UNIVERSIDAD NACIONAL DE COLOMBIA

Modelamiento de casos de malaria en la región de Ashanti-Ghana usando regresión logística, Machine Learning y discriminante lineal de Fisher

Javier Mosquera Renteria

Universidad Nacional de Colombia
Facultad de Ciencias, Departamento de Estadística
Medellín, Colombia
2024

Modelamiento de casos de malaria en la región de Ashanti-Ghana usando regresión logística, Machine Learning y discriminante lineal de Fisher

Javier Mosquera Renteria

Tesis presentada como requisito parcial para optar al título de:
Magíster en Ciencias-Estadística

Director:
Juan Carlos Salazar Uribe, Ph.D.
Líneas de investigación:
Análisis Multivariado de Datos, Analítica y Bioestadística.

Universidad Nacional de Colombia
Facultad de Ciencias, Departamento de Estadística
Medellín, Colombia
2024

Si quieres vivir una vida feliz, átala a una meta, no a una persona o un objeto.

La crisis es necesaria para que la humanidad avance. Solo en momentos de crisis, surgen las grandes mentes.

- Albert Einstein

Agradecimientos

A Dios, primero que todo, ya que sin Él nada es posible, y a mis padres y hermanos que siempre estuvieron apoyándome en esta meta propuesta, a pesar de las dificultades, que no fueron un obstáculo para seguir adelante con mis objetivos propuestos.

A la Escuela de Estadística de la Universidad Nacional de Colombia Sede Medellín y a los profesores que me dictaron los cursos a lo largo de mis estudios, quienes con sus formas de enseñar me facilitaron las herramientas necesarias que fueron de gran utilidad para entender las temáticas de los cursos siguientes, y al científico de laboratorio Ellis Kobina Paintsil, quien me prestó los datos para poder realizar la tesis. En especial, al profesor Juan Carlos Salazar Uribe, quien, además de acompañarme y asesorarme en todo el proceso de mi trabajo, fue un gran profesor y motivador, para que de esta manera pudiera terminar con éxito este trabajo de tesis.

Modelamiento de casos de malaria en la región de Ashanti-Ghana usando regresión logística, Machine Learning y discriminante lineal de Fisher

Resumen

A nivel mundial, se han logrado importantes avances en la reducción de casos de malaria; sin embargo, la enfermedad sigue siendo un problema desafiante en la salud pública de Ghana. Aproximadamente, entre el 40 y el 60 % de las personas son hospitalizadas debido a esta enfermedad. A pesar de ello, el examen mediante el uso del microscopio sigue siendo el mejor método en todo el mundo para determinar si un paciente tiene o no el parásito de la malaria. Este estudio buscó determinar si los parámetros hematológicos, la edad y el género de los pacientes podrían usarse para predecir la malaria mediante el uso de modelos de regresión logística, naive Bayes, análisis discriminante lineal de Fisher y K vecinos más cercanos. Con estos modelos, que forman parte del aprendizaje automático, se buscó determinar en qué medida se podría estimar la probabilidad de que un paciente tenga o no la enfermedad de la malaria. Se evaluó qué tan buenas alternativas son estos modelos para estimar la probabilidad de tener la enfermedad. Utilizando R, se determinó la capacidad predictiva de los modelos considerados, así como la elección del mejor modelo según algún criterio estadístico, mediante el uso de datos reales. En este estudio, se observó que la prevalencia de casos de malaria fue del 25.95 %, siendo los niños menores de 5 años los más afectados, alcanzando el 29.98 % (206 de 687), seguidos por niños entre 5 y 14 años de edad, con un 45.30 % (164 de 362). El mejor modelo de los cuatro se utilizará para mejorar el diagnóstico de la malaria en pacientes; en este caso, el mejor modelo por su interpretabilidad y mayor capacidad predictiva de casos de malaria fue la regresión logística, demostrando un área bajo la curva de 81.5 %. La especificidad y sensibilidad fueron del 74.6 % y 79.89 %, respectivamente, con un valor predictivo positivo del 39.8 % y un valor predictivo negativo del 94.6 %.

Palabras claves: Regresión logística, Machine Learning, discriminante de Fisher, estadística, malaria.

Modeling of malaria cases in the Ashanti-Ghana region using logistic regression, Machine Learning, and Fisher's linear discriminant

Abstract

Globally, significant progress has been made in reducing malaria cases; However, the disease remains a challenging public health problem in Ghana. Approximately, between 40 and 60 %

of people are hospitalized due to this disease. Despite this, examination using a microscope remains the best method worldwide to determine whether or not a patient has the malaria parasite. This study sought to determine whether hematological parameters, age, and gender of patients could be used to predict malaria by using logistic regression, naïve Bayes, Fisher's linear discriminant analysis, and K-nearest neighbors models. With these models, which are part of machine learning, we sought to determine to what extent the probability of a patient having or not having malaria disease could be estimated. It was evaluated how good alternatives these models are for estimating the probability of having the disease. Using R, the predictive capacity of the considered models was determined, as well as the choice of the best model according to some statistical criterion, through the use of real data. In this study, it was observed that the prevalence of malaria cases was 25.95 %, with children under 5 years of age being the most affected, reaching 29.98 % (206 of 687), followed by children between 5 and 14 years of age. age, with 45.30 % (164 of 362). The best model of the four will be used to improve malaria diagnosis in patients; In this case, the best model due to its interpretability and greater predictive capacity for malaria cases was logistic regression, demonstrating an area under the curve of 81.5 %. The specificity and sensitivity were 74.6 % and 79.89 %, respectively, with a positive predictive value of 39.8 % and a negative predictive value of 94.6 %.

Keywords: Logistic regression, Machine Learning, Fisher discriminant, statistics, malaria.

Contenido

Lista de Figuras

Lista de Tablas

Introducción

La malaria es causada por el parásito Plasmodium, el cual se transmite a través de la picadura del mosquito Anopheles (Aliyu et al., 2021). Esta enfermedad sigue siendo una importante carga en todo el mundo, con más de 200 millones de casos aproximadamente y más de 400,000 muertes cada año (Poostchi et al., 2023). Por ejemplo, Ghana ha sido uno de los países que ha luchado intensamente contra esta epidemia. Los donantes de sangre que están sanos pueden portar el parásito Plasmodium sin mostrar signos de malaria. La sangre de dichos donantes representa un riesgo para las personas que reciben transfusiones (Adusei and Owusu-Ofori, 2018). Aproximadamente del 40 al 60 % de las personas hospitalizadas lo están debido a esta enfermedad. Sin embargo, el uso de microscopios sigue siendo la mejor alternativa utilizada en todo el mundo como método para determinar si un paciente tiene o no el parásito de la malaria. Ghana es un país que carece de suficientes microscopios como una alternativa más eficaz para identificar el parásito transmitido por el mosquito Anopheles en la población (Paintsil et al., 2019).

Se acepta ampliamente que la malaria es una enfermedad de los pobres (Sach and Malaney, 2002). Algunos médicos piensan que dejar pasar un caso de malaria o paludismo es un problema mucho mayor de lo que se piensa; por lo tanto, tratan a casi la totalidad de los pacientes que tengan fiebre con el tratamiento de la malaria (Leslie et al., 2012).

El elevado número de casos de malaria lleva al despilfarro en el sistema de salud, lo que implica un mayor gasto de los pacientes en el hospital y un alto ausentismo en el trabajo (Hume et al., 2008). En el diagnóstico de la malaria, las pruebas de diagnóstico rápido (PDR) son una forma rápida de detectar el parásito en las personas que posiblemente lo tengan, pero las PDR son costosas en comparación con el uso del microscopio, y también es posible que las pruebas rápidas no puedan detectar la infección por el parásito (Landier et al., 2016). En el trabajo de investigación de Stephen et al. (2021) se utilizaron modelos para predecir el brote de malaria utilizando técnicas de machine learning. Estos modelos fueron Naive Bayes, Support Vector Machine (SVM), Regresión Lineal, Regresión Logística y K Vecinos Más Cercanos.

Las creencias y prácticas relacionadas con la malaria están asociadas con la cultura y pueden relacionar su eficacia con las prácticas de control (Adera, 2003). Los síntomas de esta enfermedad están relacionados con la fiebre y generalmente suelen estar asociados con escalofríos, sudoración y dolor de cabeza (Zuluaga and Trujillo, 2010). En los estudios y análisis realizados se mostró que los mayores efectos están dados por las altas temperaturas, la humedad y la precipitación pluvial sobre el paludismo, donde estos se evidencian en los meses de septiembre, marzo y octubre (Ankamah et al., 2018).

La producción de anticuerpos anti-eritropoyetina puede estar relacionada con la patogenia de la anemia implicada por la malaria por Plasmodium falciparum durante el embarazo (Nkansah et al., 2023). Esta enfermedad sigue siendo muy preocupante en las mujeres embarazadas, donde se realiza un control cercano desde el comienzo del embarazo hasta el parto, con el objetivo de evitar resultados neonatales y maternos adversos (Anabire et al., 2023).

En el estudio de investigación realizado por Aliyu and Bada (2023) sobre la predicción de brotes de malaria en Nigeria, se diseñó un modelo mejorado que emplea el algoritmo de Naive Bayes y una red neuronal artificial. Dicho modelo fue entrenado con un extenso conjunto de datos recopilados en el estado de Kebbi entre 2020 y 2022, teniendo en cuenta variables como la temperatura, la humedad, la cantidad de precipitaciones, así como el número de casos registrados, tanto de malaria como de personas sin la enfermedad.

En el artículo desarrollado por Sajana and Narasingarao (2018) para la clasificación de la enfermedad de malaria desequilibrada, se propuso un estudio comparativo sobre la clasificación de datos desequilibrados de la malaria utilizando el clasificador naive Bayes. Se presentó un estudio clínico de 165 pacientes de diferentes grupos de edad recolectados en las salas médicas de Narasaraopet en la India, entre 2014 y 2017. Las variables utilizadas fueron la edad (Age), hemoglobina (Hemoglobin), glóbulos rojos (RBC), hematocrito (Hct), volumen corpuscular medio (Mcv), hemoglobina corpuscular media (Mch), concentración de hemoglobina corpuscular media (Mchc), plaquetas (Platelets), glóbulos blancos (WBC), granulocitos (Granuls), linfocitos (Lymphocytes), monocitos (Monocytes) y malaria (Malaria).

En el estudio llevado a cabo por Mahmoudi et al. (2006) con el fin de descubrir nuevos fármacos antipalúdicos, se emplearon índices topológicos como descriptores estructurales que se correlacionaron con la actividad antipalúdica mediante modelos de regresión múltiple y análisis discriminante lineal. Se desarrollaron dos ecuaciones discriminantes (FD_1) y (FD_2) que permitieron una clasificación exitosa; estos modelos se basaron en 27 fármacos contra un clon susceptible a la cloroquina (3D7) de falciparum. Los fármacos estudiados incluyen monensina, nigericina, vincristina, vindesina, etilhidrocupreína y salinomicina con IC_{50s}.

En la investigación realizado por Opoku-Ansah et al. (2019) sobre la identificación óptica del subproducto de la malaria Plasmodium falciparum y la estimación de la densidad parasitaria, se introdujo el empleo de técnicas ópticas. Estas técnicas fueron utilizadas para calcular la densidad parasitaria (PD) de Plasmodium falciparum en muestras de sangre infectadas. Se recolectaron muestras de sangre de individuos infectados con falciparum antes y después de recibir tratamiento. Durante la investigación, se observaron diferencias en las características de absorción óptica entre las muestras de sangre infectadas y las no infectadas. Se identificó una disminución en las densidades ópticas conforme aumentaba la PD. Además, se propuso un modelo de clasificación lineal basado en el enfoque de Fisher que utiliza intensidades de píxeles para estimar la PD como una alternativa a la estimación manual, lo que podría mejorar el diagnóstico y tratamiento de la malaria.

En el trabajo de investigación realizado por Seyoum et al. (2023) sobre el uso de mosquiteros tratados con insecticida (MIT) y factores asociados entre hogares con niños menores de cinco años en África Oriental, el objetivo fue evaluar el uso de MIT y los factores relacionados en estos hogares. Se utilizaron los conjuntos de datos más recientes de la Encuesta Demográfica y de Salud, y se aplicó un modelo de regresión logística binaria multinivel para identificar los factores asociados con el uso de MIT.

Según la investigación llevada a cabo por Devi et al. (2021) sobre la detección de malaria mediante el uso de aprendizaje automático, se ha desarrollado un método capaz de identificar el parásito de la malaria en imágenes de frotis de sangre espesa. Este método consta de dos pasos principales. Primero, se aplica la detección mínima global iterativa (IGMS), que se basa en la fuerza y permite una detección rápida para identificar parásitos emergentes. Luego, se utiliza el algoritmo de los K vecinos más cercanos (KNN).

En una investigación reciente realizada por Irmanita et al. (2021) acerca de la clasificación de las complicaciones de la malaria, se desarrolló un sistema de predicción para identificar la malaria grave utilizando el método de árbol de clasificación y regresión (CART) y la probabilidad de complicaciones de malaria mediante el método naive Bayes. En la primera etapa del estudio, se determinaba si los pacientes presentaban síntomas de malaria grave o no, según el modelo construido. En la segunda etapa, si un paciente era clasificado como malaria grave, se predecía la probabilidad de complicaciones por malaria.

En un enfoque innovador de aprendizaje conjunto para identificar la malaria, los investigadores Qadri et al. (2023) propusieron estudiar el desarrollo de un modelo capaz de diagnosticar tempranamente la enfermedad, utilizando imágenes de glóbulos rojos parasitados y no infectados. Para los experimentos del estudio, se aplicaron métodos basados en redes neuronales, como Neural Search Architecture Network (NASNet), y se comparó su rendimiento con técnicas de aprendizaje automático.

Este estudio tiene como objetivo evaluar la capacidad predictiva de los modelos de los K vecinos más cercanos, naive Bayes, análisis discriminante lineal de Fisher y regresión logística. Se utilizarán parámetros hematológicos, la edad y el género de pacientes de la región de Ashanti en Ghana. Además de evaluar los cuatro modelos con fines predictivos, se seleccionará el modelo con el mejor desempeño entre los considerados según ciertos criterios estadísticos.

En el primer capítulo se abordará la teoría de los modelos lineales, que juegan un papel importante y motivador a la hora de aplicar el modelo de regresión logística (LR), naive Bayes (NB), K vecinos más cercanos (KNN) y el análisis discriminante lineal de Fisher (LDA). En el segundo capítulo se aborda un análisis descriptivo de los datos de malaria. En el tercer capítulo se muestra una ilustración de los cuatro modelos con fines predictivos y se hace una comparación de los resultados de los modelos. En el cuarto capítulo se presentan las conclusiones y recomendaciones. La escritura de este trabajo se realizará mediante la herramienta Overleaf (Overleaf, 2023).

A continuación, se introducirá la estimación de máxima verosimilitud, un método esencial para obtener estimadores en modelos estadísticos, destacando su aplicación en la regresión lineal. En el capítulo también se profundizará en el teorema de Bayes, fundamental para entender el enfoque bayesiano y sus aplicaciones en modelos como el discriminante lineal de Fisher, regresión logística y Naive Bayes.

En el apartado dedicado al análisis discriminante lineal de Fisher, se explorarán sus objetivos principales: la separación o discriminación de grupos y la predicción o asignación de objetos a categorías específicas. Se integrará el teorema de Bayes para calcular probabilidades condicionales y se presentará una detallada derivación matemática.

El modelo de regresión logística ocupará un lugar destacado, especialmente en el contexto de estudios de casos de malaria. Se describirá cómo este modelo se utilizará para calcular probabilidades condicionales, que asignarán valores cualitativos a los resultados de las pruebas de malaria y utilizarán parámetros hematológicos como variables predictoras.

Naive Bayes emergerá como una opción viable en situaciones donde la estimación de una distribución sea desafiante debido a la falta de datos. El capítulo expondrá la aplicación del clasificador Naive Bayes, que asumirá independencia entre variables y proporcionará un equilibrio práctico entre sesgo y varianza.

Finalmente, se explorará el método de los K vecinos más cercanos, destacando su utilidad en la predicción de respuestas. Este enfoque se basará en la proximidad de observaciones en el espacio de características.

1. Marco teórico

En este capítulo se explicarán los métodos y técnicas para la construcción de los modelos que se aplicarán a lo largo de los otros capítulos. Primero, se precisarán las funciones de cada modelo con el objetivo de conocer los alcances que tendrán a lo largo de este capítulo.

Luego, se consideran metodologías importantes como el modelo de regresión lineal, el teorema de Bayes, regresión lineal múltiple, estimación de máxima verosimilitud, vector de medias y matriz de covarianzas. Este material es considerado relevante para el desarrollo y aplicación del modelo de regresión logística, el K vecinos más cercanos, el naive Bayes y el discriminante lineal de Fisher a lo largo de este trabajo.

1.1. Regresión lineal múltiple

El análisis de regresión es una técnica estadística para modelar la relación que podría existir entre variables de respuesta y covariables o regresores.

En un modelo donde interviene más de un regresor, se fórmula el modelo de regresión lineal múltiple y se escribe de acuerdo con la siguiente expresión:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon, \tag{1-1}$$

donde y es llamada variable respuesta, x_1, x_2,..., x_p son los regresores en el modelo, β_0, β_1,..., β_p, son los parámetros del modelo y ϵ es el término de error aleatorio, donde $\epsilon \sim N(0, \sigma^2)$.

El adjetivo lineal en un modelo de regresión lineal indica que el modelo es lineal respecto a los coeficientes de regresión β_0, β_1, ..., β_p, y no respecto a los regresores o predictores x_1, x_2, ..., x_p (Wackerly et al., 2010).

En clasificación se utiliza una idea de regresión similar a esta, pero la respuesta es categórica (cualitativa) (Montgomery et al., 2012).

En un modelo de regresión lineal, se tienen los siguientes supuestos:

1. **Supuesto de media cero**: En cualquier conjunto de valores de $x_1, x_2, ..., x_p$, la media de los valores poblacionales es igual a cero.

