PREDICTIVE ANALYSIS OF PEDIATRIC IN HOSPITAL MALARIA MORTALITY USING MACHINE LEARNING FOR EARLY CLINICAL INTERVENTION

DENNIS ARTHUR NYANZI

M24M19/008

A DISSERTATION SUBMITTED TO THE FACULTY OF ENGINEERING, DESIGN AND TECHNOLOGY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF A DEGREE OF MASTER OF SCIENCE IN DATA SCIENCE AND ANALYTICS OF UGANDA CHRISTIAN UNIVERSITY

April, 2026

UGANDA CHRISTIAN UNIVERSITY

A Centre of Excellence in the Heart of Africa

Declaration

I, *Nyanzi Dennis Arthur*, declare to the best of my knowledge that the work in this dissertation is original and has not been submitted before to any university/institution. It is the result of my effort.

Signature: _____

Date: 28-04-2026

Ethical Generative AI Usage Declaration

A. **Declaring the non-use of AI for content generation**

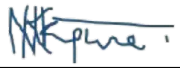
I, *Nyanzi Dennis Arthur*, declare that no content generated by AI technologies has been used in this dissertation.

Signature: _____ Date: 28-04-2026

Approval

This dissertation has been submitted for examination with approval of the following advisors/supervisor.

Advisor: *Dr. Kimbugwe Nasser*
Department of Computing & Technology
Faculty of Engineering Design & Technology
Uganda Christian University

Signature: 

Date: 28-04-2026

Abstract

Malaria remains a major public health challenge in Uganda especially among the children. Pediatric patients are more vulnerable because of weaker immune systems and faster disease progression. Delayed assessment increases mortality rates. This study aimed at using machine learning to predict mortality among children admitted with malaria. An experimental computational research design was adopted on retrospective secondary data. Five supervised machine learning models were developed: Decision Tree Classifier, Random Forest Classifier, Naive Bayes, Logistic Regression and Gradient Boosting model. These models were developed to predict mortality among pediatric malaria patients. Models were evaluated using precision, recall, F1 score, Confusion Matrix, Learning curves and Receiver Operating Characteristic Area Under Curve. K fold cross validation was applied during training and Grid search was used for hyperparameter tuning. The results showed that the Naive Bayes classifier performed best followed by the Logistic Regression model having achieved the highest receiver operating characteristic area under curve of 0.899 and 0.885 respectively. The models were able to accurately predict the likelihood of mortality among children admitted with malaria through a combination of both clinical and laboratory factors. Machine learning demonstrated potential in early identification of high risk pediatric malaria patients for timely clinical intervention.

Dedication

I dedicate my research to my family for their support and encouragement during my research journey

Acknowledgments

I would like to express my gratitude to my supervisor for his guidance through out my research journey.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

Child mortality remains a very crucial global health issue, especially when it comes to limited resource settings where there are inadequate health care systems that are even inaccessible, malnutrition is high and other socioeconomic disparities add so much to the mortality rates among pediatrics. According to the World Health Organization (WHO), there are reports that infectious diseases like malaria, poor sanitation and limited maternal health care services have had a very high impact when it comes to the number of child deaths worldwide. The development of predictive analytics using machine learning has highly impacted in early identification of children with high mortality risks and some studies like the one for Chandra Sekhar 2025 explored the application of machine learning techniques including logistic regression, support vector machines, random forest model to predict child mortality risks with basis on health indicators and the target of the study was to improve on the data driven decision making approaches in health settings [1]. There are various challenges that health care systems interface on the global scale and these include the widely growing patient numbers in health facilities, the high costs for medication and the complexity involved in the parasite detection of some rare species of malaria therefore due to such demanding issues, there was need for fast innovation and invention of solutions to handle the patient care better and ensure that the operational efficiency is streamlined. Due to the challenges in the system, predictive models have emerged as a new approach to pave way for transformation through use of the complex methods of data analytics to forecast patient requirements, optimize resources and also combat medical errors. The original methods used in disease diagnostics are important but have a challenge of limitation in capturing the intricate patterns with the big datasets thus welcoming the new era of machine learning which has techniques of picking insights from the datasets therefore the traditional diagnostic methods are not being replaced but rather supplemented for high performance rates. Due to the early detection of machine learning, this has enabled achievement of success factors in early disease detection, treatment planning and hospital re-admission prevention through deep learning analysis of broad and vast datasets thus providing real time reports of analysis for proper diagnostic accuracy. Some of the examples on how these machine learning models have helped in diagnostics is one in which neural networks have proved precision in cancer detection in medical imaging where insights are provided by models that radiologists could have missed out on, also used in prediction of patient deterioration, allowing timely interventions [2]. The major goal of this study was to assess the various machine learning classifiers in prediction of patient outcomes of pediatrics with malaria by use of the vari-

ous data collected from the patients using the traditional techniques from the laboratory and the clinic. The motivation for this research was inspired by experiences shared by pediatricians working in government hospitals where high patient volumes often make it difficult to identify critically ill children. In some cases the very sick children are attended to late thus leading to delayed intervention.

