends of Climatic Variability and Malaria in Ghana Using Vector Autoregression. *Hindawi*, 2018(6124321), 1-11.

Bishop, C. M.(2006).Pattern Recogniton and Machine Learning- 738p.

Bun, Y., Ma, X., Lam, K. F. & Milligan, P. (2020). Estimation of the primary, secondary and composite effects of malaria vaccines using data on multiple clinical malaria episodes. *Elsevier*, 38(**32**), 4964-4969.

Dalrymple, Ú., Arambepola, R., Gething, P.W & Cameron, E. (2018). How long do rapid diagnostic tests remain positive after anti-malaria treatment?. *Malaria Journal*, 17(**1**), 1-13.

Deress, T & Girma, M. (2019). Plasmodium falciparum am Plasmodium vivax Prevalence in Ethiopia: A Systematic Review and Meta-Analysis. *Hindawi*, 2019(7065064), 1-12.

Gaw, S-L., Hromatka, B.S., Ngeleza, S., Buarpung, S., Ozorslan, N., Tshefu, A & Fisher, S.J. (2019). Differential Activation of Fetal Hofbaver Cells in Primigravidas Is Associated with Decreased Birth Weight in Symptomatic Placental Malaria. *Hindawi*, 2019(378174), 1-10.

Hume, J.C., Barnish, G., Mangal, T., Armázio, L., Streat, E. & Bates, I. (2008). Household cost of malaria overdiagnosis in rural Mozambique. *Malaria Journal*, 7(**1**), 1-8.

James, G., Witten, D., Hastie, T & Tibshirani, R. (2021). An Introduction to Statistical Learning with Applications in R. 607 p.

Johnson, R.A & Wichern, D.W.(2007). Applied Multivariate Statistical Analysis. 773 p.

Kobina, E., Yaw, A., Glover, M & Akwasi, M. (2019). Analysis of Haematological Parameters as Predictors of Malaria Infection Using a Logistic Regression Model: A case Study of a Hospital in The Ashanti Region of Ghana. *Hindawi*, 2019(1486370), 1-7.

Ladier, J., Parker, D.M., Myin, A., Carrara, V.I., Maung, K., Bonnington, C.A., Pukrittayakamee, S., Delmas, G & Nosten, F.H. (2016). The role of early detection and treatment in malaria elimination. *Malaria Journal*, 15(**1**), 1-8.

Lamine. & Sy, A. (2021). On the Efficiency of Machine Learning Models in Malaria Prediction.*Public Health and Informatics*, 281, 437-441. doi: 10.3233/SHTI210196

Lave, E., Aseidu, J & Adjei, E. (2017). A Weather-Based Prediction Model of Malaria Prevalence in Amenfi West District, Ghana. *Hindawi*, 2017(7820454), 1-8.

Muhamedhussein, M.S., Ghosh, S., Khanbhai, K., Maganga, E., Nagri, Z. & Manji. (2019). Prevalence and Factors Association with Acute Kidney Injury among Malaria Patients in Dar es Salaam: A Cross-Sectional Study. *Hindawi*, 2019(4396108), 1-7.

R Core Team. (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. [Consultada en 2023]. Disponible en: https://www.R-project.org/.

RStudio Team. (2023). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA. [Consultada en 2023]. Disponible en: http://www.rstudio.com/.

Suseela, Devi.S., Samantha, E., Priyadharshini, B., & Jetlin, C.P. (2021). Malaria Detection Using Machine Learning with K Nearest Neighbour Algorithm. *International Journal of Scientific Development and Research*, 6(2455-2631), 457-460.

Walpole, R.E., Myers, R.H., Myers, S.L & Ye, K.(2012). Probabilidad y estadística para ingeniería y ciencias. 816p.

Wackerly, D., Mendenhall, W & Scheaffer, R. (2008). Estadística matemática con aplicación. 911 p.