2. **Supuesto de varianza constante:** Para cualquier combinación de valores de $x_1, x_2, ..., x_p$, la varianza de la población de los posibles valores del término de error no depende de

la combinación de valores de $x_1, x_2, ..., x_p$. En otras palabras, las diferentes poblaciones de valores potenciales del término de error que corresponden a las distintas combinaciones de valores de $x_1, x_2, ..., x_p$ tienen varianzas equivalentes. Denotamos la varianza constante como σ^2.

3. **Supuesto de normalidad:** Para cualquier conjunto de valores dados de $x_1, x_2, ..., x_p$, la población de los valores del término de error se distribuye normal.

4. **Supuesto de independencia:** Cualquier término de error ϵ es independiente de cualquier otro valor de ϵ (Bowerman et al., 2009).

1.2. Estimación de máxima verosimilitud

De la misma manera que en un modelo de regresión lineal simple, se muestra que los estimadores de máxima verosimilitud son los mismos estimadores de mínimos cuadrados, siempre y cuando los errores en el modelo sean normales e independientes. El modelo en forma matricial es:

$$y = X\beta + \epsilon,$$

donde $\epsilon \sim N(0, \sigma^2 I)$. La función de densidad normal es

$$f(\epsilon_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}\epsilon_i^2\right).$$

La densidad conjunta de ϵ_1, ϵ_2, ..., ϵ_n, escrita como $\prod_{i=1}^{n} f(\epsilon_i)$, es la función de verosimilitud y está dada por:

$$L(\beta|X) = \prod_{i=1}^{n} f(\epsilon_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}\epsilon_i^2\right) = (2\pi)^{-n/2}\,\sigma^{-n} \exp\left(-\frac{1}{2\sigma^2}\epsilon^T\epsilon\right),$$

reemplazando $y - X\beta = \epsilon$, la función de verosimilitud transformada es:

$$L(\beta|X) = (2\pi)^{-n/2}\,\sigma^{-n} \exp\left[-\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)\right].$$

Ahora, tomando el logaritmo natural a ambos miembros, se tiene que:

$$\ln L(\beta|X) = -\frac{n}{2}\ln(2\pi) - n\ln(\sigma) - \frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta).$$

En la expresión anterior, σ tiene un valor máximo si se minimiza $(y - X\beta)^T(y - X\beta)$, lo que indica que se puede considerar (Montgomery et al., 2012).

$$(y - X\beta)^T(y - X\beta) = 0.$$

La solución de máxima verosimilitud para el vector β es:

$$\hat{\beta} = \frac{y}{X} = \frac{X^T y}{X^T X} = (X^T X)^{-1} X^T y.$$

Si derivamos el $\ln L(\beta|X)$ respecto a σ^2 e igualamos a cero, se tiene:

$$\frac{\partial}{\partial \sigma^2} \ln L(\beta|X) = -\frac{n}{2\sigma^2} + \frac{2}{(\sigma^2)^2}(y - X\beta)^T(y - X\beta) = 0,$$

entonces,

$$\frac{n}{2\sigma^2} = \frac{2}{(\sigma^2)^2}(y - X\beta)^T(y - X\beta),$$

simplificando queda,

$$n(\sigma^2) = (y - X\beta)^T(y - X\beta),$$

donde,

$$\hat{\sigma}^2 = \frac{(y - X\hat{\beta})^T(y - X\hat{\beta})}{n}.$$

1.3. Regla de Bayes

La estadística bayesiana representa un conjunto de herramientas de gran utilidad en el análisis de datos obtenidos experimentalmente en diversas situaciones de ciencias, ingeniería, entre otras.

Antes de comenzar a hablar de la regla de Bayes, abordaremos la regla de probabilidad total.

En el siguiente enunciado, presentamos la regla de probabilidad total, que es de gran utilidad a la hora de calcular probabilidades usando la regla de Bayes. Si $\{B_1, B_2, ..., B_k\}$ representan una partición del espacio muestral S, tal que $P(B_i) \neq 0$ para $i = 1, 2, ..., k$, entonces, para un evento cualquiera $A \subset S$ (Walpole et al., 2012),

$$P(A) = \sum_{i=1}^{k} P(B_i \cap A) = \sum_{i=1}^{k} P(B_i)P(A|B_i)$$

El teorema de Bayes establece lo siguiente:

Si $\{B_1, B_2, ..., B_k\}$ constituyen una partición del espacio muestral S, donde $P(B_i) \neq 0$ para $i = 1, 2, ..., k$, entonces, para un evento cualquiera A contenido en S, tal que $P(A) \neq 0$ (Walpole et al., 2012),

$$P(B_r|A) = \frac{P(B_r \cap A)}{\sum_{i=1}^{k} P(B_i \cap A)} = \frac{P(B_r)P(A|B_r)}{\sum_{i=1}^{k} P(B_i)P(A|B_i)} = \frac{P(A|B_r)P(B_r)}{P(A)}.$$

Con esta expresión, se puede obtener una fórmula para la probabilidad condicional $P(Y = 1|X)$ que se conoce como discriminante lineal de Fisher (LDA), que se muestra más adelante, y para el método Naive Bayes.

Para este trabajo sobre estudios de casos de malaria en la región de Ashanti-Ghana, es de gran importancia el uso de la regla de Bayes al calcular probabilidades condicionales mediante el análisis discriminante lineal de Fisher y el naive Bayes.

1.4. Modelos considerados

En este apartado se presenta aspectos del modelo de regresión logística, el K vecinos más cercanos, naive Bayes y discriminante lineal de Fisher.

1.4.1. Modelo de regresión logística

El modelo de regresión logística, miembro de la familia GLM (generalized linear model), permite modelar situaciones donde la variable respuesta no necesariamente sigue una distribución normal ni es continua. En los modelos lineales generalizados, la variable respuesta solo necesita pertenecer a la familia exponencial, que incluye las distribuciones normal, de Poisson, binomial, exponencial, gamma, Erlang, Weibull, exponencial generalizada, Pareto y Rayleigh. Por otro lado, el modelo lineal con error normal representa solo un caso del modelo lineal generalizado (Montgomery et al., 2012).

En esta tesis, el modelo de regresión logística se usó para calcular las probabilidades de que un paciente tenga o no el parásito de la malaria, en función de las variables glóbulos blancos, plaquetas, linfocitos, recuentos mixtos de células, neutrófilos, la edad y el género, entre otras. Debido a que los resultados de las pruebas de malaria son datos cualitativos, a estos datos se les asignaron los valores 0 y 1, donde a un paciente que tenga el parásito por malaria se le asignó el valor 1; en caso contrario, se le asignó el valor 0.

Los parámetros hematológicos, edad y género representaron las variables en el modelo, donde estas variables se incluyeron mediante el vector de p componentes aleatorias $X = (X_1, X_2, ..., X_p)$.

El modelo de regresión logística, para calcular la probabilidad condicional de que un paciente tenga malaria o no, dadas las covariables de interés, está dado por la ecuación (Díaz et al., 2018).

$$P(Y = 1|X = x) = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)},$$

donde $P(Y = 1|X = x)$ representa la probabilidad condicional de que una persona tenga el parásito de la malaria ($Y = 1$) dadas las covariables proporcionadas en $X = (X_1, X_2, ..., X_p)$.

1.4.2. Análisis discriminante lineal de Fisher

El análisis discriminante lineal de Fisher tiene dos objetivos principales. Por un lado, está la separación o discriminación de grupos, y por otro lado, está la predicción o asignación de un objeto en uno de varios grupos definidos, con base en las variables que lo identifican (Johnson and Wichern, 2007).

Al igual que en el modelo de regresión logística, se utilizó el análisis discriminante lineal de Fisher como otra herramienta para estudiar las probabilidades de que un paciente tenga malaria o no.

En el análisis discriminante lineal de Fisher, fue fundamental el uso de la regla de Bayes, ya que la probabilidad de que un paciente tenga o no el parásito de la malaria se calcula mediante esta regla. La probabilidad de que un paciente tenga malaria, usando el teorema de Bayes para clasificación, se obtiene de la siguiente manera y de acuerdo con el teorema de Bayes,

$$P(Y = 1|X = x) = \frac{P(X = x|Y = 1)P(Y = 1)}{P(X = x)}. \tag{1-2}$$

Por el teorema de probabilidad total, y definiendo $\pi_\ell = P(Y = \ell)$ y $f_\ell(x) = P(X = x|Y = \ell)$, donde $\ell \in \{0, 1\}$, se tiene que:

$$\begin{aligned} P(X = x) &= P(X = x \wedge Y = 1) + P(X = x \wedge Y = 0) \\ &= P(Y = 1)P(X = x|Y = 1) + P(Y = 0)P(X = x|Y = 0) \\ &= \pi_1 f_1(x) + \pi_0 f_0(x) \\ &= \sum_{\ell=0}^{1} \pi_\ell f_\ell(x). \end{aligned}$$

Por lo tanto, el resultado anterior se puede escribir como,

$$P(X = x) = \sum_{\ell=0}^{1} \pi_\ell f_\ell(x). \tag{1-3}$$

También, se tiene que,

$$P(X = x|Y = 1)P(Y = 1) = f_1(x)\pi_1. \tag{1-4}$$

Reemplazando (1-3) y (1-4) en (1-2), se tiene:

$$P(Y = 1|X = x) = \frac{f_1(x)\pi_1}{\sum_{\ell=0}^{1} \pi_\ell f_\ell(x)} \tag{1-5}$$

Donde, por ejemplo, $X = (X_1, X_2, ..., X_p)$ representa el vector aleatorio de parámetros hematológicos, la edad y el género mencionados anteriormente. π_k es la probabilidad de que un paciente al que se le toma la muestra sea del grupo de los que tienen malaria o de los que no tienen el parásito. La probabilidad para cualquiera de los dos grupos está dada por $\pi_k = n_k/n$, con $k = 0, 1$.

El vector que tiene como variables aleatorias los parámetros hematológicos, la edad y el género se asume que tiene una distribución normal multivariada, lo cual se puede expresar como $X \sim N_p(\mu, \Sigma)$. Aquí, $E(X) = \mu$ es la media del vector X y $Cov(X) = \Sigma$ es la matriz de covarianzas $p \times p$ de X. La función de densidad normal multivariada se puede escribir como (James et al., 2021):

$$f(x) = \frac{1}{(2\pi)^{p/2}|\Sigma_1|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right] \tag{1-6}$$

Ahora, si se reemplaza la expresión (1-6) en (1-5), se tiene:

$$P(Y=1|X=x) = \frac{\pi_1 \frac{1}{(2\pi)^{p/2}|\Sigma_1|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)\right]}{\sum_{\ell=0}^{1} \pi_\ell \frac{1}{(2\pi)^{p/2}|\Sigma_\ell|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_\ell)^T\Sigma_\ell^{-1}(x-\mu_\ell)\right]}$$

Asumiendo varianza constante en los dos grupos relacionados con malaria, $\Sigma_0 = \Sigma_1 = \Sigma$, se tiene:

$$P(Y=1|X=x) = \frac{\pi_1 \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right]}{\sum_{\ell=0}^{1} \pi_\ell \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_\ell)^T\Sigma^{-1}(x-\mu_\ell)\right]}.$$

Para las k clases, sea $f_k(x) = P(X = x|Y = k)$ y π_k es la probabilidad previa de que una observación elegida aleatoriamente provenga de la k-ésima clase. Interesa la siguiente probabilidad:

$$P_k(x) = P(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{\ell=0}^{1} \pi_\ell f(x)}, \text{ por el teorema de Bayes.}$$

Se clasificará una observación en la clase para la cual $P_k(x)$ es mayor. El clasificador de Bayes asigna una observación $X = x$ a la clase para la cual $P_k(x)$ es la mayor probabilidad. Ahora, suponga que

$$f_k(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k)\right]$$

donde $\Sigma_0 = \Sigma_1 = \Sigma$, entonces

$$P_k(x) = \frac{\pi_k \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k)\right]}{\sum_{\ell=0}^{k} \pi_\ell \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_\ell)^T\Sigma^{-1}(x-\mu_\ell)\right]}$$

Un sujeto $X = x$ podría clasificarse en cualquiera de las k categorías. Para hacerlo, debemos calcular todas las probabilidades $P_1(x), P_2(x), ..., P_k(x)$ y luego seleccionar la mayor. Supongamos que $P_k(x)$ es la mayor probabilidad entre $\{P_1(x), P_2(x), ..., P_k(x)\}$, entonces:

$$P_k(x) \geq P_{k^*}(x), \ \forall \ k \neq k^*$$

Por lo tanto,

$$\ln P_k(x) \geq \ln P_{k^*}(x)$$

$$\ln[P_k(x)] = \ln \left\{ \frac{\pi_k \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k)\right]}{\sum_{\ell=0}^{K} \pi_\ell \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_\ell)^T\Sigma^{-1}(x-\mu_\ell)\right]} \right\}$$

$$= \ln \left\{ \left[\pi_k \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \right] \exp\left[-\frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k) \right] \right\}$$
$$- \ln \left\{ \left[\sum_{\ell=0}^{K} \pi_\ell \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \right] \exp\left[-\frac{1}{2}(x-\mu_\ell)^T\Sigma^{-1}(x-\mu_\ell) \right] \right\}$$

$$= \ln(\pi_k) + \ln\left[\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \right] - \frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k)$$
$$- \ln \left\{ \left[\sum_{\ell=0}^{K} \pi_\ell \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \right] \exp\left[-\frac{1}{2}(x-\mu_\ell)^T\Sigma^{-1}(x-\mu_\ell) \right] \right\}$$

$$= \ln(\pi_k) - \frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_k) - \ln \left\{ \left[\sum_{\ell=0}^{K} \pi_\ell \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \right] \exp\left[-\frac{1}{2}(x-\mu_\ell)^T\Sigma^{-1}(x-\mu_\ell) \right] \right\}$$
$$+ \ln\left[\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \right]$$

$$= \ln(\pi_k) - \frac{1}{2}x^T\Sigma^{-1}x + x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k$$
$$- \ln \left\{ \left[\sum_{\ell=0}^{K} \pi_\ell \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \right] \exp\left[-\frac{1}{2}(x-\mu_\ell)^T\Sigma^{-1}(x-\mu_\ell) \right] \right\} + \ln\left[\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \right]$$

$$= x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \ln(\pi_k) - \ln \left\{ \left[\sum_{\ell=0}^{K} \pi_\ell \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \right] \exp\left[-\frac{1}{2}(x-\mu_\ell)^T\Sigma^{-1}(x-\mu_\ell) \right] \right\}$$
$$+ \ln\left[\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \right] - \frac{1}{2}x^T\Sigma^{-1}x$$

$$= x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \ln(\pi_k) - C$$

$$= \delta_k(x) - C, \quad \text{donde } C \text{ es una constante.}$$

Análogamente,

$$\ln P_{k^*}(x) = \delta_{k^*} - C.$$

Por otro lado, $P_k(x) \geq P_{k^*}(x)$ es equivalente a,

$$\delta_k(x) - C \geq \delta_{k^*}(x) - C,$$

donde,

$$\delta_k(x) \geq \delta_{k^*}(x),\ \forall\ k \neq k^*.$$

1.4.3. Naive Bayes

Dado que estimar una distribución requiere una gran cantidad de datos, utilizar naive Bayes se vuelve una buena opción en una amplia gama de situaciones. Fundamentalmente, el supuesto de naive Bayes conduce a cierto sesgo, pero disminuye la varianza, lo que lleva a un clasificador que funciona adecuadamente en la práctica, manteniendo un buen equilibrio entre varianza y sesgo (James et al., 2021).

Recuerde el teorema de Bayes teniendo en cuenta los casos de malaria:

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

donde,

- $P(Y = 1|X)$: es la probabilidad de estar en la clase 1, dado el perfil hematológico de los pacientes X.
- $P(X|Y = 1)$: es la probabilidad de observar el vector de perfiles hematológicos X, dado que está en la clase 1 (malaria).
- $P(Y = 1)$: probabilidad previa de que ocurra la clase 1 (malaria).
- $P(X)$: probabilidad previa de observar el vector de perfiles hematológicos X.

El clasificador de naive Bayes para los casos de malaria asume que las componentes del vector de hematología X son independientes y que son igualmente importantes. Con este supuesto, la probabilidad de observar el vector hematológico $X = (X_1, ..., X_p)$ dado que pertenece a la clase de los que tienen malaria es:

$$P(X|Y = 1) = P(X_1|Y = 1) \times P(X_2|Y = 1) \times \cdots \times P(X_p|Y = 1).$$

$P(X_i|Y=1)$ es la probabilidad de que la clase 1 genere el vector observado para las componentes del vector hematológico i, para $i = 1, 2, ..., p$.

Teniendo en cuenta que el vector de hematología $X = (X_1, X_2, ..., X_p)$, el naive Bayes se puede aplicar usando:

$$\begin{aligned} P(Y=1|X) &= \frac{P(X|Y=1)P(Y=1)}{P(X)} \\ &= \frac{P(X_1|Y=1) \times P(X_2|Y=1) \times \cdots \times P(X_p|Y=1) \times P(Y=1)}{P(X)} \\ &= \frac{\prod_{i=1}^{p} P(X_i|Y=1) \times P(Y=1)}{P(X)} \propto \prod_{i=1}^{p} P(X_i|Y=1) \times P(Y=1). \end{aligned}$$

La ecuación anterior se utilizó para clasificar a los pacientes en uno de los dos grupos, es decir, con malaria o sin la enfermedad.