1.2 Background

Malaria remains one of the leading causes of sickness and deaths globally, particularly among children under five years of age in Africa. According to the World Health Organization (WHO), approximately 249 million malaria cases and 608,000 deaths were reported globally in 2022 across more than 90 countries. Uganda continues to carry a high malaria burden, with approximately 12.7 million reported cases and 17,556 deaths annually. Children are more vulnerable due to their underdeveloped immune systems, leading to increased risk of severe malaria complications and mortality. Several malaria diagnostic techniques are currently being implemented in Ugandan hospitals, including Polymerase Chain Reaction (PCR), Rapid Diagnostic Tests (RDTs), and microscopic examination of blood smears[3]. Although these approaches are widely used, they are often associated with high operational expenses and delayed turnaround times. In many Ugandan health facilities, especially in resource limited settings, these challenges contribute to delayed diagnosis and poor clinical decision making for high risk patients. Predictive modelling in healthcare has evolved over several decades, initially relying on statistical techniques and rule-based systems such as logistic regression and clinical decision support systems like MYCIN. These traditional approaches played an important role in disease diagnosis and treatment planning, particularly in cardiovascular and infectious diseases. However, they often struggled to handle complex, high-dimensional clinical datasets and lacked the adaptability required for evolving healthcare environments. Advancements in computational power, electronic health records, and big data technologies have accelerated the adoption of machine learning techniques in healthcare. Unlike traditional statistical models, machine learning algorithms can automatically learn patterns, relationships, and hidden insights from large datasets without requiring extensive manual programming [4]. This capability has enabled improved analysis of structured and unstructured clinical data, including laboratory findings, medical images, and clinical notes. Machine learning applications have increasingly been used in disease prediction, mortality risk assessment, and early detection of severe illness among pediatric patients. Previous studies have demonstrated the usefulness of machine learning models in predicting child mortality and identifying important risk factors associated with poor clinical outcomes. These studies commonly applied preprocessing techniques such as feature scaling, encoding, and data balancing to improve predictive performance. Model evaluation was frequently conducted using metrics such as accuracy, confusion matrices, receiver operating characteristic (ROC) curves, precision, recall, F1-score, and Matthews Correlation Coefficient. Despite these advancements, several gaps remain unresolved. Many previous studies focused primarily on image-based diagnosis or general disease classification rather than predicting mortality outcomes among pediatric malaria patients using routine clinical and laboratory parameters [5]. Additionally, most existing models were developed using datasets from non-Ugandan populations, limiting their applicability to local healthcare settings. In Uganda, many hospital systems still lack intelligent predictive tools integrated into electronic medical records that can support early identification of high-risk malaria patients and assist clinicians in timely resource allocation and intervention. Furthermore, some previous machine learning studies relied heavily on single-model

approaches or achieved limited generalizability due to insufficient feature selection, class imbalance, small datasets, and inadequate external validation [6]. There is therefore a need for more robust supervised machine learning models capable of accurately predicting mortality among pediatric malaria patients using locally relevant clinical and laboratory data. This study aims to address these gaps by developing and evaluating supervised machine learning models, including Random Forest, Gradient Boosting, Logistic Regression, Decision Tree, and Naïve Bayes classifiers, for mortality prediction among pediatric malaria patients using clinical and laboratory parameters obtained from Ugandan healthcare settings. The study contributes toward improving early risk detection, supporting clinical decision-making, and enhancing resource allocation for critically ill children in malaria-endemic regions.

1.3 Problem Statement

Since malaria is one of the leading causes of morbidity in Uganda with 90-95% of population at risk, it's contributing to 13% of under-five mortality and 10% for under fifteen [7]. The traditional clinical diagnostics are so efficient but have a gap of early detection of patients who are highly vulnerable to high mortality levels thus creating need for implementation of a data driven approach to supplement their performance with emphasis put on analysis of diverse patient data for accurate prediction of disease severity in the low resource settings [8]. Unlike in the developed countries where patients are monitored efficiently with adequate resources in place, the low resource settings have high patient inflows with limited resources available therefore creating need for a data driven approach to assess which patients might need early interventions to be saved and more so there are various silent factors that can go unnoticed by the traditional methods but will highly be noticed by the machine learning model since outcome prediction is based on data patterns. Therefore, this study enabled the training of predictive algorithms for early intervention support to improve the outcomes of the patients during hospitalization and the outcome was a validated machine learning classification model with accuracy in prediction of hospitalization outcomes like recovery, mortality in pediatric malaria patients through use of the relevant clinical data and demographic parameters.

1.4 Research Objectives

1.4.1 General Objective

To develop and evaluate a machine learning model used in prediction of pediatric malaria patient mortality during hospitalisation.

1.4.2 Specific Objective

(a) To determine the correlation and predictive significance of clinical and laboratory parameters associated with mortality among hospitalized malaria children for development of a machine learning model.

(b) To develop a machine learning model that accurately predicts patient mortality during hospitalization.

(c) To evaluate the classification model predicting patient mortality during hospitalization.

1.5 Research Questions

(a) What are the factors that contributed significantly to mortality of malaria children?

(b) What is the best machine learning model in prediction of mortality among malaria children?

(c) What is the performance of the machine learning algorithms across various evaluation metrics in prediction of malaria patient mortality?

1.6 Conceptual framework

The dataset included various independent variables with the target variable as patient outcome. The independent variables include; Demographics; Age, sex. Clinical factors include; Chills, Fever, Fatigue, Sweats, Headache. Socio economic status; Region. Laboratory factors; White blood cells, Hemoglobin, Platelets, Parasitemia.

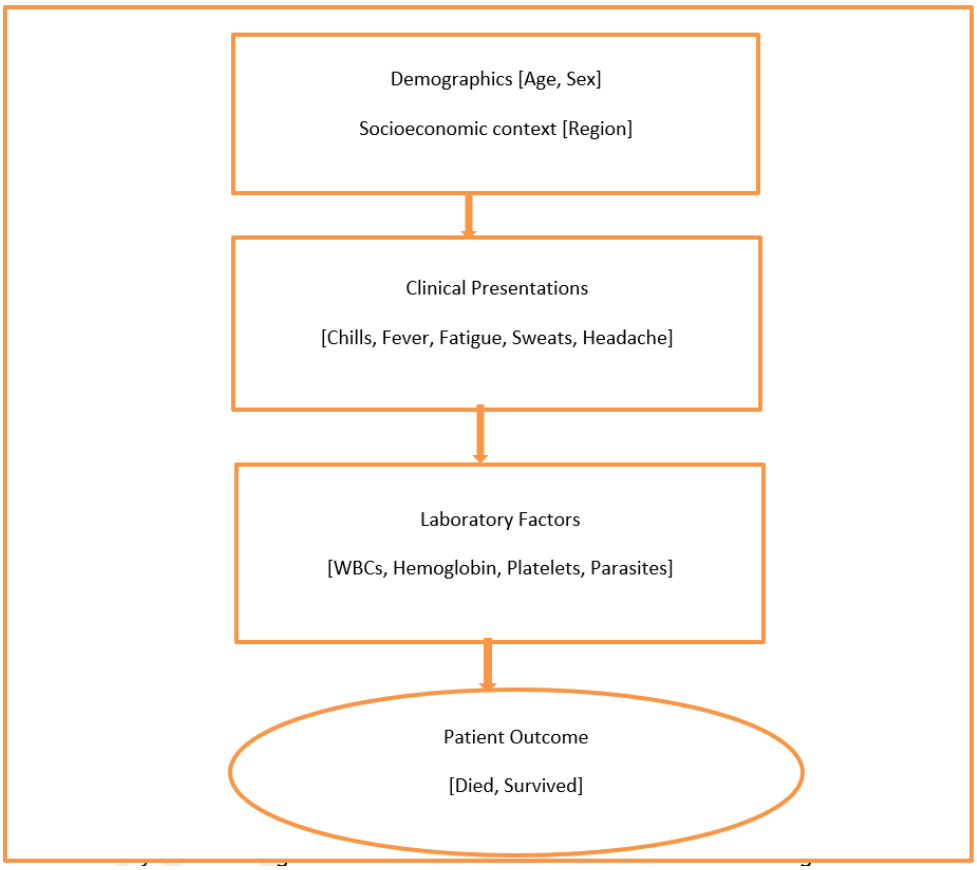


Figure 1.1: Conceptual Framework Diagram

1.7 Justification

From the clinical perspective this model was trained to fast-track patients with abnormal parameter values from the laboratory tests so as to take appropriate measures like transfusion and administration of fluids to enable them gain stability. Additionally, the children admitted with thrombocytopenia and anemia were transfused with platelets and whole blood respectively to ensure that they stabilize. [9] From the academic perspective, this project provided additional evidence that machine learning algorithms could be used in identification of children with high risk factors of mortality during hospitalization due to the abnormal laboratory factors seen after they are hospitalized in the low resource settings. The model could easily reflect those children with such issues so that the health care workers could figure out ways on how to handle the clinical decisions[9].