1.4.4. K vecinos más cercanos

Frecuentemente, se quiere predecir respuestas aplicando el clasificador o regla de Bayes, pero al utilizar datos reales, se desconoce la distribución condicional de Y dado X, lo que se vuelve difícil en la práctica. Algunos enfoques buscan estimar la distribución de Y dado X y luego clasificar una observación en un grupo o clase con probabilidades grandes. Un método relevante es el de K vecinos más cercanos (KNN), donde para un entero positivo K y una observación de prueba x, el método KNN identifica primero los K puntos más cercanos a x en los datos de entrenamiento que se representan con $\mathcal{N}_0$, en este caso los datos de entrenamiento corresponden a 1660 observaciones de malaria, lo que equivale al 80 % de las 2076 observaciones. Luego se estima la probabilidad para la clase j teniendo en cuenta el conjunto de test o de validación, en este caso el conjunto de validación corresponde a 416 observaciones de malaria, lo que equivale al 20 % de las 2076 observaciones totales, mediante la siguiente expresión (James et al., 2021):

$$P(Y=j|X=x) = \frac{1}{K}\sum_{i \in \mathcal{N}_0} I(y_i = j),\ j \in \{0,1\}. \tag{1-7}$$

La probabilidad de clasificación para cualquiera de los dos grupos de malaria, para los K vecinos más cercanos a x, está dada por la ecuación (1-7), donde $\mathcal{N}_0$ denota el conjunto de puntos del grupo de malaria más cercanos a x, e $I(y_i = j)$ se define como:

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

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

Por último, el KNN clasifica la observación de prueba x al grupo con la mayor probabilidad $P(Y = j|X = x)$.

Para hallar las K distancias más cercanas al punto x, se utiliza la distancia de Euclides, que se expresa como sigue: dados dos puntos en $\mathbb{R}^p$, con $X_h = (X_{h1}, ..., X_{hp})$ y $X_i = (X_{i1}, ..., X_{ip})$, se define la distancia euclidiana como (Monroy and Rivera, 2012):

$$d_{hi} = \left[\sum_{j=1}^{p}(X_{hj} - X_{ij})^2\right]^{1/2}. \tag{1-8}$$

Para la ecuación (1-8), h e i en X representan dos observaciones cualquiera del conjunto de datos de malaria. Esta ecuación se utilizó para hallar las distancias de todas las observaciones relacionadas con malaria respecto a una observación x que se clasificó en uno de los dos grupos de malaria.

En este capítulo, se exploraron diversos métodos y técnicas fundamentales para la construcción de modelos que se aplicarán a lo largo de la tesis. En primer lugar, se abordaron los conceptos esenciales de la regresión lineal múltiple, que constituye una herramienta clave para modelar la relación entre variables de respuesta y covariables o regresores. Se destacaron las suposiciones asociadas con este modelo, como el supuesto de media cero, la varianza constante, la normalidad y la independencia de los errores.

Posteriormente, se realizó un breve resumen de la estimación de máxima verosimilitud como un enfoque para encontrar los parámetros óptimos en modelos de regresión lineal. Se describió el proceso de maximizar la función de verosimilitud para obtener estimadores eficientes, especialmente relevantes en el contexto de la regresión lineal múltiple.

A continuación, se introdujo la regla de Bayes como un marco teórico fundamental para el análisis estadístico, destacando la regla de probabilidad total como un concepto clave. Se exploró el teorema de Bayes, fundamental para el desarrollo de modelos de clasificación, como el discriminante lineal de Fisher y el naive Bayes.

En la sección de modelos, se presentaron tres enfoques específicos que se aplicarán en el estudio de casos de malaria: el modelo de regresión logística, el análisis discriminante lineal de Fisher y el naive Bayes. Se detallaron las ecuaciones y supuestos asociadas con cada uno de estos modelos, destacando su relevancia en el contexto de la clasificación de pacientes con o sin malaria.

Finalmente, se mencionó el método de K vecinos más cercanos como otra herramienta valiosa para la clasificación, basado en la idea de asignar a un punto la clase predominante entre sus vecinos más cercanos. Cada uno de estos métodos se aplicará y comparará en el estudio de casos de malaria en la región de Ashanti-Ghana, contribuyendo así a un análisis integral y robusto de la problemática abordada.

En el capítulo 2, se llevará a cabo un análisis descriptivo centrado en un conjunto de datos provenientes del Departamento de Laboratorio del Hospital St.Patrick's en la región de Ashanti, Ghana. La información recopilada se abarca desde enero de 2018 hasta junio de 2018 e incluye datos hematológicos, edad y género de más de dos mil pacientes que se sometieron a pruebas de malaria.

En el siguiente capítulo, también se presentará una descripción detallada de las covariables que se utilizarán en los modelos, destacando variables clave como la edad, hemoglobina, plaquetas y linfocitos. Además, se llevará a cabo un análisis descriptivo mediante el uso de tablas y gráficos para visualizar la distribución de estas variables entre los pacientes diagnosticados con malaria y aquellos que no presentaron la enfermedad.

Asimismo, se explorarán las relaciones entre variables mediante técnicas como diagramas de caja y bigotes, y matrices de dispersión. La visualización de la distribución de las variables en función del género y la edad también se abordará, proporcionando una visión integral de la diversidad de datos y posibles patrones.

El análisis detallado sentará las bases para comprender la relevancia de cada covariable en la predicción de la presencia de malaria, permitiendo así la identificación de factores clave y la mejora de las estrategias de diagnóstico y prevención de esta enfermedad.

2. Análisis descriptivo de los datos de malaria

Comprender la malaria a nivel mundial y nacional es crucial para los gobiernos transitorios, ya que a través de políticas de estado se pueden implementar prácticas que contrarresten la propagación de esta enfermedad en todo el mundo, especialmente en aquellas naciones o países donde la malaria es endémica debido a su clima tropical, ubicación geográfica y situación de pobreza extrema. En esta sección, se realizará una descripción detallada de las covariables o regresores en el modelo, acompañada de un análisis descriptivo mediante el uso de tablas y gráficos. Todo ha sido mediante una base de datos proporcionada por cortesía del científico de laboratorio Ellis Kobina Paintsil, datos que se obtuvieron del Departamento de laboratorio del Hospital St.Patrick's de Ghana.

2.1. Fuentes de información y análisis descriptivo

2.1.1. Fuentes de información

Para calcular las probabilidades utilizando los modelos de regresión logística, naive Bayes, K vecinos más cercanos y discriminante lineal de Fisher mediante el software estadístico R (R Core Team, 2023) y RStudio (RStudio Team, 2023), se emplearon datos del laboratorio departamental del Hospital de St. Patrick's en la región de Ashanti, Ghana. Los datos fueron recopilados desde enero de 2018 a junio de 2018 mediante la revisión de los registros hematológicos, que incluyeron información sobre análisis de sangre y pruebas de malaria (Paintsil et al., 2019). De entre estos datos, se seleccionaron las variables hematológicas, edad y género, para incorporarlas en los modelos, con el propósito de determinar las probabilidades de que un paciente diagnosticado con malaria realmente padeciera la enfermedad. El objetivo principal fue demostrar la capacidad predictiva de los modelos en la identificación del parásito causante de la malaria en pacientes diagnosticados por el personal de salud. En consecuencia, estos modelos pueden ser utilizados con frecuencia en diagnósticos futuros de malaria.

El número de datos utilizados en la aplicación de los modelos fue 2076, recopilados de pacientes en la región de Ashanti. De estos, 1200 correspondieron a mujeres y 876 a hombres, cuyas edades oscilan entre 1 y 102 años. El diagnóstico para determinar la presencia del parásito de la malaria en los pacientes se llevó a cabo mediante pruebas de diagnóstico rápido (PDR) y se confirmó utilizando la técnica del método Giemsa, realizada por expertos microscopistas. Los análisis de hemogramas se llevaron a cabo utilizando el analizador automático Mindray

BC 3000plus (Paintsil et al., 2019).

En específico, las covariables usadas en este trabajo son las siguientes:

- Edad (Age) en años cumplidos
- Género (Sex), donde 0 = Mujer y 1 = Hombre
- Hemoglobina (Hb)
- Glóbulos blancos (Wbc)
- Plaquetas (Plt)
- Linfocitos (Lymph)
- Recuentos mixtos de células (Mxd)
- Neutrófilos (Neut)

2.1.2. Análisis descriptivo

Según la figura 2-1, de un total de 2076 pacientes sometidos a la prueba de malaria, 539, equivalentes al 25.95 %, dieron positivo para la enfermedad. En distintos grupos de edad, se observó que los niños menores de 5 años presentaron la mayor prevalencia de malaria, alcanzando el 29.98 % (206 de 687), seguidos por los niños con edades entre 5 y 14 años, con un 45.30 % (164 de 362). Por otro lado, las edades que registraron menor prevalencia de malaria fueron: entre 15 y 24 años, con un 29.6 % (69 de 302), y mayores de 24 años, con un 13.8 % (100 de 725), siendo los pacientes entre 15 y 24 años los que tuvieron la menor prevalencia.

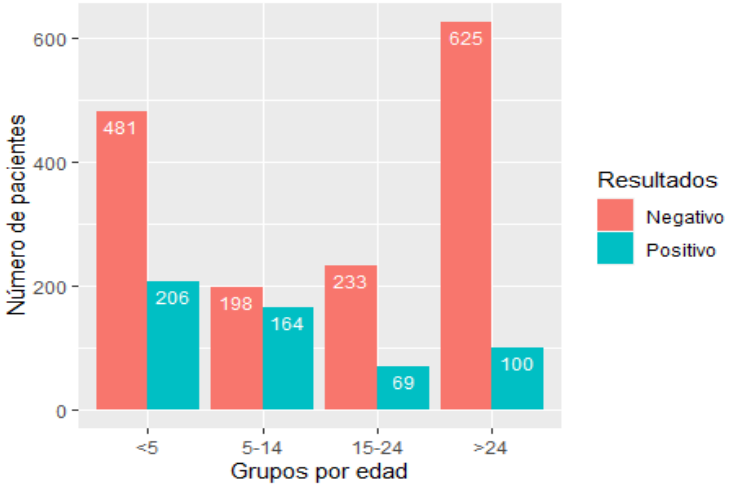

Figura 2-1.: Gráfico de barras por grupos de edad para los casos de malaria en la región de Ashanti. Fuente: Elaboración propia.

En el gráfico de cajas y bigotes, figura 2-2a, se observa que la variable hemoglobina (Hb) para los pacientes que tuvieron malaria varió entre 3.10 y 16.7 g/dL (gramos por decilitros), y el 50 % central de los pacientes tuvo la hemoglobina entre 8.50 y 11.70 g/dL, existiendo algunos valores de la hemoglobina anormalmente grandes, ya que estos valores aparecen alineados con los bigotes superior e inferior. La distribución para la Hb de los pacientes que tuvieron malaria es asimétrica en la parte inferior de la caja, porque la zona de abajo en el área central de la caja es mayor que la parte superior, y la mediana corresponde al valor de 10.20 g/dL, siendo la media de 10.02 g/dL. Observando el diagrama para los pacientes que no tuvieron malaria, se puede ver que la hemoglobina varía entre 2.5 y 17.60 g/dL, y el 50 % central de los pacientes tuvo la hemoglobina entre 10.10 y 12.50 g/dL, donde de la misma manera se observaron algunos valores anormales de la hemoglobina. La distribución para la Hb de los pacientes que no tuvieron la enfermedad es simétrica, ya que la zona inferior y superior de la caja tienen casi igual área, con una mediana de 11.30 g/dL, siendo la media de 11.24 g/dL.

En la variable hematológica plaquetas (Plt) para los pacientes que tuvieron malaria, figura 2-2b, se puede observar que varió entre 17 y 780 G/L (G = giga y L = litro), y el 50 % central de los pacientes tuvo las plaquetas entre 103.5 y 221.5 G/L, existiendo algunos valores de las plaquetas anormales en el bigote de la parte superior de la caja. La distribución para las plaquetas de los pacientes que tuvieron malaria es simétrica, ya que la parte superior e inferior de la caja tienen igual área, con una mediana correspondiente a 156 G/L, siendo la media de 174.8 G/L. Para los pacientes que no tuvieron la enfermedad, se puede ver que las plaquetas tuvieron una variación entre 34 y 941 G/L, y el 50 % central de los pacientes tuvo las plaquetas entre 195 G/L y 352 G/L con algunos valores anormales en el bigote superior de la caja. La distribución para la variable Plt de los pacientes que no tuvieron la enfermedad es asimétrica en la parte superior de la caja, ya que la zona superior tiene mayor área que la de abajo, con una mediana de 2059 G/L, siendo la media de 284.6 G/L.

Para la variable hematológica linfocito (Lymph), figura 2-2c, en los pacientes que tuvieron malaria, varió entre 6.10 y 84.60 %, y el 50 % central de los pacientes tuvo los linfocitos entre 17.40 y 38.95 %, existiendo algunos valores de los Lymph anormales; exactamente dos valores son anormales, como se observa en el bigote de la parte superior de la caja. La distribución para los Lymph en los pacientes que tuvieron malaria es simétrica, ya que la parte superior e inferior de la caja tienen casi igual área, con una mediana que corresponde a 27.8 %, siendo la media de 29.16 %. Para los pacientes que no tuvieron la enfermedad, se puede ver que los Lymph tuvieron una variación entre 4.3 y 80.7 %, y el 50 % central de los pacientes tuvo los linfocitos entre 22.5 % y 49.9 %, sin ningún valor anormal. La distribución para la variable Lymph de los pacientes que no tuvieron la enfermedad es asimétrica en la parte inferior de la caja, ya que la parte inferior de la caja tiene mayor área que la de arriba, con mediana de 34.6 %, siendo la media de 34.7 %.

Para la variable edad (Age), figura 2-2d, en los pacientes que tuvieron malaria varió entre 1 y 94 años, y el 50 % central de los pacientes tuvo edades entre 3 y 20 años, existiendo algunos valores anormales de la edad en el bigote de la parte superior de la caja. La distribución para las edades de los pacientes que tuvieron malaria es asimétrica en la parte superior de la

caja, ya que la zona de arriba en el área centrada de la caja es mayor que la parte inferior, con una mediana correspondiente a 7 años, siendo la media de 13.9 años. Para los pacientes que no tuvieron la enfermedad, se puede ver que las edades tuvieron variabilidad entre 1 y 102 años, y el 50 % central de los pacientes tuvo edades entre 3 y 35 años con algunos valores anormales de las edades. La distribución para la variable edad de los pacientes que no tuvieron la enfermedad es asimétrica en la parte inferior de la caja, con una mediana de 20 años, siendo la media 22.67 años. Los niños a quienes se les diagnosticó con más frecuencia la enfermedad tenían aproximadamente 1 año.

Para las tablas que se observarán a continuación, la tabla 2-1 muestra un resumen de las medidas para el Mínimo, Q_1, Q_2, Q_3 y el Máximo para el grupo de los que tuvieron malaria, y la tabla 2-2 muestra un resumen para el grupo de los que no tuvieron la enfermedad.

Variables	Mín.	1^{er} cuantil.	Mediana.	Media.	3^{er} cuantil.	Máx.
Hemoglobina (Hb)	3.10	8.50	10.20	10.02	11.70	16.70
Plaquetas (Plt)	17.00	103.50	156.00	174.80	221.50	780.00
Linfocitos (Lymph)	6.10	17.40	27.80	29.16	38.95	84.60
Edad (Age)	1.00	3.00	7.00	13.88	19.50	94.00

Tabla 2-1.: Valores para el Mínimo, Q_1, Q_2, Q_3 y el Máximo, para las variables del grupo con malaria. Fuente: Elaboración propia.

Variables	Mín.	1^{er} cuantil.	Mediana.	Media.	3^{er} cuantil.	Máx.
Hemoglobina (Hb)	2.50	10.10	11.30	11.24	12.50	17.60
Plaquetas (Plt)	34.00	195.00	259.00	284.60	352.00	941.00
Linfocitos (Lymph)	4.30	22.50	34.60	34.70	45.90	80.70
Edad (Age)	1.00	3.00	20.00	22.67	35.00	102.00

Tabla 2-2.: Valores para el Mínimo, Q_1, Q_2, Q_3 y el Máximo, para las variables del grupo sin malaria. Fuente: Elaboración propia.

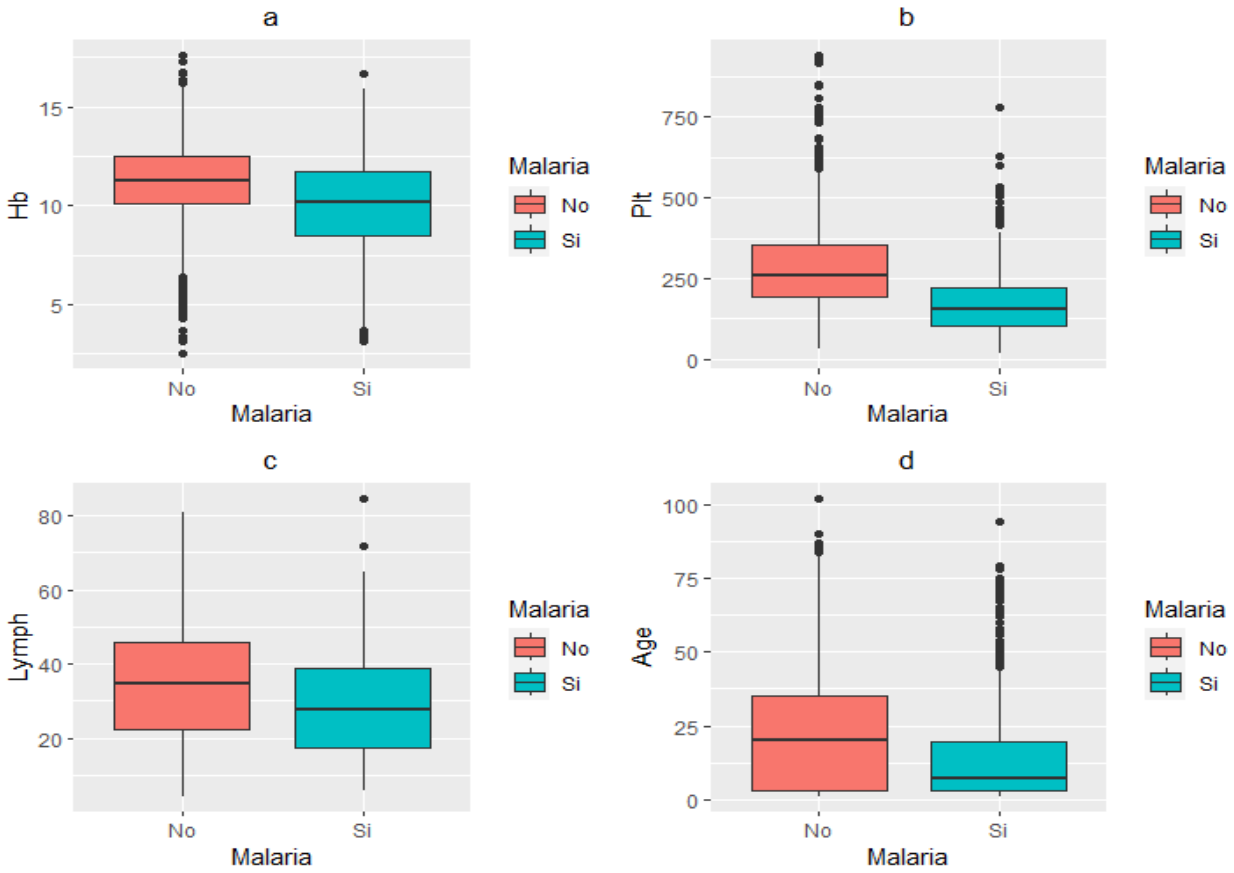

Figura 2-2.: Gráfico de caja y bigote de las variables Hb, Plt, Lymph y Age para los pacientes relacionados con malaria. Fuente: Elaboración propia.

Para la matriz de dispersión o figura 2-3, se puede observar en la nube de puntos de la entrada (2,1) una correlación débil, es decir, la correlación entre la variable Age y Hb, es débil y corresponde al valor 0.273; además, se puede observar en la nube de puntos que no es fácil ajustar una línea recta. Para la nube de puntos en los pacientes que no tuvieron malaria y los que tuvieron la enfermedad, se observa que las correlaciones también son débiles en las variables Age y Hb con valores de 0.195 y 0.390. Los puntos rojos corresponden a los pacientes que no tuvieron malaria y los verdes a los pacientes que tuvieron la enfermedad. Esto parece indicar niveles de colinealidad bajos que no comprometen seriamente el proceso de estimación.

En la nube de puntos de la entrada (4,2), se puede observar una correlación nula, lo que quiere decir que no hay correlación entre la variable hemoglobina (Hb) y plaquetas (Plt), y su valor correspondiente es de -0.033; además, no es fácil ajustar una línea recta a la nube de puntos. Para la nube de puntos en los pacientes que no tuvieron malaria y los que tuvieron el parásito, se observa que las correlaciones son débiles con valores de -0.220 y 0.118.

Observando la nube de puntos de la entrada (3,2), se puede observar una correlación casi nula, lo que muestra que no hay correlación entre las variables Hb y Lymph, y su valor corresponde a -0.050. Para la nube de puntos, se observa que no es fácil ajustar una línea recta. En la nube de puntos para los pacientes que no tuvieron malaria y los que fueron positivos para la enfermedad, se observó correlación negativa, nula y moderada con valores de -0.006 y -0.313. Las interpretaciones para las nubes de puntos de las entradas (3,1), (4,1), y (4,2) respectivamente son similares a las anteriores.

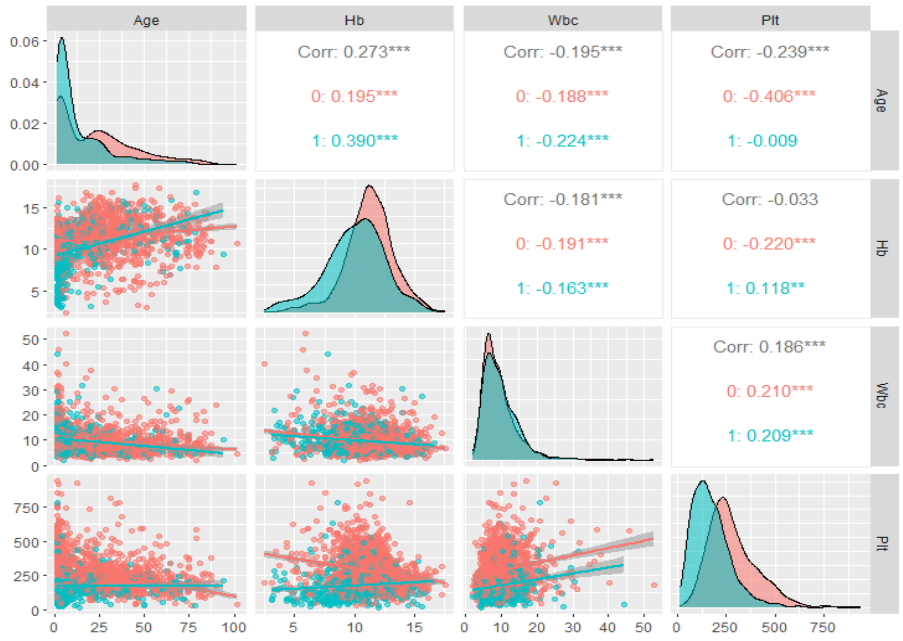

Figura 2-3.: Matriz de dispersión para los datos de malaria. Fuente: Elaboración propia.

En el gráfico de densidad para la hemoglobina, figura 2-4, se observa que los datos están bastante dispersos respecto a la media para ambos grupos, es decir, tanto en aquellos que tienen malaria como en los que no tienen la enfermedad. Aunque en el grupo de los que tienen malaria, la dispersión es ligeramente mayor y con una centralidad destacada en los datos. En los pacientes que no tuvieron malaria, se aprecia una mayor concentración en hemoglobinas específicas en comparación con las hemoglobinas de los pacientes que tuvieron la enfermedad; esto se debe a que la gráfica de densidad para los pacientes que no tuvieron malaria tiene el pico más alto que en los pacientes que tuvieron la enfermedad. Al analizar el QQ plot para la hemoglobina en ambos grupos de malaria, se nota que los puntos no están tan alejados de las líneas rectas diagonales, lo que indica que los datos para la hemoglobina parecen provenir de una distribución normal.

Para las plaquetas en la figura 2-4, se observan los datos bastante dispersos respecto a la media para ambos grupos; sin embargo, el grupo de aquellos que no tuvieron malaria muestra una mayor dispersión que el grupo de los que estuvieron enfermos con malaria. Además, notamos que la gráfica de densidad para las plaquetas en los pacientes que tuvieron la enfermedad tiene el pico más alto que en los pacientes que no tuvieron malaria, indicando una mayor concentración en plaquetas específicas en el grupo afectado por la malaria. En el QQ plot para las plaquetas en los dos grupos de malaria, se aprecia que una buena parte de los datos está desalineada con respecto a las líneas diagonales para ambos grupos, lo que sugiere que los datos parecen no provenir de una distribución normal.

Observando el gráfico de densidad para los linfocitos, figura 2-5, se nota que su dispersión es similar en el grupo de aquellos que tuvieron malaria y en el grupo de los que no tuvieron la enfermedad. En el gráfico de densidad para los linfocitos, se pudo observar que los linfocitos tienen el pico más alto en los pacientes que tuvieron malaria, indicando una mayor concentración de linfocitos específicos en el grupo que tuvo la enfermedad en comparación con el grupo que no tuvo malaria. Observando los QQ plot para los linfocitos en los dos grupos de malaria, se aprecia que la mayoría de los puntos no están tan desalineados de las líneas diagonales, lo que muestra que los datos provienen de una distribución aproximadamente normal.

Para la edad (Age), figura 2-5, se puede observar que los datos tienen poca dispersión tanto en el grupo de aquellos que no tienen malaria como en el grupo de los que tienen la enfermedad, aunque las edades para el grupo de los que no tienen malaria están ligeramente más dispersas. En el gráfico de densidad se destaca un pico más alto en las edades de los pacientes que tuvieron malaria que en aquellos que no tuvieron la enfermedad, indicando una mayor concentración de edades específicas en el grupo afectado por la malaria. Observando los QQ plot para las edades en los dos grupos de malaria, se aprecia que la mayoría de los puntos se ven desalineados de las diagonales para ambos grupos, lo que sugiere que los datos posiblemente no sean de una distribución normal.

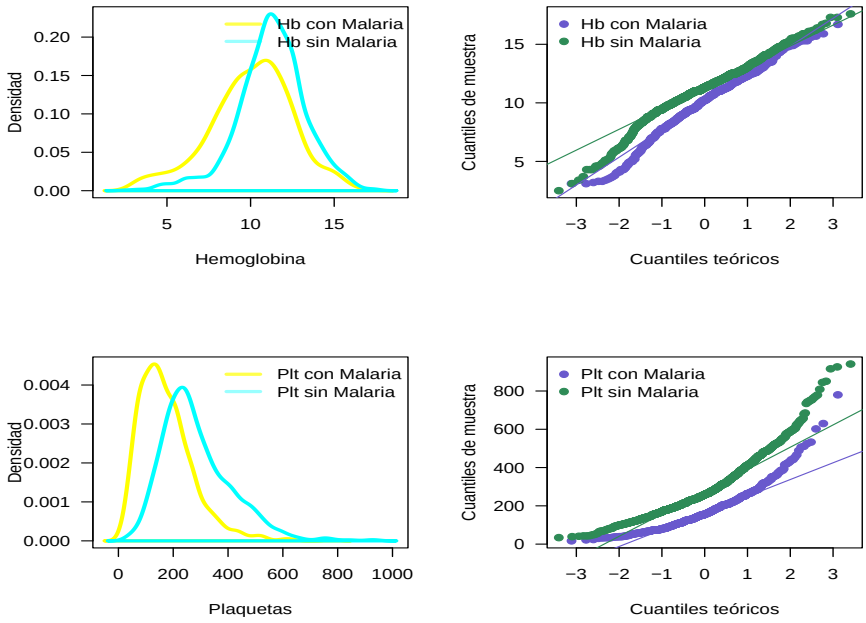

Figura 2-4.: Densidades y QQ plots de las variables Plt y Hb de acuerdo a la variable Malaria. Fuente: Elaboración propia.

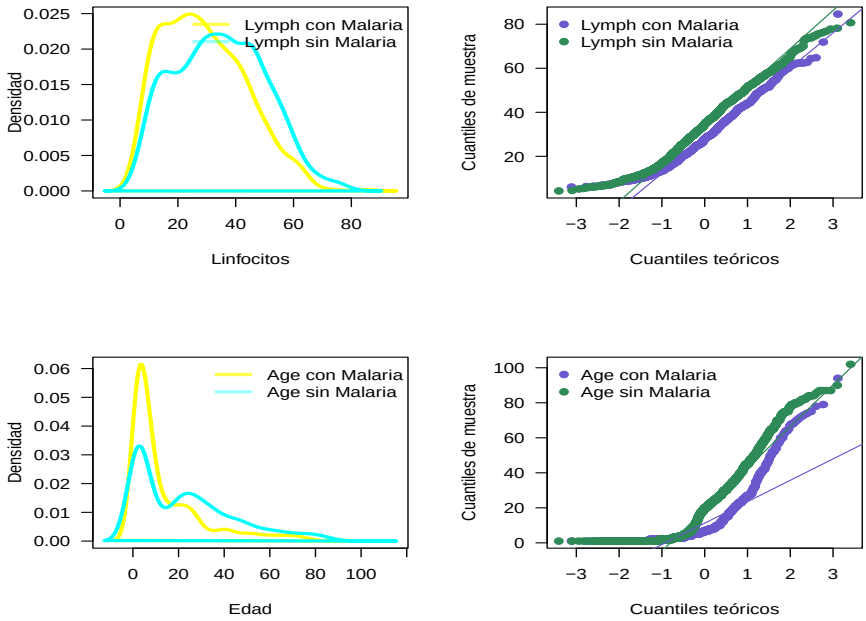

Figura 2-5.: Densidades y QQ plots de las variables Lymph y Age de acuerdo a la variable Malaria. Fuente: Elaboración propia.

En el gráfico de cajas y bigotes, figura 2-6, se observa que la variable (Age) para los pacientes del género femenino que fueron diagnosticadas con malaria tuvieron edades entre 1 y 102 años. El 50 % central de las pacientes que fueron diagnosticadas con malaria tuvieron edades entre 4 y 33 años. En el diagrama se observan algunos valores anormales. La mediana y la media para las edades de las mujeres que fueron diagnosticadas con el parásito son de 20 años y 22.65 años, respectivamente. Para el grupo de hombres que fueron diagnosticados con el parásito, se observa que las edades variaron entre 1 y 87 años. El 50 % central de los pacientes que fueron diagnosticados con malaria tuvieron edades entre 2 y 28 años, con mediana de 7 años y media de 17.28 años.

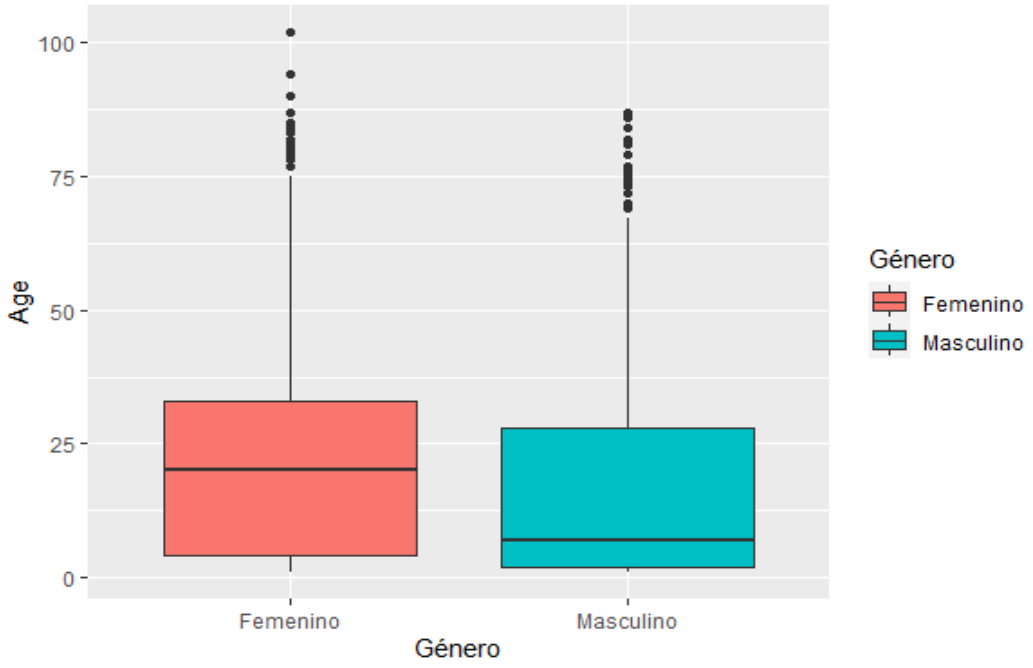

Figura 2-6.: Diagrama de cajas y bigotes por género para las edades. Fuente: Elaboración propia.

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

Tabla 2-3.: Valores para el Mínimo, Q_1, Q_2, Q_3, el Máximo y prueba de t de Student, para los diagramas de cajas y bigotes de la figura 2-6. Fuente: Elaboración propia.

Al realizar la prueba t de Student para las edades en ambos géneros y observar que el valor p es menor a 0.05, se afirma que existen diferencias significativas entre las muestras en cuanto a sus medias (consultar tabla 2-3). Es decir, las edades de las mujeres difieren en comparación

con las edades de los hombres.

Observando el gráfico de densidad, figura 2-7, para las edades en hombres y mujeres diagnosticados con malaria, se percibe que la dispersión de las edades en el género femenino es mayor que en el género masculino. Además, se nota que el gráfico de densidad para los hombres tiene un pico más alto que el de las mujeres, indicando una mayor concentración de edades específicas en los los hombres en comparación con las mujeres. En el QQ plot para las edades en ambos grupos, se observa que la mayoría de los puntos para los dos grupos no están próximos a la línea recta diagonal, lo que sugiere que los datos de las edades no provienen de una distribución normal.

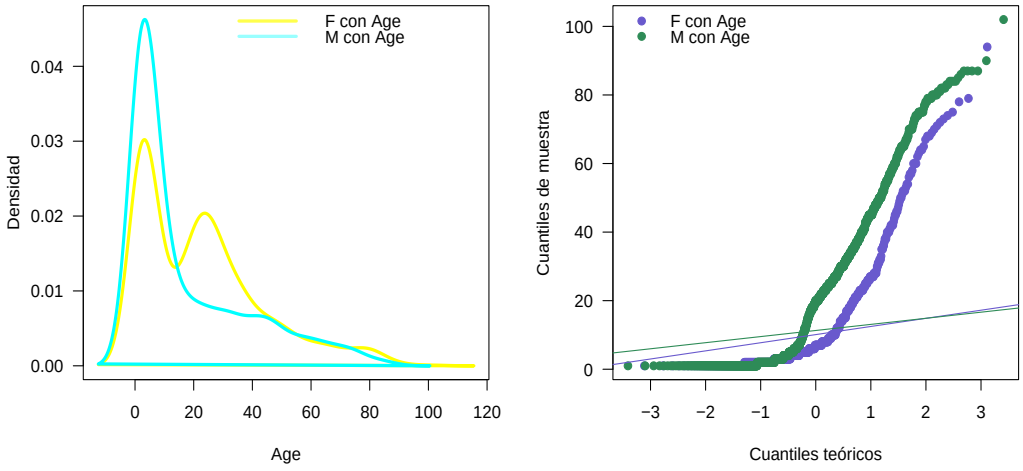

Figura 2-7.: Densidades y QQ plots de la variable Age de acuerdo a la variable género. Fuente: Elaboración propia.

En este capítulo, se llevó a cabo un análisis descriptivo detallado de los datos relacionados con la malaria, utilizando información recopilada en el Departamento de laboratorio del Hospital St. Patrick's en la región de Ashanti, Ghana. La base de datos incluyó variables hematológicas, edad y género de pacientes sometidos a pruebas de diagnóstico rápido y confirmados mediante la técnica del método Giemsa.

El análisis por género mostró diferencias en la distribución de las edades entre hombres y mujeres diagnosticados con malaria, proporcionando una perspectiva adicional sobre las características de los pacientes en función de su género.

En resumen, este análisis descriptivo de los datos de malaria proporciona una base sólida para comprender las características de los pacientes y las variables clave asociadas con la enfermedad. Estos hallazgos son fundamentales para informar estrategias de prevención y tratamiento, así como para el desarrollo y mejora de modelos de diagnóstico.