Chapter 2

Literature Review

2.1 Literature review

Accurate prediction of patients at risk of mortality is a very crucial decision when patients are hospitalized and there was need to alert health care workers about the patients at high mortality risk so as to prompt early interventions by inputting the conditions likely associated with mortality into the machine learning model.

2.2 Machine learning in child mortality prediction with malaria

Predictive analysis using machine learning came up as a very strong method in identification of risk factors having relationship with child mortality in resource constrained environments. The previous studies showed that machine learning models including logistic regression, support vector machines and random forest were effective in making analysis of large datasets and ensuring that patterns and insights were developed in child health outcomes. These algorithms enabled researchers and policy makers in identification of risk populations with implementation of target interventions in reduction of mortality rates [9].

2.3 Data- Driven risk factor identification

The traditional epidemiological methods have highly focused on known risk factors of child mortality like malaria and recent machine learning studies have put emphasis on the relevance of identification of both proximal and distal risk factors through use of big data. A study by Bizzego et al applied machine learning techniques to the multiple indicators cluster survey (MICS) dataset and there was predetermination of crucial parameters predicting under 5 child mortality some of which involved maternal education, household wealth and sanitation conditions[9]. This study's finds were key in showing the role of biological, socioeconomic and environmental factors in shaping child health outcomes [2].This study implemented the use of a random forest model where pneumonia plus congenital anomalies showed the highest impact to the model performance though its biggest gap was that the model's accuracy was less than 80% thus this study implemented hyper parameter tuning and also the other thing was about the model being trained on very many socio economic factors where during missing data treatment there was imputa-

tion with mean and median and this study is using more clinical parameters that cannot be imputed for model accuracy [10]. A study conducted by Sheraz Fatima 2024 looked at incorporation of machine learning into patient care for enhancing the outcomes of the patients and operational efficiency through utilization of big data analytics, these models helped in the derivation of proper insights from the various datasets for early detection of morbidity and enforcement of preventive health care measures. The study found out that medical diagnosis in patient outcome prediction had been reinforced through accuracy and speed plus through detection of outliers and insights that clinicians could miss and this study used deep learning machines (neural networks) trained on medical imaging datasets which proved to have surpassed the conventional methods in detection of abnormalities like cerebral edema. The predictive analytics showed hope in prediction of patient re-admissions, disease outbreaks and outcomes through the help of a mixture of historical data and real time surveillance [11]. The conclusion was machine learning model application had shown an improvement in the operational aspects of health care organizations thus enabling the clinicians handle large patient volumes with high quality care [7]. The gap in this research that my study covered was the fact that this study included laboratory and clinical parameters in the model training unlike sheraz Fatima study that used only imaging data in mortality prediction of malaria children. In the study by Joy et al, the emphasis was on the ability of smart systems to diagnose diseases in real time through the use of machine learning and the applicability was through use IOT sensors and electronic health records working hand in hand to provide real time patient monitoring for enhancement of patient outcomes for ailments like neurodegenerative disorders which was a very important aspect. For full utilization of the capabilities of machine learning models, it's highly advisable that health care systems use the data without any issues [12]. In a study conducted by Mathun Sarker, it presented a robust predictive model with capability of forecasting patient diseases with high accuracy levels with benchmarking on input information and various parameters thus enabling use of extensive datasets with various patient populations. This study implemented different machine learning techniques like logistic regression which had an accuracy of 0.796875, K nearest neighbors with an accuracy of 0.7864, XG Boost model with an accuracy of 0.78125, pytorch with an accuracy of 0.7337 where logistic regression was more accurate in-patient outcome prediction. This adoption of machine learning models in the hospital created a baseline for a more efficient patient care centered medical eco system that is supporting the complex medical needs [13]. The gap in this model created was that its confusion matrix seemed to capture a lot more false predictions therefore making the logistic regression model unreliable because it could not accurately differentiate between the true and false predictions and more so it's accuracy was below 80% because hyperparameter tuning wasn't done. A study by Addisalem Workie aimed at predicting the under-five death of children using the best performance supervised learning model where the random forest was seen to be the best model followed by the J48 algorithm with 89% sensitivity and an area under the curve of 77.8% with some factors identified that led to children failure to celebrate their fifth birthday as delay for early breast feeding [14]. The gap identified in this study was that the model that was created had a low accuracy, low area under curve thus the new model created aimed to incorporate additional features and optimize hyper parameters. A study by Lauren Erdman looked at development of machine learning models in predicting mortality using routinely collected observational health data from 0-59 months old children admitted to intensive care units in Bangladesh and these models were trained on biochemistry and hematology data and this study used 5 machine learning models and the best models were the random forest models and the gradient boosting model and the random forest was seen to be the best with the highest