En el capítulo 3, se realizará un análisis de los resultados obtenidos a partir de la aplicación de los cuatro modelos con fines predictivos para determinar la presencia de malaria en pacientes. Los modelos que se evaluarán son: regresión logística, análisis discriminante lineal de Fisher, naive Bayes y K vecinos más cercanos.

Se describirán los pasos a seguir para la aplicación de estos modelos, incluyendo la selección de variables significativas y la división del conjunto de datos en un conjunto de entrenamiento (80 %) y un conjunto de prueba (20 %). Posteriormente, se presentarán y explicarán las ecuaciones de los cuatro modelos.

El capítulo abordará también la evaluación de la capacidad discriminativa de modelos predictivos, utilizando las covariables hemoglobina (Hb), plaquetas (Plt), linfocitos (Lymph) y edad (Age). Se analizará la capacidad de las variables para predecir la presencia del parásito de la malaria en los pacientes, evaluando áreas bajo la curva ROC, puntos de corte, sensibilidad y especificidad.

A continuación, se procederá con la interpretación de los resultados, destacando los coeficientes estimados para cada variable en los modelos que así lo permitan, así como la matriz de confusión y la curva ROC para cada uno de ellos. Se analizarán las métricas de desempeño, tales como la exactitud, sensibilidad, especificidad, valor predictivo positivo y valor predictivo negativo.

Finalmente, se llevará a cabo la comparación de las capacidades predictivas de los cuatro modelos mediante la evaluación de las áreas bajo la curva ROC, y se presentarán los errores cuadráticos medios (MSE) como medida de rendimiento.

3. Aplicación de los cuatro modelos con fines predictivos

En este capítulo, presentaremos los resultados de los modelos de K vecinos más cercanos, naive Bayes, análisis discriminante lineal de Fisher y regresión logística. Este análisis constituye el aporte más importante de este trabajo.

Para ilustrar los modelos mencionados anteriormente, se seleccionaron 1660 observaciones, lo que equivale al 80 % de 2076, como conjunto de entrenamiento (train), y 416 observaciones, que representan el 20 % restante, como muestra de prueba (test). Para predecir la presencia de malaria utilizando estos modelos, primero se ajusto cada uno de ellos con el conjunto de entrenamiento. Luego, los modelos entrenados se utilizaron para realizar predicciones con el conjunto de datos de prueba. La curva ROC es una herramienta extremadamente valiosa para evaluar y comparar el rendimiento de modelos de clasificación binaria en términos de sensibilidad y especificidad a través de varios umbrales de clasificación.

La curva ROC, como se observará en los gráficos más adelante, idealmente se sitúa en la esquina superior izquierda, indicando una alta tasa de verdaderos positivos y una baja tasa de falsos positivos. La línea punteada de 45 grados, representa el clasificador con "ninguna información" (James et al., 2021).

La curva ROC es una herramienta muy valiosa para evaluar y comparar el rendimiento de modelos de clasificación binaria en términos de sensibilidad y especificidad a través de varios umbrales de clasificación. En este estudio, se procede a construir una curva ROC de los modelos marginales asociados con las variables respuesta consideradas importantes. Al observar la curva ROC (Figura 3-1) y la Tabla 3-1, se aprecia que las plaquetas (Plt) exhibieron el área más grande bajo la curva, alcanzando 0.772, lo cual se considera buena área. A continuación, se ubicaron la hemoglobina (Hb), los linfocitos (Lymph) y la edad (Age), las cuales mostraron un área bajo la curva considerada modesta. Destaca que las variables con mayor especificidad y sensibilidad fueron la hemoglobina, con una especificidad del 0.81, y las plaquetas, con una sensibilidad del 0.71, igualando la sensibilidad de la edad.

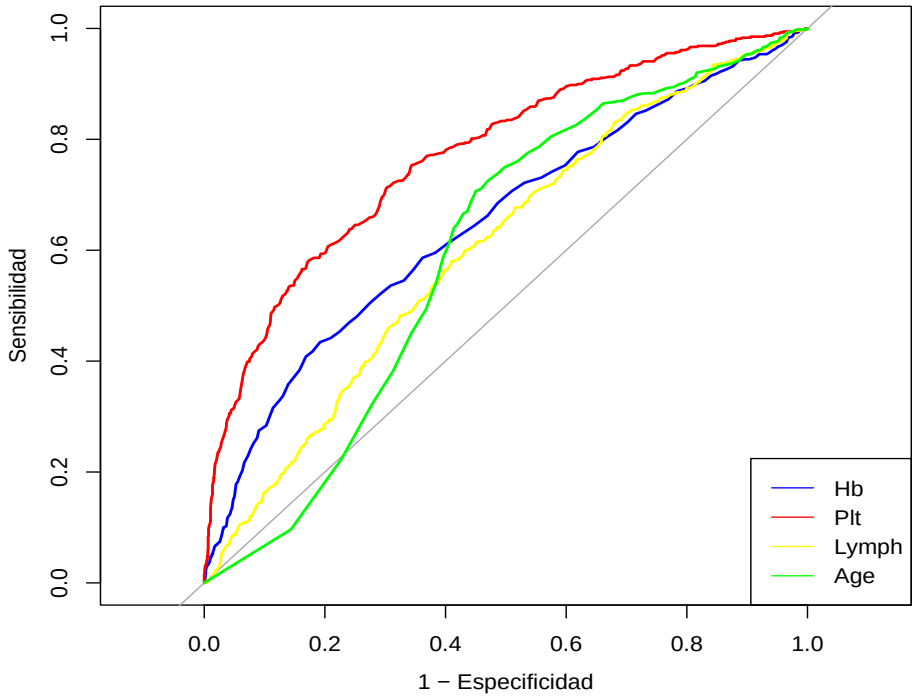

Figura 3-1.: Gráfico de la curva ROC para los predictores significativos en los modelos, basado en modelos marginales. Fuente: Elaboración propia.
AUC: Area Under Curve.

Variables	AUC	Punto de corte	Especificidad	Sensibilidad
Hemoglobina (Hb)	0.650	0.75	0.81	0.43
Plaquetas (Plt)	0.772	209.50	0.70	0.71
Linfocitos (Lymph)	0.605	30.55	0.59	0.58
Edad (Age)	0.600	15.50	0.55	0.71

Tabla 3-1.: Evaluación de la capacidad discriminativa. Fuente: Elaboración propia.

Con fines ilustrativos, se procede a construir cuatro gráficos de contorno de acuerdo al material del texto de (James et al., 2021). Estos gráficos se obtienen usando cada clasificador con 2 variables y con la ayuda de R. Se observa la naturaleza lineal y no lineal de los clasificadores.

En los gráficos de contornos con su frontera de decisión para el modelo de regresión logística (LR), figura 3-2; el K vecinos más cercanos (KNN), figura 3-3; el naive Bayes (NB), figura 3-4; y el análisis discriminante lineal (LDA), figura 3-5; observamos cómo clasifican los cuatro modelos. Los puntos azules en los gráficos representan a los pacientes que fueron diagnosticados con malaria, y los puntos naranjas representan a los pacientes que no tuvieron malaria. La línea roja es la frontera de decisión.

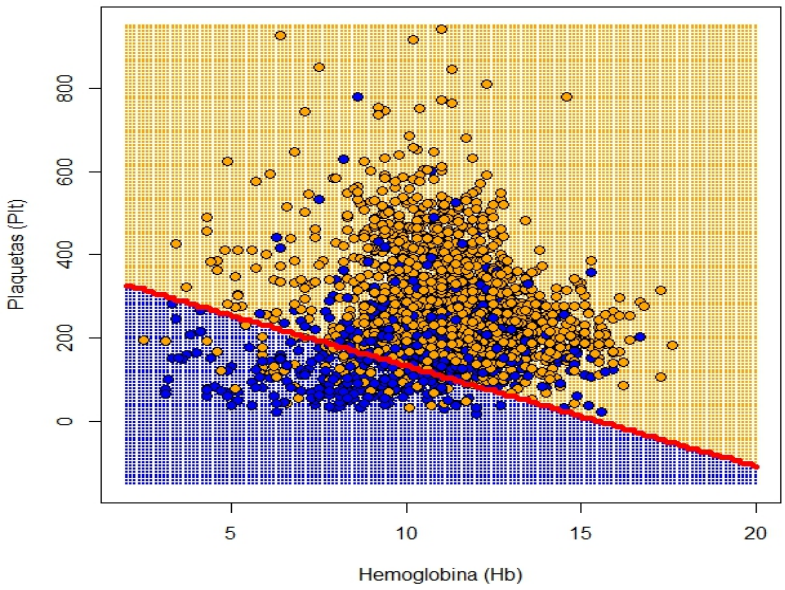

Figura 3-2.: Gráfico de contorno para LR. Fuente: Elaboración propia.

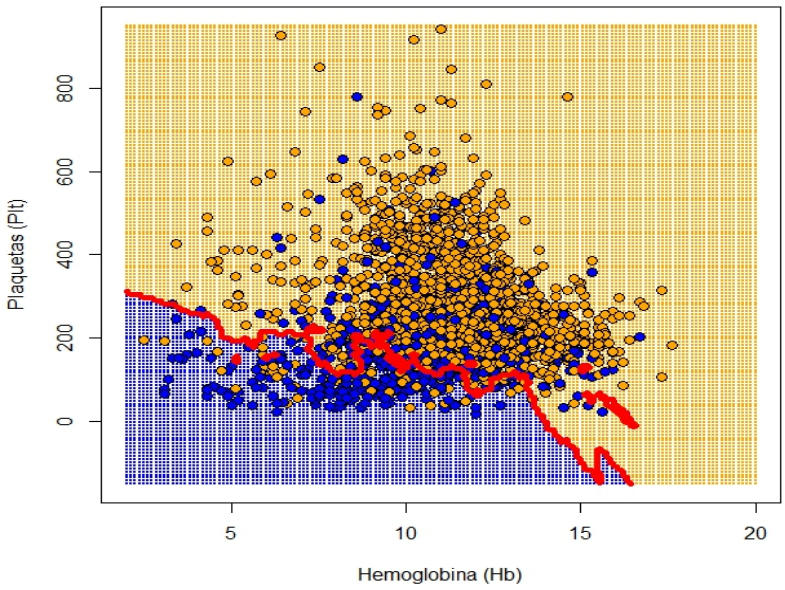

Figura 3-3.: Gráfico de contorno para KNN. Fuente: Elaboración propia.

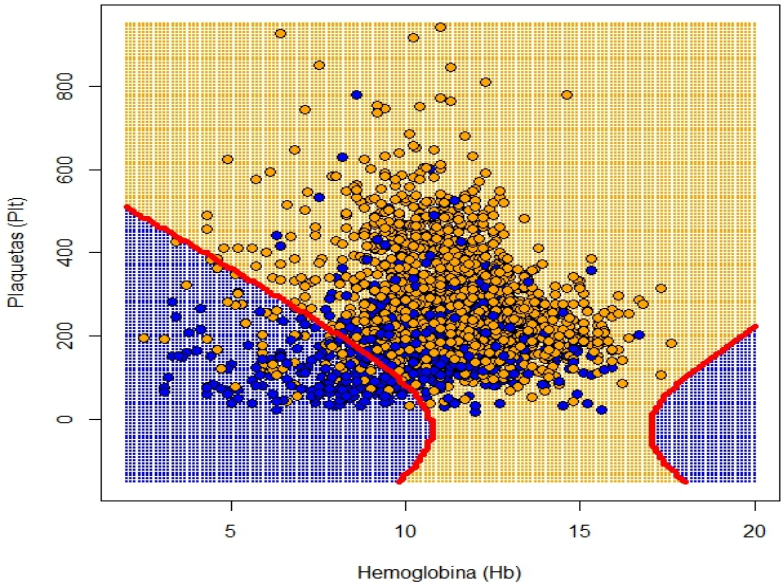

Figura 3-4.: Gráfico de contorno para NB. Fuente: Elaboración propia.

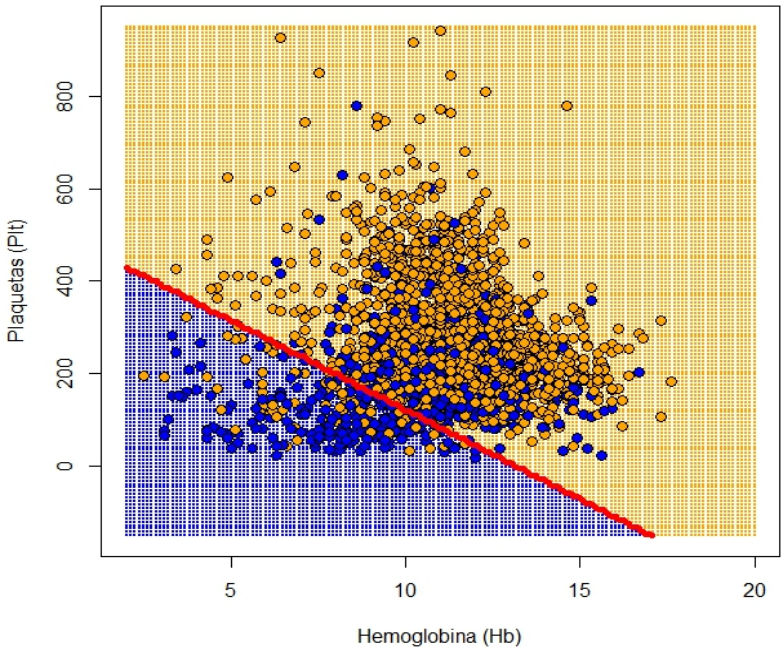

Figura 3-5.: Gráfico de contorno para LDA. Fuente: Elaboración propia.

3.1. Resultados para el modelo de regresión logística

De acuerdo a Paintsil et al. (2019) el modelo de regresión logística se ajusto utilizando las variables significativas en el modelo, que en este caso fueron la hemoglobina (Hb), las plaquetas (Plt), los linfocitos (Lymph) y la edad (Age). Por lo tanto, la ecuación del modelo de regresión logística para estas variables se expresa de la siguiente manera:

$$P(Y = 1|X = x) = \frac{e^{\beta_0+\beta_1\times Age+\beta_2\times Hb+\beta_3\times Plt+\beta_4\times Lymph}}{1 + e^{\beta_0+\beta_1\times Age+\beta_2\times Hb+\beta_3\times Plt+\beta_4\times Lymph}}.$$

Donde $P(Y = 1|X)$ es la probabilidad condicional de que una persona tenga malaria ($Y = 1$), dado el valor de las covariables denotadas como $X = (Hb, Plt, Lymph, Age)$. Por otro lado, $\beta_0, \beta_1, ..., \beta_4$ representan los parámetros en el modelo. Este modelo se ajusto usando R (R Core Team, 2023) y (RStudio Team, 2023).

3.2. Interpretación de los resultados para el modelo de regresión logística

En la tabla 3-2, la estimación del intercepto es de 5.175. El valor z de 12.507 indica cuántas desviaciones estándar se encuentra la estimación del intercepto con respecto al valor nulo. El valor p asociado es prácticamente cero, lo que sugiere una fuerte evidencia en contra de la hipótesis nula de que el coeficiente del intercepto es igual a cero.

En cuanto a la hemoglobina (Hb), su coeficiente estimado es -0.175. El valor z indica que existe evidencia significativa para rechazar la hipótesis nula de que el coeficiente es igual a cero. El signo negativo indica que a medida que aumenta Hb el logit de tener malaria disminuye.

La variable de plaquetas (Plt) muestra un coeficiente negativo significativo, indicando que al aumentar el valor de Plt, se espera una disminución en el logaritmo del odds de la variable de respuesta.

De manera similar, la variable linfocitos (Lymph) también presenta un coeficiente negativo significativo. Por último, la edad (Age) tiene un coeficiente negativo significativo, sugiriendo que a medida que la edad aumenta, la log-odds de tener malaria disminuye.

	Estimación	Error estándar	Valor z	$P(>\|z\|)$
Intercepto	5.175	0.414	12.507	$< 2e^{-16}$
Hb	-0.176	0.032	-5.564	$2.63e^{-8}$
Plt	-0.011	0.001	-14.365	$< 2e^{-16}$
Lymph	-0.032	0.001	-6.817	$9.31e^{-12}$
Age	-0.042	0.004	-9.814	$< 2e^{-16}$

Tabla 3-2.: Estimación del modelo de regresión logística LR. Fuente: Elaboración propia.

La matriz de confusión para el modelo de regresión logística presentado en la tabla 3-3 muestra a los pacientes que padecieron la enfermedad de la malaria, aquellos que no la tuvieron, y las predicciones del modelo para ambos grupos en relación con la enfermedad. En la tabla, se observa que del total de pacientes que tuvieron el parásito, en este caso, 118, el modelo clasifica a 47 con malaria, lo que equivale al 39.83 % del total, y a 71 sin malaria, lo que representa el 60.17 %. En cuanto a los pacientes que no tuvieron la enfermedad, que fueron 298 en total, el modelo predijo que 16 personas sí la tuvieron, equivalente al 5.37 %, y 282 personas que no la tuvieron, representando el 94.63 %, donde cero (0) indica los pacientes sin la enfermedad y uno (1) a los pacientes que la tuvieron. Teniendo en cuenta que la exactitud es la suma de observaciones que el modelo predice correctamente para ambos grupos, dividido por el total de observaciones, en este caso, $(282 + 47)/416$, lo cual equivale al 79.1 %. El modelo es capaz de clasificar correctamente el 79.1 % de las observaciones, ya sea con malaria o sin la enfermedad.

	Predichos		
Observados	Sin paludismo (0)	Con paludismo (1)	Total
Sin paludismo (0)	282	16	298
Con paludismo (1)	71	47	118
Total	353	63	416

Tabla 3-3.: Matriz de confusión para el modelo de regresión logística. Fuente: Elaboración propia.

Para el modelo de regresión logística se construye la curva ROC junto a su AUC representada en la figura 3-6. El área bajo la curva ROC es del 81.5 %, lo que indica una discriminación considerada buena. Es importante destacar que la diagonal, también conocida como la línea de no discriminación, divide el espacio de la curva ROC de manera porcentual.

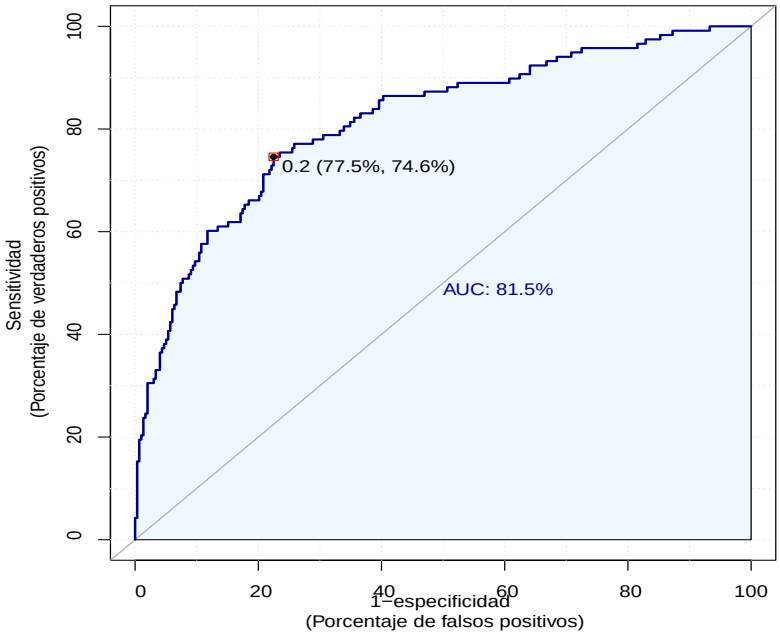

Figura 3-6.: Curva ROC y AUC para el modelo de regresión logística. Fuente: Elaboración propia.

En la tabla 3-4 se observa que aproximadamente el 79.1 % de las predicciones fueron precisas, con un intervalo de confianza de (74.9 %, 82.9 %). En relación con la tasa de no información, en este caso, si el modelo no proporcionara información, acertaría aproximadamente el 84.9 %. En este contexto, el valor de kappa es 0.401, lo que indica un nivel moderado. La tasa de verdaderos positivos, también conocida como sensibilidad, alcanzó el 74.6 %. La especificidad, o tasa de verdaderos negativos, que fueron correctamente identificados por el modelo, fue del 79.9 %. La proporción de casos con malaria predichos correctamente con respecto al total de casos positivos para malaria fue del 39.8 %, mientras que la proporción de casos sin la enfermedad predichos correctamente con respecto al total de casos sin malaria fue del 94.6 %.

La prevalencia de casos con malaria reales en la muestra se situó en un 15.1 %, y la tasa de detección del modelo fue del 11.3 %. Además, el modelo identificó como malaria aproximadamente el 28.4 % de la muestra.

La exactitud equilibrada (Balanced Accuracy), que considera tanto la sensibilidad como la especificidad, alcanzó un 77.2 %. Estos resultados ofrecen una visión integral del rendimiento del modelo de regresión logística en términos de precisión y capacidad para prever tanto casos positivos como negativos.

Accuracy	0.791	Pos Pred Value	0.398
95 % CI	(0.749, 0.829)	Neg Pred Value	0.946
No Information Rate	0.849	Prevalence	0.151
Kappa	0.401	Detection Rate	0.113
Sensitivity	0.746	Detection Prevalence	0.284
Specificity	0.799	Balanced Accuracy	0.772

Tabla 3-4.: Resumen de los estadísticos para la matriz de confusión del modelo LR. Fuente: Elaboración propia.

3.3. Resultados para el análisis discriminante lineal de Fisher

En el análisis discriminante lineal de Fisher, se aplicó utilizando las variables hemoglobina (Hb), plaquetas (Plt), linfocitos (Lymph) y edad (Age), en el siguiente modelo:

$$P(Y=1|X=x)=\frac{\pi_1 \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right]}{\sum_{\ell=0}^{1}\pi_\ell \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_\ell)^T\Sigma^{-1}(x-\mu_\ell)\right]},$$

donde $P(Y=1|X=x)$ es la probabilidad de que un paciente tenga el parásito dado el vector $X=(Hb, Plt, Lymph, Age)$.

$\pi_k = n_k/n$ con $k \in \{0,1\}$, representa las probabilidades de los dos grupos, es decir, del grupo de pacientes que tuvo malaria y el grupo que no tuvo la enfermedad. Σ es la matriz de varianzas-covarianzas de los grupos, y μ_ℓ con $\ell \in \{0,1\}$, es el vector de medias para los grupos. Por lo tanto, Σ y μ_ℓ se escriben como:

$$E(X_\ell)=\begin{bmatrix} E(Age_\ell) \\ E(Hb_\ell) \\ E(Plt_\ell) \\ E(Lymph_\ell) \end{bmatrix} = \begin{bmatrix} \mu_{\ell 1} \\ \mu_{\ell 2} \\ \mu_{\ell 3} \\ \mu_{\ell 4} \end{bmatrix} = \mu_\ell$$

y

$$\Sigma = E(X_\ell-\mu_\ell)(X_\ell-\mu_\ell)' = Cov(X_\ell) = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44} \end{bmatrix}$$

3.4. Interpretación de los resultados para el discriminante lineal de Fisher

Para el análisis discriminante lineal de Fisher, en la matriz de confusión presentada en la tabla 3-5, se observa que de un total de 118 pacientes que tuvieron la enfermedad por ma-

laria, el modelo clasificó a 51 pacientes con la enfermedad, lo que equivale al 43.2 %, y a 67 personas sin paludismo, equivalente al 56.8 %. Sin embargo, para los pacientes que no tuvieron la enfermedad, que fueron un total de 298, el modelo clasificó a 15 con malaria, lo que equivale al 5.03 % de 298, y a 283 sin malaria, con un 94.97 %. El modelo es capaz de clasificar correctamente $(283 + 51)/416$ que equivale al 80.3 % de las observaciones con malaria o sin la enfermedad.

	Predichos		
Observados	Sin malaria (0)	Con malaria (1)	Total
Sin malaria (0)	283	15	298
Con malaria (1)	67	51	118
Total	350	66	416

Tabla 3-5.: Matriz de confusión para el análisis discriminante lineal de Fisher. Fuente: Elaboración propia.

En el análisis discriminante lineal de Fisher, en la figura 3-7, al observar que el área bajo la curva ROC es del 81.3 %, se concluye que la discriminación es buena.

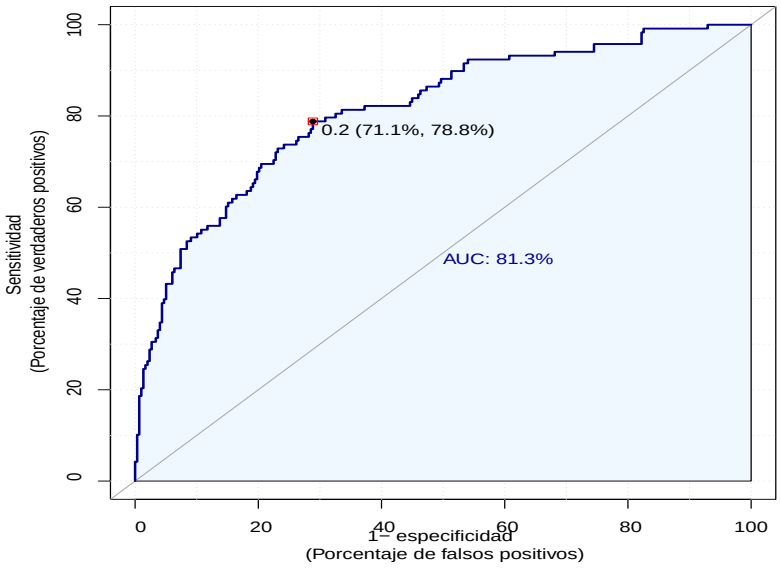

Figura 3-7.: Curva ROC y AUC para el análisis discriminante linear de Fisher. Fuente: Elaboración propia.

Para el análisis discriminante lineal de Fisher (LDA), en la tabla 3-6 que se muestra a continuación, podemos observar las probabilidades para cada grupo de pacientes a los que se les realizo la prueba de malaria. La probabilidad de que un paciente pertenezca al grupo de los que no tienen la enfermedad (cero) fue del 75 %, mientras que la probabilidad de que un paciente sea del grupo de enfermos por malaria (uno) fue del 25 %.

Clase	Sin malaria (0)	Con malaria (1)
Probabilidad	0.75	0.25

Tabla 3-6.: Probabilidades previas por grupos para el LDA. Fuente: Elaboración propia.

La tabla 3-7 muestra los promedios para cada variable en cada grupo de pacientes evaluados para la enfermedad de la malaria.

	Hb	Plt	Lymph	Age
0	11.249	281.628	34.842	22.858
1	9.996	170.803	29.116	13.615

Tabla 3-7.: Promedios para las variables por clase para el LDA. Fuente: Elaboración propia.

En la tabla 3-8, se muestran los coeficientes de discriminación lineal para el discriminante lineal de Fisher. En el caso del coeficiente -0.182 asociado a la hemoglobina (Hb), el valor negativo indica que a medida que el nivel de Hb disminuye, la función discriminante lineal también disminuye. Esto sugiere una relación inversa entre la hemoglobina y la variable que se está discriminando.

Para las plaquetas, el coeficiente es -0.007, y al igual que en el caso anterior, un coeficiente negativo indica que a medida que los valores de las plaquetas (Plt) disminuyen, la probabilidad estimada con la función discriminante lineal también disminuye.

En cuanto al coeficiente -0.023 asociado a los linfocitos (Lympl), de manera análoga, la disminución en los valores de los linfocitos se asocia con una reducción en la función discriminante lineal.

En el caso de la edad, el coeficiente es -0.029. Aquí, el coeficiente negativo sugiere que a medida que la edad disminuye, la probabilidad estimada con la función discriminante lineal también disminuye.

Variables	LD1	Variables	LD1
Hb	-0.182	Lymph	-0.023
Plt	-0.007	Age	-0.029

Tabla 3-8.: Coeficientes de discriminación para el modelo de Fisher. Fuente: Elaboración propia.

En la tabla 3-9, en relación con los casos de malaria, se observa que la precisión, que es la proporción aproximada de predicciones correctas en relación con el total de observaciones, fue de 0.803. En este caso, alrededor del 80.3 % de las predicciones son correctas.

En relación con la tasa de no información, en este caso, si el modelo no proporcionara información, acertaría aproximadamente el 84.1 %

El valor de kappa de 0.440 indica una concordancia moderada entre las predicciones y las observaciones reales. La tasa de verdaderos positivos es de 0.773, lo cual sugiere que alrededor del 77.3 % de los casos de malaria se identificaron correctamente. La proporción de casos negativos correctamente identificados fue del 0.809, indicando que aproximadamente el 80.9 % de los casos negativos se identificaron correctamente.

La proporción de predicciones positivas que son verdaderamente positivas en este caso fue de 0.432, lo que significa que alrededor del 43.2 % de las predicciones con malaria son correctas. La proporción de predicciones negativas que son verdaderamente negativas es de 0.950, indicando que aproximadamente el 95.0 % de las predicciones sin malaria son correctas.

La prevalencia es del 0.159, lo que representa la proporción de la clase con la enfermedad. En este caso, alrededor del 15.9 % de las observaciones pertenecen a la clase con malaria. La proporción de casos positivos correctamente identificados en relación con el total de casos positivos es de 0.123, lo que significa que aproximadamente el 12.3 % de los casos con malaria se detectan correctamente. La proporción de casos detectados en relación con el total de observaciones es de 0.284, indicando que alrededor del 28.4 % de las observaciones se clasifican con la enfermedad.

La precisión equilibrada, que es el promedio de sensibilidad y especificidad, es de 0.791, sugiriendo que el modelo tiene alrededor del 79.1 % de precisión equilibrada en términos de rendimiento general, teniendo en cuenta ambas clases.

Accuracy	0.803	Pos Pred Value	0.432
95 % CI	(0.761, 0.840)	Neg Pred Value	0.950
No Information Rate	0.841	Prevalence	0.159
Kappa	0.440	Detection Rate	0.123
Sensitivity	0.773	Detection Prevalence	0.284
Specificity	0.809	Balanced Accuracy	0.791

Tabla 3-9.: Resumen de los estadísticos para la matriz de confusión del LDA. Fuente: Elaboración propia.

3.5. Resultados para el naive Bayes

Teniendo en cuenta el vector de hematología y edad $X = (Hb, Plt, Lymph, Age)$, para el naive Bayes, tenemos:

$$P(Y=1|X) = \frac{P(X|Y=1)P(Y=1)}{P(X)}$$
$$\approx \frac{P(Hb|Y=1) \times P(Plt|Y=1) \times P(Lymph|Y=1) \times P(Age|Y=1) \times P(Y=1)}{P(X)}$$

donde

- $P(Y=1|X)$: Es la probabilidad de estar en la clase 1, dado $X = (Hb, Plt, Lymph, Age)$.
- $P(X|Y=1)$: Es la probabilidad de observar el vector $X = (Hb, Plt, Lymph, Age)$.
- $P(Y=1)$: Es la probabilidad para el grupo de malaria teniendo en cuenta los parámetros del vector X.
- $P(X)$: Es la probabilidad de observar $X = (Hb, Plt, Lymph, Age)$.

Con la ecuación anterior se clasificaron pacientes con malaria en uno de los dos grupos, usando NB.

3.6. Interpretación de los resultados para el naive Bayes

En la matriz de confusión presentada en la tabla 3-10, observamos que el naive Bayes, para un total de 118 pacientes que tuvieron malaria, el modelo predijo a 45 pacientes con el parásito, equivalente al 38.1 %, y a 73 pacientes sin la enfermedad, representando el 61.9 % restante. De un total de 298 pacientes sin la enfermedad, el modelo clasificó a 19 con malaria y a 279 sin la enfermedad, lo que equivale al 6.38 % y 93.62 %, respectivamente. El modelo es capaz de clasificar correctamente $(279 + 45)/416$ que equivale al 77.9 % de las observaciones con malaria o sin la enfermedad.

	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 3-10.: Matriz de confusión para el naive Bayes. Fuente: Elaboración propia.

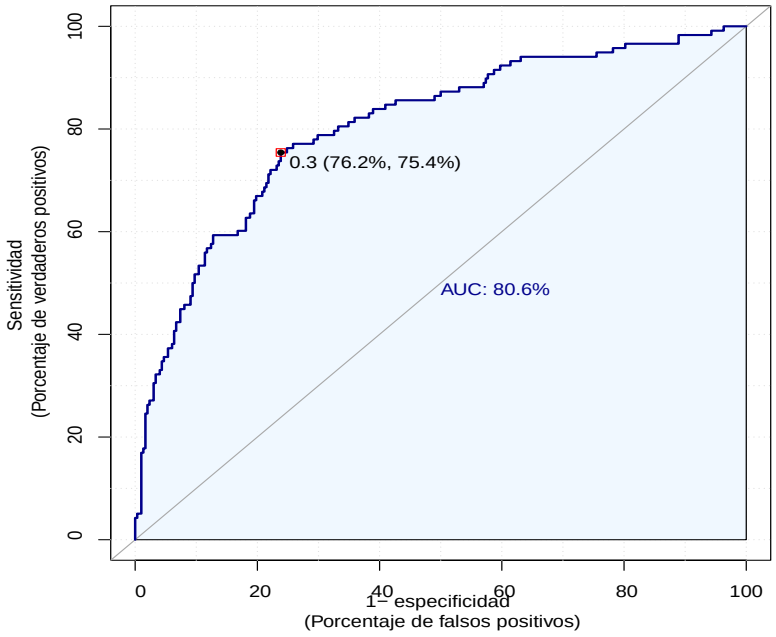

Figura 3-8.: Curva ROC y AUC para el clasificador naive Bayes. Fuente: Elaboración propia.

En el gráfico anterior, figura 3-8, para el naive Bayes (NB), al observar que el área bajo la curva AUC es del 80.6 %, se concluye que la discriminación es buena.

En la tabla 3-11 se presenta un resumen de los estadísticos para la matriz de confusión correspondiente al modelo de naive Bayes (Tabla 3-10). En relación con la exactitud o precisión, que representa 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 73.6 % y 81.8 %. La tasa de no información para los pacientes sin malaria fue del 84.6 %.

El coeficiente kappa, ajustado por azar y con un valor de 0.369, refleja una concordancia moderada. La sensibilidad, que es la proporción de pacientes con malaria correctamente clasificados (también conocida como tasa de verdaderos positivos), alcanza el 70.3 %. La especificidad, o tasa de verdaderos negativos, que fueron correctamente identificados por el modelo, fue del 79.3 %. En cuanto al valor predictivo positivo, que es la proporción de pacientes clasificados con malaria que son realmente positivos para malaria, se observa un valor del 38.1 %, indicando una confiabilidad relativamente baja en las clasificaciones de pacientes con malaria.

El valor predictivo para pacientes sin malaria, que mide la proporción de pacientes clasificados como negativos para la enfermedad que son realmente negativos, alcanza un 93.6 % que es muy alto. La prevalencia, que es la proporción de pacientes con malaria, es del 15.4 %. La tasa de detección, que representa la proporción de pacientes con malaria correctamente identificados, se sitúa en el 10.8 %. La prevalencia de detección (Detection Prevalence),

que indica la proporción de pacientes clasificados con malaria, es del 28.4 %. Finalmente, la exactitud equilibrada, obtenida como el promedio de la sensibilidad y la especificidad, es del 74.8 %.

Accuracy	0.779	Pos Pred Value	0.381
95 % CI	(0.736, 0.818)	Neg Pred Value	0.936
No Information Rate	0.846	Prevalence	0.154
Kappa	0.369	Detection Rate	0.108
Sensitivity	0.703	Detection Prevalence	0.284
Specificity	0.793	Balanced Accuracy	0.748

Tabla 3-11.: Resumen de los estadísticos para la matriz de confusión del NB. Fuente: Elaboración propia.

3.7. Resultados para el K vecinos más cercanos

Para hallar las K distancias más cercanas a un punto que se clasificó a uno de los dos grupos de malaria, se tomaron 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 calcula la distancia euclidiana como

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

Para la expresión anterior, h e i en X representan dos observaciones cualquiera distintas, del conjunto de datos de malaria.