area under curve of 0.87 and the top mortality predictors were carbon dioxide, potassium, creatinine and total calcium levels. My study aimed at improving on the area under curve of the model to ensure that the model could generalize well on unseen data and more so incorporate new data features [15]. In a study by Muriithi et al, the study noted that malaria was still a killer disease with great threat imposed in various parts of the world where the purpose of this study was to predict and mitigate malaria occurrence in kenya and machine learning models were used to predict the final malaria test results. In this study, the results indicated that presence of plasmodium falciparum was found to be the most important feature in classifying final test results followed by region, endemic. The use of machine learning could complement existing efforts and provide a robust data driven approach to predict malaria risks in Kenya thus leading to improvement in the public health outcomes [16]. Yadav et al study used the clinical data in prediction of machine learning malaria mortality among patients and its major focus was on investigation and validation of the efficacy of various machine learning models in mortality prediction with emphasis put on malaria parasite presence based on clinical signs and symptoms. This study was conducted in Senegal and it utilized 2 databases for malaria children and the findings showed that the random forest, support vector machine with Gaussian Kernel and artificial neural networks delivered promising and accurate results on both datasets with accuracy of 92%, recall of 85% and F1 score of 89% for all the models thus concluding that machine learning models can easily detect the existence or the absence of malaria according to medical information [17].

2.4 Importance of feature selection and model performance

The efficacy of machine learning models in child mortality prediction highly relies on the levels of data pre processing done when wrangling the dataset like feature scaling plus encoding techniques. Research by Zhang and Li showed that if feature selection methods were added into the pre processing levels, this improves the accuracy and reduces on the model over fitting thus making better predictions. When models are trained studies recommend that evaluation metrics should be employed to evaluate the performance like the receiver operating characteristics (ROC) and Matthews Correlation Coefficient (MCC) in assessment of model performance [18].

2.5 Challenges in implementation machine learning models for malaria mortality prediction

There are several challenges that affect the implementation of ML models in health care due the absence of quality data as many datasets contain missing values and inconsistences which highly affect the performance of the model accuracy plus the difficulty involved when it comes to interpretability of complex machine learning models who highly rely on actionable insights. These machine learning models are highly affected by the nature of the datasets used and more so the dataset having various input features added because the patient outcomes highly depend on the various inputs used in the model which could have not been collected in the datasets [19].

2.6 Outcomes of research

The study done had a target of reduction on the mortality rates among the children hospitalized with severe malaria and this looked at the factors affecting patient outcome and later on leading to deaths for example a child might have hemoglobin levels that are high from the complete blood count but due to some underlying factors like the acidosis because of high potential of hydrogen ions in the blood, the child might deteriorate later on therefore with these machine learning models in place all the necessary features are captured and early intervention will be given to the patients.

2.7 Gaps in previous research

A model developed by Bizzego et al had limited accuracy and poor tuning since the random forest model that he developed achieved only an accuracy level of 75% meaning that its performance was suboptimal and the major gap here was that there was no hyper parameter tuning used and more so the model was trained on socio economic imputation techniques that reduced the model's precision therefore the novelty of this research addressed this by use of hyper parameter tuning and use of non-imputable clinical features thus improving model accuracy and robustness. A study by Sheraz Fatima used much of the medical imaging data for prediction where critical features for the clinic and laboratory were not captured thus limiting the comprehensive general picture of malaria mortality prediction therefore this research addressed this through integrating the laboratory and clinical features and then developing a multi modal approach that favored malaria mortality prediction. The other study by Mathun Sarker that found out that the logistic regression was the best model to use because of the high accuracy levels later on got to realize that it's confusion matrix captured more false predictions rates thus lacking reliability and generalization therefore this study looked at inclusion of various metrics like precision, recall, f1 score, AUC to ensure better classification models are adopted. Lauren Erdman study used biochemistry data and trained a random forest model though its AUC was big 0.87, the model was biased by only patients with malaria parasites therefore this study included patients whose parasites cleared but were still vulnerable to mortality to ensure that the model is not biased on a particular group of people [20].

2.8 Future Directions and policy implications

Future research should focus on the improvements and the interpretability of the machine learning models plus integration of real time data sources through use of explainable AI techniques that can help out between model outputs and clinical decision making thus making predictions more accessible to non-stakeholders.

Chapter 3

Methodology

Materials and Methods In conducting the analysis for prediction, we utilized a robust set of tools and software to ensure accuracy and reliability. All the statistical procedures conducted in the study were implemented using the most recent version of python 3.14.3, a widely recognized open-source statistical tool known for its versatility and extensive data analysis capabilities. Specifically, the predictive modeling was performed using the scikit learn package in Python. The package provided a comprehensive framework for building, training and evaluating predictive models making it particularly suitable for our study's objectives.