3.8. Interpretación de los resultados para el K vecinos más cercanos

En el modelo de K vecinos más cercanos, la matriz de confusión de la tabla 3-12 muestra que, de un total de 64 pacientes que tuvieron la enfermedad de la malaria, el modelo predice el parásito en 46 pacientes, equivalente al 71.9 %, mientras que el 28.1 % de los 18 pacientes restantes no fueron clasificados como positivos para la enfermedad. En cuanto a los pacientes que no tuvieron la enfermedad (352 pacientes en este caso), el modelo predijo la enfermedad en 72 pacientes con una probabilidad del 20.5 %, y para los 280 pacientes restantes, el modelo no predijo la enfermedad con una probabilidad del 79.5 %. El modelo es capaz de clasificar correctamente $(280 + 46)/416$ que equivale al 78.4 % de las observaciones con malaria y sin la enfermedad.

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

Tabla 3-12.: Matriz de confusión para el K vecinos más cercanos. Fuente: Elaboración propia.

Observando el gráfico, figura 3-9, en el caso de los K vecinos más cercanos (KNN), el área bajo la curva AUC es del 70.3 %, indicando una discriminación moderada.

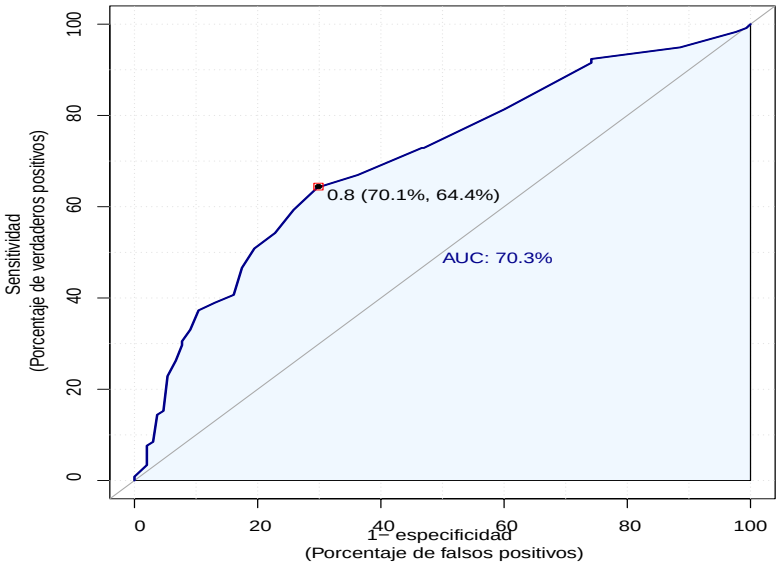

Figura 3-9.: Curva ROC y AUC para K vecinos más cercanos. Fuente: Elaboración propia.

En la tabla 3-13 se presenta un resumen de los estadísticos para la matriz de confusión correspondiente al método de los K vecinos más cercanos, para la tabla 3-12. En este escenario, el modelo exhibe una exactitud del 78.4 %, lo que implica que aproximadamente el 78.4 % de las predicciones son correctas. Además, la exactitud del modelo se sitúa entre el 74.1 % y el 82.2 % con un nivel de confianza del 95 %.

En relación con la tasa de no información, en este caso, si el modelo no proporcionara información, acertaría aproximadamente el 71.6 %. El valor de kappa de 0.382 indica una concordancia moderada entre las predicciones del modelo y las observaciones reales.

La sensibilidad, también conocida como tasa de verdaderos positivos, revela la proporción de pacientes positivos que el modelo logra identificar correctamente. En este contexto, el modelo

identifica con precisión alrededor del 39.0 % de los pacientes con malaria. Por otro lado, la especificidad, o tasa de verdaderos negativos, muestra que el modelo identifica correctamente cerca del 94.0 % de los pacientes que no tienen la enfermedad.

El valor predictivo positivo, que representa alrededor del 71.9 % de los pacientes clasificados con malaria, refleja la proporción de pacientes que son verdaderamente positivos. En cuanto al valor predictivo negativo, este alcanza aproximadamente el 79.5 %, indicando la proporción de pacientes clasificados sin malaria que son verdaderamente negativos.

La prevalencia, que corresponde a la proporción de pacientes con malaria en la muestra total, es del 15.4 %. En términos de la tasa de detección en este escenario, el modelo identifica correctamente alrededor del 11.1 % de los pacientes con la enfermedad de la malaria. La prevalencia de detección muestra que el modelo clasifica aproximadamente el 28.4 % de los pacientes con la enfermedad.

Finalmente, la exactitud equilibrada en este caso es del 66.5 %, proporcionando una medida general del rendimiento del modelo en ambas clases.

Accuracy	0.784	Pos Pred Value	0.719
95 % CI	(0.741, 0.822)	Neg Pred Value	0.795
No Information Rate	0.716	Prevalence	0.154
Kappa	0.382	Detection Rate	0.111
Sensitivity	0.390	Detection Prevalence	0.284
Specificity	0.940	Balanced Accuracy	0.665

Tabla 3-13.: Resumen de los estadísticos para la matriz de confusión del KNN. Fuente: Elaboración propia.

En la tabla 3-14 se presenta un resumen de la capacidad de predicción de los cuatro modelos evaluados para los casos de malaria en la región de Ashanti de Ghana.

	Área bajo la curva ROC (AUC)
Regresión logística (LR)	81.5 %
Análisis discriminante lineal (LDA)	81.3 %
Naive Bayes (NB)	80.6 %
K vecinos más cercanos (KNN)	70.3 %

Tabla 3-14.: Áreas bajo la curva ROC para el LR, LDA, NB y KNN. Fuente: Elaboración propia.

En la tabla 3-15 se presenta un resumen de los errores cuadráticos medios para los cuatro modelos propuestos.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(e_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

	Error cuadrático medio (MSE)
Regresión logística (LR)	0.209
Análisis discriminante lineal (LDA)	0.197
Naive Bayes (NB)	0.221
K vecinos más cercanos (KNN)	0.216

Tabla 3-15.: Error cuadrático medio para el LR, LDA, NB y KNN. Fuente: Elaboración propia.

En la tabla 3-16 se muestra la exactitud con la que se adaptan matemáticamente a los datos los modelos: Regresión logística, naive Bayes, análisis discriminante lineal de Fisher y K vecinos más cercanos. En la tabla 3-17, se muestra la sensibilidad y especificidad de los modelos.

	Exactitud del modelo	95 % CI
Regresión logística (LR)	0.791	(0.749, 0.829)
Análisis discriminante lineal (LDA)	0.803	(0.761, 0.840)
Naive Bayes (NB)	0.779	(0.736, 0.818)
K vecinos más cercanos (KNN)	0.774	(0.741, 0.822)

Tabla 3-16.: Exactitud para los modelos LR, LDA, NB y KNN. Fuente: Elaboración propia.

	Especificidad	Sensibilidad
Regresión logística (LR)	0.7460	0.7989
Análisis discriminante lineal (LDA)	0.7727	0.8086
Naive Bayes (NB)	0.7030	0.7926
K vecinos más cercanos (KNN)	0.7188	0.7955

Tabla 3-17.: Especificidad y Sensibilidad de los modelos LR, LDA, NB y KNN. Fuente: Elaboración propia.

3.9. Comparación de los resultados

Después de comparar los resultados de los modelos de regresión logística, análisis discriminante lineal de Fisher, naive Bayes y K vecinos más cercanos en la tarea de predecir casos de malaria, se pueden resumir algunas observaciones clave:

- Capacidad Predictiva:

1. El modelo de regresión logística ha demostrado la mejor capacidad predictiva, con un área bajo la curva (AUC) del 81.5 %. Esto sugiere que es más efectivo en discriminar entre casos positivos y negativos.
2. El análisis discriminante lineal de Fisher también ha mostrado una buena capacidad predictiva con un AUC del 81.3 %.
3. El modelo naive Bayes ha alcanzado un AUC del 80.6 %, lo que indica una capacidad aceptable pero ligeramente inferior en comparación con los dos primeros modelos.
4. El K vecinos más cercanos ha mostrado la capacidad predictiva más baja con un AUC del 70.3 %.

- Interpretabilidad:
  1. La regresión logística y el análisis discriminante lineal de Fisher suelen ser más interpretables debido a su estructura y relación directa con probabilidades.
  2. El naive Bayes, aunque simple, puede ser menos intuitivo en términos de interpretación de las relaciones entre variables, pese a ser más flexible por no ser lineal.
  3. El K vecinos más cercanos tiende a ser menos interpretable, ya que depende de la similitud en el espacio de características. Pero podría llegar a tener mucho poder predictivo con pocas variables.

- Sensibilidad y Especificidad:
  1. La regresión logística y el análisis discriminante lineal de Fisher han mostrado sensibilidades y especificidades buenas.
  2. El naive Bayes ha tenido una sensibilidad aceptable pero con una especificidad ligeramente superior.
  3. El K vecinos más cercanos ha mostrado sensibilidad bastante baja y especificidad ligeramente alta.

- Errores Cuadráticos Medios (MSE):
  1. La regresión logística y análisis discriminante lineal de Fisher, tienen errores cuadráticos medios bajos (0.209 y 0.197, respectivamente), indicando buen poder predictivo en el test data.
  2. El naive Bayes, presenta un MSE aceptable de 0.221.
  3. El K vecinos más cercanos (KNN) con un error cuadrático medio de 0.216 es similar al naive Bayes (NB) con un error cuadrático medio de 0.221, indicando un ajuste adecuado, pero ligeramente inferior a los dos primeros modelos.

- Balance entre Sensibilidad y Especificidad:
  1. La regresión logística y análisis discriminante lineal de Fisher, logran un equilibrio eficiente entre sensibilidad y especificidad.

2. El naive Bayes mantiene un equilibrio, aunque con valores ligeramente inferiores.
3. El K vecinos más cercanos (KNN) muestra una sensibilidad más baja en comparación con la especificidad.

- Se podría llevar a cabo un estudio de simulación para obtener mayor poder en las conclusiones.

En resumen, la regresión logística y el análisis discriminante lineal de Fisher son modelos destacados en esta tarea, cada uno con sus ventajas y limitaciones. La elección del mejor modelo dependerá de las prioridades específicas, la interpretabilidad requerida y la naturaleza del problema médico en cuestión.

En este capítulo, se presentaron los resultados de cuatro modelos con fines predictivos aplicados al diagnóstico de la malaria. Los modelos incluyeron regresión logística (LR), análisis discriminante lineal de Fisher (LDA), Naive Bayes (NB) y K vecinos más cercanos (KNN). Se utilizó un conjunto de datos dividido en un 80 % de entrenamiento y un 20 % de prueba, y se evaluaron las capacidades predictivas de cada modelo utilizando métricas como el área bajo la curva ROC, la matriz de confusión, entre otros.

Para el modelo de regresión logística, se destacaron variables significativas como la hemoglobina (Hb), las plaquetas (Plt), los linfocitos (Lymph) y la edad (Age). Se proporcionaron coeficientes estimados y se interpretaron sus efectos en la probabilidad condicional de tener malaria. Además, el capítulo abordó la evaluación de la capacidad discriminativa de modelos predictivos, utilizando las covariables mencionadas. Se analizó la capacidad de las variables para predecir la presencia del parásito de la malaria en los pacientes, evaluando áreas bajo la curva ROC, puntos de corte, sensibilidad y especificidad.

El análisis discriminante lineal de Fisher se aplicó utilizando las mismas variables mencionadas, y se presentaron los coeficientes de discriminación lineal. Este modelo también se evaluó mediante la curva ROC y se destacó su capacidad de discriminación.

Para el clasificador Naive Bayes, se utilizó el vector de variables hematológicas y edad para predecir la presencia de malaria. Se calcularon las probabilidades condicionales y se evaluó su rendimiento con la curva ROC.

El modelo de K vecinos más cercanos se implementó calculando las distancias euclidianas entre puntos en el espacio de características. Se presentó la matriz de confusión y se evaluó la capacidad predictiva del modelo.

Finalmente, se realizó una exhaustiva comparación de los cuatro modelos con el objetivo de evaluar el rendimiento de cada uno de ellos.

El cierre de este capítulo destaca la importancia de evaluar no solo la precisión global de los modelos, sino también aspectos como la sensibilidad, especificidad y otros estadísticos

detallados en la evaluación del rendimiento. Además, se enfatiza la necesidad de considerar el contexto clínico y las implicaciones prácticas al seleccionar un modelo para aplicaciones médicas.

En el capítulo cuatro se presentarán las conclusiones de los modelos mencionados en este informe, con el objetivo de consolidar y dar sentido a los resultados obtenidos, proporcionando así una visión general y clara.

4. Conclusiones y recomendaciones

4.1. Conclusiones

Evaluando el desempeño de los cuatro modelos, teniendo en cuenta el conjunto de datos de malaria para escoger el mejor de ellos, se obtienen los siguientes resultados:

- El modelo de regresión logística ha demostrado tener una capacidad considerable para predecir la presencia del parásito de la malaria en pacientes, con un área bajo la curva (AUC) del 81.5 %. Además, su interpretabilidad es alta, lo que facilita la comprensión de las relaciones entre las variables predictoras y la variable de respuesta. La sensibilidad y especificidad del modelo son razonables. El error cuadrático medio es bajo, indicando un buen ajuste del modelo, y mostrando un rendimiento superior en comparación con los otros modelos.

- El modelo de análisis discriminante lineal de Fisher también ha mostrado una buena capacidad de predicción, con un AUC del 81.3 %. Aunque la sensibilidad es alta, la especificidad es moderada. El error cuadrático medio es bajo, indicando un buen ajuste. En términos de interpretabilidad, este modelo puede ser más complejo que la regresión logística, pero sigue siendo comprensible.

- Aunque el modelo naive Bayes ha mostrado una capacidad aceptable con un AUC del 80.6 %, su rendimiento es ligeramente inferior en comparación con los anteriores. El error cuadrático medio es aceptable. Sin embargo, naive Bayes es conocido por su simplicidad y eficiencia en ciertos contextos. La interpretabilidad es buena, pero puede ser menos intuitiva en comparación con la regresión logística..

- El modelo K vecinos más cercanos ha demostrado una capacidad predictiva más baja con un AUC del 70.3 %. La sensibilidad y especificidad son moderadas. El error cuadrático medio es adecuado y similar al naive Bayes, pero ligeramente inferior a los dos primeros modelos. En cuanto a la interpretabilidad, KNN tiende a ser menos interpretable, ya que depende de patrones en los datos que pueden no ser fácilmente explicables.

- La sensibilidad, especificidad, el error cuadrático medio y el área bajo la curva ROC son métricas cruciales a considerar en el contexto médico. La elección del modelo debe basarse en cómo estas métricas se alinean con los objetivos clínicos y los costos asociados con los errores de clasificación.

- Dada la naturaleza médica de la aplicación, la colaboración con profesionales de la salud y expertos en bacteriología es fundamental para garantizar la interpretación correcta y la aplicabilidad clínica de los modelos.
- La regresión logística y el análisis discriminante lineal de Fisher suelen ser más interpretables que otros modelos más complejos. La factibilidad de implementación en un entorno clínico también es esencial.

4.2. Recomendaciones

- Con el análisis realizado al conjunto de datos de los pacientes que tuvieron malaria en la región de Ashanti en Ghana, se puede dar referencia partiendo de los resultados obtenidos para la toma de decisiones en el área de la medicina, especialmente en bacteriología, donde se quiere obtener resultados óptimos de predicción del parásito de la malaria.
- Se recomienda el uso de modelos de regresión logística en el proceso de predicción o diagnóstico del parásito de la malaria en pacientes que presenten la sintomatología de la enfermedad, dado que, de acuerdo a los resultados obtenidos, se tiene en cuenta que la regresión logística es un modelo bien establecido y comprendido que se puede entender mejor en el área médica. Su estructura logística facilita la interpretación de las relaciones entre variables predictoras y la variable de resultados.

A. Apéndice

A.1. Glosario

Malaria: La malaria, también conocida como paludismo, es una enfermedad transmitida por un grupo de parásitos en la sangre llamados protozoos del género Plasmodium. Estos parásitos son transmitidos a través de la picadura de mosquitos hembra del género Anopheles, que se crían en aguas superficiales de poca profundidad (Nájera et al., 2009).

Curva ROC: La curva ROC es una herramienta muy valiosa para evaluar y comparar el rendimiento de modelos de clasificación binaria en términos de sensibilidad y especificidad a través de varios umbrales de clasificación (James et al., 2021).

Sensibilidad: a sensibilidad de un test es la probabilidad de que dicho test arroje un resultado positivo, considerando la premisa de que el paciente sea efectivamente positivo (Milton, 2007).

Especificidad: La especificidad de un test es la probabilidad de que dicho test arroje un resultado negativo, teniendo en cuenta la premisa de que el paciente sea ciertamente negativo (Milton, 2007).

AUC: En inglés, AUC, que significa "Área Bajo la Curva", se define como la superficie bajo la curva. Cuanto mayor sea el AUC, mayor será la eficacia del clasificador (James et al., 2021).

Valor predictivo positivo: El valor predictivo positivo de un test es la probabilidad de que una persona sea verdaderamente positiva, dado que el resultado del test es positivo (Milton, 2007).

Valor predictivo negativo: El valor predictivo negativo de un test es la probabilidad de que una persona sea verdaderamente negativa, dado que el resultado es negativo (Milton, 2007).

A.2. Ecuaciones para los estadísticos utilizados en el tercer capítulo

Teniendo en cuenta la tabla A-1, se presentará un resumen de los estadísticos utilizados en el tercer capítulo.