3.1 Study area

Uganda is located in the Eastern part of Africa with an estimated population of 51,384,894 people. It has a total land area of 241,038 square kilometers and it has longitude and latitude of 1.3733° N, 32.2903° E. Being a malaria endemic country it has various hospitals where our study was performed from Mulago national referral hospital as the case study. Mulago National referral is located in Kawempe North Division of Kampala in the capital city of Uganda which offers various services like pediatric, surgery and internal medicine services with their subspecialties. Despite their size and significance, the hospitals are often overwhelmed by the high volume of patients seeking medical care, far exceeding their capacity [21]. This frequently results in patients being treated in overcrowded conditions, with some even having to sleep on the floor due to limited bed space.

3.2 Data Source

The specific contribution of our study based in utilizing a clinical dataset to predict malaria mortality among pediatric patients. This extended the existing knowledge by leveraging advanced analytical methods for a targeted and nuanced understanding the factors that led to mortality of malaria children in hospital. The clinical dataset was obtained from World Health Organization open databank containing both severe malaria anemia cases plus the outcomes of each participant. The variables in the dataset included patient outcome which was the target variable, demographics like age, sex, clinical factors included chills, fever, fatigue, sweats, headache. Laboratory markers included hemoglobin, platelets, white blood cells, parasitemia and finally the socio-economic factors included region. The target variable for the study was the outcome of hospitalization which was classified under 0 for died and 1 for survived. During class distribution

assessment, the dataset showed class imbalance where survival cases were observed to outnumber the mortality cases by 676 to 44 respectively.

3.3 Study design and population

The study adopted an experimental computational research design using retrospective clinical and laboratory data to train and evaluate machine learning models for mortality prediction among pediatric malaria patients. This design was good because the study involved training, testing and comparing multiple supervised machine learning models to determine the best performing model. The study experiment evaluated how different models performed when exposed to the same dataset and predictor variables. Since the dataset used was previously collected from the hospital thus bringing about the retrospective element. The study involved patients admitted from 1 week to 17 years of age. All the children with severe malaria had plasmodium falciparum on blood smear, children with severe malaria anemia had hemoglobin level of less than 5g/DL [24]. The sample size of 720 patients was determined through the use of the William Conchran sample size determination formular;

$$n = \frac{Z^2 P(1-P)}{E^2}$$

Where: n= sample size, N = population size, Z score corresponding to the desired confidence level, P = proportion of the population with the characteristic of interest (use 0.5 if unknown) and finally E = Margin of the error (3.65%).

3.4 Data cleaning and pre-processing

Rigorous steps were taken in assurance of dataset integrity and reliability. There were some missing variables found but since they were not 30% of the dataset they were dropped and this being a clinical dataset imputation with mean, median would not be advisable as it affects the findings. The dataset underwent a comprehensive normalization process in enhancement of machine learning effectiveness. This step was critical in enhancing quality and effectiveness of machine learning algorithms. Normalization was applied to the numerical independent variables using the min-max normalization technique thus transforming features into a shared range mitigating the impact of larger numeric values dominating the model's learning process.

$$\frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

Where Xmax and Xmin represented the variables maximum and minimum values respectively. Without normalization, features with larger numeric values would have exerted undue influence on the learning algorithms thus overshadowing the contributions of smaller scale features. The relevance of normalizing data was more in its ability to create a proper environment for numeric features in order to ensure that every variable had a proportionate contribution to the learning process thus contributing much to the overall robustness and reliability of the machine learning models used in malaria mortality predictions in the hospital. Additionally, the dataset was analyzed for duplicates to ensure that the data observations that repeat each other were dropped to enable the model become reliable and not to over fit on a particular group of patients. More so the dataset was analyzed for outliers and they were found though along the normal ranges therefore they were not dropped. This thorough preprocessing laid foundation for the training of better machine learning models.

3.5 Feature engineering and selection.

First of all, the categorical variables in the dataset were changed to numeric using the one hot encoder and the scaling was used to ensure that the variables make sense and incase of class imbalance the synthetic minority over sampling technique was used to ensure that the classes balance to ensure that the model is not biased towards the majority class and capable of learning patterns from both classes of patient outcomes. This approach prevented the model from over fitting, enhanced the reliability of performance estimates and ensured a resource efficient use of available data. Initially the class imbalance on the training set was 541 for class survived and 35 for class died but after resampling both classes got an equal distribution by 541 per class. Statistical analysis was performed on the features in the dataset to enable the determination of which some of them had a positive relationship with the target variable and since the target variable was categorical, the analysis of variance test was implemented to check for the correlation between the variables and the chi square was also used to assess the relationship between the categorical target and the other variables. For the case of visualization, the box plots were used to show the relationships between numeric and categorical features, then the histograms were used to show the nature of distribution of the quantitative variables and the bar plots were used to show the trends and patterns in the categorical variables. Multicollinearity was checked for in the dataset using variance inflating factor though it wasn't seen anywhere.