	Predichos		
Observados	Sin paludismo (0)	Con paludismo (1)	Total
Sin paludismo (0)	d	c	$c+d$
Con paludismo (1)	b	a	$a+b$
Total	$b+d$	$a+c$	n

Tabla A-1.: Matriz de confusión usada en los cuatro modelos. Fuente: Elaboración propia.
Estadísticos:

$$Accuracy = \frac{a+d}{n} \quad No\ Information\ Rate = \frac{b+d}{n} \quad Sensitivity = \frac{a}{a+c}$$

$$Specificity = \frac{d}{b+d} \quad Pos\ Pred\ Value = \frac{a}{a+b} \quad Neg\ Pred\ Value = \frac{d}{c+d}$$

$$Prevalence = \frac{a+c}{n} \quad Detection\ Rate = \frac{a}{n} \quad Detection\ Prevalence = \frac{a+b}{n}$$

$$Balanced\ Accuracy = \frac{Sensitivity + Specificity}{2} = \frac{a(b+d)+d(a+c)}{2(a+c)(b+d)}$$

A.3. Códigos en R

A.3.1. Código para el modelo de regresión logística

```
# Librerías

library(dplyr)
library(plyr)
library(forcats)
library(readxl)
library(readr)
library(car)
library(pROC)
library(gmodels)
library(caret)
require(rpart)
library(reshape)

DATOS_SOBRE_MALARIA<-read_csv("DATOS SOBRE MALARIA.csv")

# Construyendo el modelo de regresión logística (glm) apartir de los datos,
# tenemos:
```

```
df<-DATOS_SOBRE_MALARIA
dim(df)

set.seed(123)
smp_size <- floor(0.8 * nrow(df))
smp_size

train_ind <- sample(seq_len(nrow(df)), size = smp_size)
train_ind

train<-df[train_ind,1:12 ]
train
# dim(train)

test<-df[-train_ind,1:12 ]
test

modelo_glm = glm(Malaria ~ Hb + Plt + Lymph + Age, data =train,family = binomial)
modelo_glm

# Prediciendo las probabilidades para el test, tenemos:

probabi_test <- modelo_glm %>% predict(test, type = "response")
probabi_test

contrasts(factor(test$Malaria))

# El siguiente código clasifica las personas en función de sus
# probabilidades previstas, si es mayor que 0.5  se le asigna el 1,
# y si es menor o igual se le asigna el 0.

predicted.classes <- ifelse(probabi_test > 0.5, "1", "0")
predicted.classes

Matriz_confusión = table(test$Malaria, predicted.classes)
Matriz_confusión

Tasa_error<-(Matriz_confusión[1,2]+
               Matriz_confusión[2,1])/sum(Matriz_confusión)
Tasa_error

# Gráficando la curva ROC o curva Característica Operativa del Receptor,
# tenemos:
```

```
par(pty = "s")  # Hacer cuadrado el espacio ROC
roc_graph<- roc(test$Malaria, probabi_test, plot = TRUE, legacy.axes = TRUE,
                 percent = TRUE,
                 ylab = "Sensitividad \n Porcentaje de verdaderos positivos",
                 xlab = "1-especificidad \n Porcentaje de falsos positivos",
                 col = "darkblue", lwd = 2,
                 print.auc = TRUE, auc.polygon = TRUE,
                 auc.polygon.col = "aliceblue", grid=TRUE,
                 print.thres=FALSE)
coordenada <- pROC::coords(roc_graph,"best",
                           ret=c("threshold","specificity","sensitivity"))
points(x=coordenada$specificity,
       y=coordenada$sensitivity, pch=0,col="red")
roc_graph

# Otra forma de mostrar la especificidad, sensitividad, tasa de falsos
# positivos, entre otros.

confusionMatrix(data = as.factor(test$Malaria),
                reference=as.factor(predicted.classes),  positive="1")
```

A.3.2. Código para el discriminante lineal de Fisher

```
# Librerías

library(MASS)
library(klaR)
library(tidyverse)
library(pROC)
library(gmodels)
library(caret)

DATOS_SOBRE_MALARIA<-read_csv("DATOS SOBRE MALARIA.csv")

# Seleccionando el 80% de los datos del conjunto, se tiene:

df<-DATOS_SOBRE_MALARIA
dim(df)

set.seed(123)
smp_size <- floor(0.8 * nrow(df))
smp_size
```

```
train_ind <- sample(seq_len(nrow(df)), size = smp_size)
train_ind

train<-df[train_ind,1:12 ]
train
# dim(train)

test<-df[-train_ind,1:12 ]
test
# dim(test)

y_test=test$Malaria
y_test
# length(y_test)
# table(y_test)

lda.fit <- lda(Malaria ~ Hb + Plt + Lymph + Age, data = train)
lda.fit

y_TRAIN<-lda.fit
y_TRAIN

length(y_TRAIN)
length

# Predicciones de las probabilidades en funsión de las clases 0 y 1.

y_LDA_clases<-predict(lda.fit,newdata=test)$class
y_LDA_clases

# Matriz de confusión

confussion_matrix<-table(y_test,y_LDA_clases)
confussion_matrix


Tasa_error<-(confussion_matrix[1,2]+
               confussion_matrix[2,1])/sum(confussion_matrix)
Tasa_error

# Hallando las probabilidades del modelo en el intervalo (0, 1), tenemos:

y_LDA_prob<-predict(lda.fit,newdata=test)$posterior
y_LDA_prob
```

```
# Creando un data set para las probabilidades, tenemos:

tabla_prob<-data.frame(y_LDA_prob)
tabla_prob

# Gráficando la curva ROC o curva Característica Operativa del Receptor,
# tenemos:

par(pty = "s")  # Hacer cuadrado el espacio ROC
roc_graph <- roc(test$Malaria, tabla_prob$X1, plot = TRUE, legacy.axes = TRUE,
                 percent = TRUE,
                 ylab = "Sensitividad \n Porcentaje de verdaderos positivos",
                 xlab = "1- especificidad \n Porcentaje de falsos positivos",
                 col = "darkblue", lwd = 2,
                 print.auc = TRUE, auc.polygon = TRUE,
                 auc.polygon.col = "aliceblue", grid=TRUE,
                 print.thres=FALSE)
coordenada <- pROC::coords(roc_graph,"best",
                           ret=c("threshold","specificity","sensitivity"))
points(x=coordenada$specificity,
       y=coordenada$sensitivity, pch=0,col="red")
roc_graph

# Especificidad, sensitividad, tasa de falsos positivos, entre otros.

confusionMatrix(data = as.factor(y_test),
                reference=as.factor(y_LDA_clases),  positive="1")
```

A.3.3. Código para el naive Bayes

```
# Librerías

library(e1071)
library(MASS)
library(klaR)
library(ROCR)
library(tidyverse)
library(vcd)
library(contrast)
library(pROC)
library(gmodels)
library(caret)
```

```
DATOS_SOBRE_MALARIA<- read_csv("DATOS SOBRE MALARIA.csv")

df<-DATOS_SOBRE_MALARIA
dim(df)

set.seed(123)

# Seleccionando el 80% de los datos del conjunto, se tiene:

smp_size <- floor(0.8 * nrow(df))
smp_size

train_ind <- sample(seq_len(nrow(df)), size = smp_size)
train_ind

train<-df[train_ind,1:12]
train

# dim(train)

test<-df[-train_ind,1:12]
test

# dim(test)

y_test=test$Malaria
y_test

# length(y_test)
# table(y_test)

# Modelo naive Bayes

nb.fit <- naiveBayes(Malaria ~ Hb + Plt + Lymph + Age, data = train)
nb.fit

# Probabilidades en función de las clases 0 y 1.

y_nb_clases <- predict(nb.fit , newdata=test)
y_nb_clases

length(y_nb_clases)
```

```
# Matriz de confusión

Matriz_confusión = table(y_test, y_nb_clases)
Matriz_confusión

# Tasa de error

Tasa_error<-(Matriz_confusión[1,2]+
               Matriz_confusión[2,1])/sum(Matriz_confusión)
Tasa_error

# Hallando las probabilidades del modelo en el intervalo (0, 1), tenemos:

y_nb_prob <- predict(nb.fit , newdata=test, type = "raw")
y_nb_prob

# Creando un data set para las probabilidades, tenemos:

tabla.prob<-data.frame(y_nb_prob)
tabla.prob

# Gráficando la curva ROC o curva Característica Operativa del Receptor,
# tenemos:

par(pty = "s")  # Hacer cuadrado el espacio ROC
roc_graph <- roc(test$Malaria, tabla.prob$X1, plot = TRUE, legacy.axes = TRUE,
                 percent = TRUE,
                 ylab = "Sensitividad \n Porcentaje de verdaderos positivos",
                 xlab = "1- especificidad \n Porcentaje de falsos positivos",
                 col = "darkblue", lwd = 2,
                 print.auc = TRUE, auc.polygon = TRUE,
                 auc.polygon.col = "aliceblue", grid=TRUE,
                 print.thres=FALSE)
coordenada <- pROC::coords(roc_graph,"best",
                           ret=c("threshold","specificity","sensitivity"))
points(x=coordenada$specificity,
       y=coordenada$sensitivity, pch=0,col="red")
roc_graph

# Especificidad, sensitividad, tasa de falsos positivos, entre otros.

confusionMatrix(data = as.factor(y_test),
                reference=as.factor(y_nb_clases),  positive="1")
```

A.3.4. Código para K vecinos más cercanos

```
# Librerías

library(MASS)
library(class)
library(readxl)
library(readr)
library(tidyverse)
library(pROC)
library(gmodels)
library(caret)

Datosdemalaria <- read_csv("DATOS SOBRE MALARIA.csv")

df=data.frame(Datosdemalaria)
df

normalize <- function(x) {
  norm <- ((x - min(x))/(max(x) - min(x)))
  return (norm)
}

set.seed(123)

smp_size <- floor(0.8 * nrow(df))

train_ind <- sample(seq_len(nrow(df)), size = smp_size)

train <- normalize(df[train_ind, c(2, 6:11)])
train

test <- normalize(df[-train_ind, c(2, 6:11)])
test

y_train=df[train_ind,5]
y_train

y_test=df[-train_ind,5]
y_test

fit.knn_Test<-knn(train=train, test=test,cl=y_train, k=50, prob=TRUE)
fit.knn_Test
```

```
tabla.probab<-attributes(.Last.value)$prob
tabla.probab

Predicted_test<-factor(fit.knn_Test)
Predicted_test

Matriz_confusión_Knn<-table(Predicted_test,y_test)
Matriz_confusión_Knn

Test_error<-(Matriz_confusión_Knn[1,2]+
               Matriz_confusión_Knn[2,1])/(sum(Matriz_confusión_Knn))
Test_error

# Gráficando la curva ROC o curva Característica Operativa del Receptor,
# tenemos:

par(pty = "s")  # Hacer cuadrado el espacio ROC
roc_graph <- roc(y_test, tabla.probab, plot = TRUE, legacy.axes = TRUE,
                 percent = TRUE,
                 ylab = "Sensitividad \n Porcentaje de verdaderos positivos",
                 xlab = "1- especificidad \n Porcentaje de falsos positivos",
                 col = "darkblue", lwd = 2,
                 print.auc = TRUE,
                 auc.polygon = TRUE, auc.polygon.col = "aliceblue", grid=TRUE,
                 print.thres=FALSE)
coordenada <- pROC::coords(roc_graph,"best",
                           ret=c("threshold","specificity","sensitivity"))
points(x=coordenada$specificity,
       y=coordenada$sensitivity, pch=0,col="red")
roc_graph

# Especificidad, sensitividad, tasa de falsos
# positivos, entre otros.

confusionMatrix(data = as.factor(y_test),
                reference=as.factor(Predicted_test),  positive="1")
```

Bibliografía

Adera, T. D. (2003). Beliefs and traditional treatment of malaria in Kishe settlement area, Southwest Etiopía. 41(1):25–34.

Adusei, K. A. and Owusu-Ofori, A. (2018). Prevalence of plasmodium parasitaemia in blood donors and a survey of the knowledge, attitude and practices of transfusion malaria among health workers in a hospital in Kumasi, Ghana. *Plos one*, 13(11):1–13.

Aliyu, A. and Bada, A. B. (2023). Improved Malaria Outbreak Predictive Model Using Naive Baye and Artificial Neural Network. *United International Journal for Research & Technology*, 4(9):1–14.

Aliyu, A., Yau, S., and Aliyu, F. M. (2021). Predicting Malaria Using Logistic Regression Model. *Ilorin Journal of Computer Science and Information Technology*, 4(2):1–6.

Anabire, N. G., Aculley, B., Pobee, A., Kyei-Baafour, E., Awandare, G. A., del Pilar Quintana, M., Hviid, L., and Ofori, M. F. (2023). High burden of asymptomatic malaria and anaemia despite high adherence to malaria control measures: a cross-sectional study among pregnant women across two seasons in a malaria-endemic setting in Ghana. *National Library of Medicine*, 51(1):1717–1729.

Ankamah, S., Nokoe, K. S., and Iddrisu, W. A. (2018). Modelling Trends of climatic Variability and Malaria in Ghana Using Vector Autoregression. *Malaria Research and Treatment*, 2018(1):1–12.

Bowerman, B. L., Connell, R. T., and Koehler, A. B. (2009). *Pronósticos, series de tiempo y regresión.* Cengage Learning.

Devi, S. S., Samantha, E., Priyadharshini, B., and Jetlin, C. P. (2021). Malaria Detection Using Machine Learning with k Nearest Neighbor Algorithm. *International Journal of Scientific Development and Research*, 6(3):1–4.

Díaz, L. G., Morales, M. A., and León, L. R. (2018). *Análisis estadístico de datos categóricos.* Universidad Nacional de Colombia, Unibiblos.

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

Irmanita, R., Prasetiyowati, S. S., and Sibaroni, Y. (2021). Classification of malaria complication using cart (classification and regression tree) and naive Bayes. *Jurnal RESTI*, 5(1):10–16.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2021). *An Introduction to Statistical Learning with Applications in R.* Springer.

Johnson, R. A. and Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis.* Pearson.

Landier, J., Parker, D. M., Thu, A. M., Carrara, V. I., Lwin, K. M., Bonnington, C. A., Pukrittayakamee, S., GillesDelmay, and Nosten, F. H. (2016). The role of early detection and treatment in malaria elimination. *Malaria Journal*, 15(1):1–8.

Leslie, T., Mikhail, A., Mayan, I., Anwar, M., Sayed Bakhtash, M. N., Chandler, C., Whitty, C. J. M., and Rowland, M. (2012). Overdiagnosis and mistreatment of malaria among febrile patiens at primary healthcare level in Aghanistan: Observational Study. *National Library of Medicine*, 345(2012):1–13.

Mahmoudi, N., de Julián-Ortiz, J.-V., Ciceron, L., Gálvez, J., Mazier, D., Danis, M., Derouin, F., and García-Domenech, R. (2006). Identification of new antimalarial drugs by linear discriminant analysis and topological virtual screening. *Journal of Antimicrobial Chemotherapy*, 57(3):489—-497.

Milton, J. S. (2007). *Estadística para Biología y Ciencias de la Salud.* McGRAW Hill.

Monroy, L. G. D. and Rivera, M. A. M. (2012). *Análisis estadístico de datos multivariados.* Coordinación de publicaciones, Facultad de Ciencias Diagramación en Latex.

Montgomery, D. C., Peck, E. A., and Vining, G. G. (2012). *Introduction to Linear Regression Analysis.* Wiley.

Nkansah, C., Bani, S. B., Mensah, K., Appiah, S. K., Boakye, F. O., Abbam, G., Daud, S., Agyare, E. M., Agbadza, P. E., Derigubah, C. A., Serwaa, D., Apodola, F. A., Quansah, Y., Issah, R., Dindiok, S. Y., and Chukwurah, F. E. (2023). Serum anti-eryhropoietin antibodies among pregnant women with plasmodium falciparum malaria and anaemia: A case control study in northeren Ghana. *Plos one*, 18(3):1–20.

Nájera, J. A., Bueno, A. G., and Díaz, A. B. (2009). *Malaria.* Biblioteca nacional de España.

Opoku-Ansah, J., Eghan, M. J., Anderson, B., Boampong, J. N., Edziah, R., Adueming, P. O.-W., and Amuah, C. L. Y. (2019). Optical Identification of Plasmodium falciparum Malarial Byproduct for Parasite Density Estimation. *Hindawi*, 2019(8782567):1–14.

Overleaf (2023). Overleaf: LaTeX, Evolución. El editor de LaTeX fácil de usar, online y colaborativo, Londres, Reino Unido. URL `https://es.overleaf.com/`.

Paintsil, E. K., Omari-Sasu, A. Y., Addo, M. G., and Boateng, M. A. (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. *Malaria Research and Treatment*, 2019(1):1–7.

Poostchi, M., Silamut, K., Maude, R. J., Jaeger, S., and Thoma, G. (2023). Image analysis and machine learning for detecting malaria. *Translational Research*, 194:36–55.

Qadri, A. M., Raza, A., Eid, F., and Abualigah, L. (2023). A novel transfer learning-based model for diagnosing malaria from parasitized and uninfected red blood cell images. *Decision Analytics Journal*, 9(100352):1–11.

R Core Team (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

RStudio Team (2023). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA. URL `http://www.rstudio.com/`.

Sach, J. and Malaney, P. (2002). The economic and social burden of malaria. *National Library of Medicine*, 415(6872):680–5.

Sajana, T. and Narasingarao, M. R. (2018). Classification of Imbalanced Malaria Disease Using Naive Bayesian Algorithm. *International Journal of Engineering & Technology*, 7(2.7):786–790.

Seyoum, T. F., Andualem, Z., and Yalew, H. F. (2023). Insecticide-treated bed net use and associated factors among households having under-five children in east África: a multilevel binary logistic regression analysis. *Malaria Journal*, 22(10):1–10.

Stephen, A., Akomolafe, P. O., and Ogundoyin, K. I. (2021). A Model for Predicting Malaria Outbreak Using Machine Learning Technique. *Scientific Annals of Computer Science*, 19(1):9–15.

Wackerly, D. D., Mendenhall, W., and Scheaffer, R. L. (2010). *Mathematical statistics with applications.* Cengage Learnig.

Walpole, R. E., Myers, R. H., Myers, S. L., and Ye, K. (2012). *Probability and statistics for engineering and science.* Pearson.

Zuluaga, G. C. and Trujillo, S. B. (2010). *Malaria: consideración sobre su diagnóstico.* Programa de Educación Médica Continúa Certificada Universidad de Antioquia, Edimedico.