3.6 Machine learning algorithms

The study applied 5 different machine learning models; XG Boost model, Decision tree classifier, Random Forest model, Logistic regression and naïve bayes in prediction of malaria mortality. Because of their wide spread use and high predictive accuracy, these models are versatile and very much used in classification tasks. This involved categorizing the instances into predefined classes like prediction of mortality for a child admitted in the hospital. Decision tree classifier was used which had hierarchical structures with root, branches, internal nodes and leaf nodes and through the support of a divide and conquer method, the ideal split points were found using a top-down greedy search. S was the set of samples, X the set of features, and H(X) the information content, then the DT algorithm aimed to recursively partition S into subsets Si using X and H(X) [22]. Random Forest was used an ensemble technique relying on multiple decision trees and when decision trees are grown very deeply, the training data are often over fitted leading to a significant fluctuation in classification results for a little change in the input data. They became extremely sensitive to the training data as well as error prone to the dataset as a result. The random forests often found solution to this issue by training different parts of the training dataset using multiple decision trees and this helped in reduction of bias, tolerate outliers, avoid over fitting and it was much less sensitive to the training data [23]. T was the set of decision trees, pi the prediction of each tree, and y hat the final prediction, then the random forest can be formulated as;

$$\hat{y} = \arg\max_{c_i} \sum_{j} I(p_{ij} = c_i)$$

XG Boost was used an efficient ensemble technique based on gradient boosting. It minimized a cost objective function, J, comprised of a regularization term and loss function.

The update rule for each iteration was expressed as;

$$\hat{y}_{k+1} = \hat{y}_k - \eta \frac{\partial J}{\partial \hat{y}_k}$$

The logistic regression models binary variables typically indicated whether an event would occur or not. X was the input data, w the weights, b the bias, and the the sigmoid function, then the logistic regression prediction was written as:

$$\hat{y} = \sigma(\mathbf{w} \cdot \mathbf{X} + b)$$

Naïve Bayes classifier was trained which used probability assuming that features are independent given a particular class. During machine learning model training, the study used Grid Search to hyper parameter tune the model in order to ensure that the study optimizes the model to the maximum in terms of performance [23]. In the context of training and testing data, the initial random split of the dataset into these sets showed that there was presence of class imbalance, particularly in the outcome of the patients variable and in derivation of a solution to this problem faced during the training phase, the synthetic minority over sampling technique was implemented which is a method used by creating synthetic samples for the minority class thus enabling reduction in the bias in model training and improving the performance of the classification algorithms.

3.7 Model validation and evaluation

To robustly assess the performance of each model during the training phase, we employed the K fold cross validation and this involved random division of the training data into 5 mutually exclusive folds of roughly equal size then each model was trained and tested 5 times. This process led to 5 performance estimates for each model like accuracy, precision, F1 score, Recall and the final estimate was based on the mean scores of the 5 estimates. The rationale for use of cross validation came from its effectiveness in provision of a robust evaluation of model performance through division of training data in 5 folds thus providing them with thorough exposure to diverse data patterns. The trained models were evaluated on accuracy, precision, recall, F1-score, confusion matrix, Receiver operating characteristic (ROC) curve and the Area Under the curve (AUC). Accuracy measured how the model predicted the outcome of the patients correctly and this measurement meant the division of total number of correct predictions out of all the predictions. Precision was also used which was an evaluation metric that measured the proportion of positives correctly. Recall/ sensitivity was a factor that measured the model's ability to correctly identify positive instances out of the total number positive samples. F1 score was the harmonic mean of precision and recall, it acted as the balance between the 2 parameters and their formulae as below;

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision}_i = \frac{TP_i}{TP_i + FP_i}$$

$$\text{Recall}_i = \frac{TP_i}{TP_i + FN_i}$$

$$F1\text{-}Score_i = 2 \times \frac{\text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i}$$

Confusion matrix was an evaluation metric that summarized the predictions made by the model against actual true labels [1].

Figure 3.1: Confusion matrix

It displayed the accurate and inaccurate values using the models predictions and lastly the Receiver Operating Characteristic (ROC) curve and Area under the Curve (AUC) which was plotted using the True Positive Rate (TPR) and False Positive Rate (FPR).

$$TPR = \frac{TP}{TP + FN}, \quad FPR = \frac{FP}{FP + TN}$$

AUC-ROC. ROC was the probability curve and AUC represented the degree of separation and this also showed the level a model can be able to distinguish between classes where a model having a very high AUC meant its performance was good unlike the ones with low AUC. Learning curves that ensured that the study could easily identify model over fitting through assessment of the training score curve and the performance of the cross-validation score curve therefore this was used to show the level of biasness and variance levels in the different models.

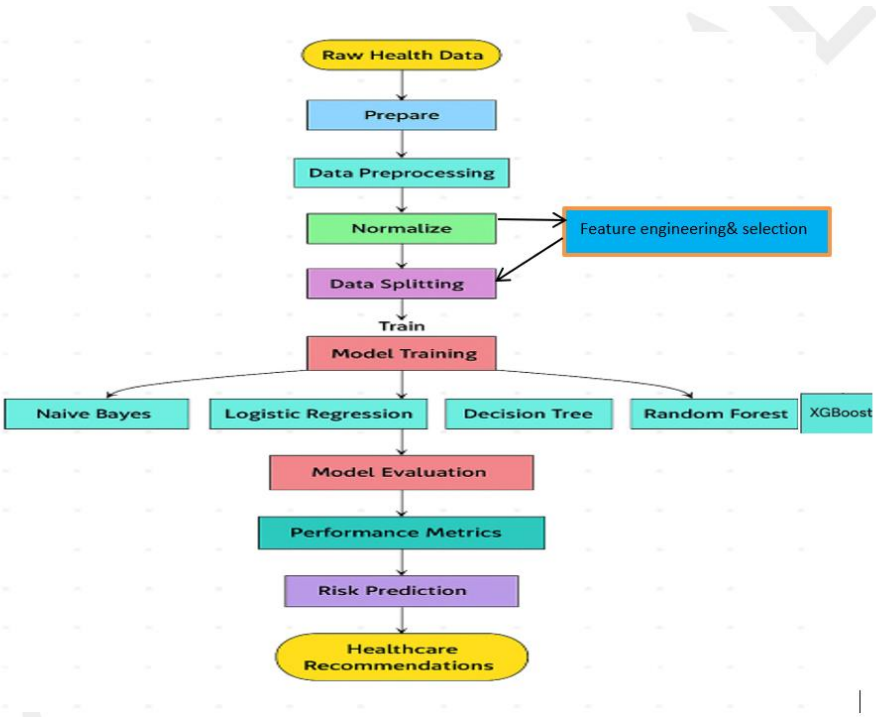


Figure 3.2: Machine learning pipeline for pediatric malaria in hospital mortality

3.8 Model interpretability and insights

To ensure the model was transparent and usable in clinical settings, interpretability tools were employed. Feature importance rankings from Random Forest and Global explanations that use permutation feature importance values were used to explain individual and global predictions [24]. These features worked through attributing each feature's contribution to the final predictions through offering visual explanations for why certain features influenced a prediction, enhancing trust and adoption by healthcare professionals for proper decision making.

3.9 Limitations to the study

The machine learning models had a very high importance in health care though their implementation and practicability was so much limited by the availability and quality of the data since the missing data values were so common plus the inconsistencies affecting the model accuracy. Then the other factor looked at the difficulty in interpretability of complex machine learning models thus causing challenges to the clinicians since they depend much on insights that are actionable [23].

3.10 Ethical considerations

The study utilized secondary data obtained from World Health Organization open access database and since it was anonymized with no personally identifiable data, there was no ethical approval or informed consent required.

Chapter 4

Results

Results and findings The dataset utilized in this study comprised of variables which had a major goal of understanding the factors that led to mortality among malaria children in the hospital. The variables encompassed both clinical and laboratory factors for various children admitted in the hospital as inpatient. Through the use of exploratory data analysis to find out the key insights from the dataset regarding the predictors of mortality among malaria some visualizations were used in the study and 5 supervised machine learning models were trained. The findings showed that the mean age was 6 from a

Table 4.1: Summary of Numeric Variables

Numeric Variables	**Mean**	**Standard Deviation**	**Range**
Age	6	3.46	0.1 to 17
White Blood Cells	1898.36	5693.41	9.19 to 40200
Hemoglobin	6.64	3.37	0.6 to 14.9
Platelets	6.17	3.53	1.5 to 22.6

range of children admitted From 0.1 Years to 17 years of age, the white blood cells had a mean of 1898.361997 cells per microliter (cells/µL) with a standard deviation of 5693.4080 cells/µL. Because of anemia, the mean hemoglobin was 6.638540 g/dl with the minimum of 0.6 and a maximum of 14.9 g/dl and finally the platelet levels had a mean of 6.172651 thousand per micro litre (10^3/µL).

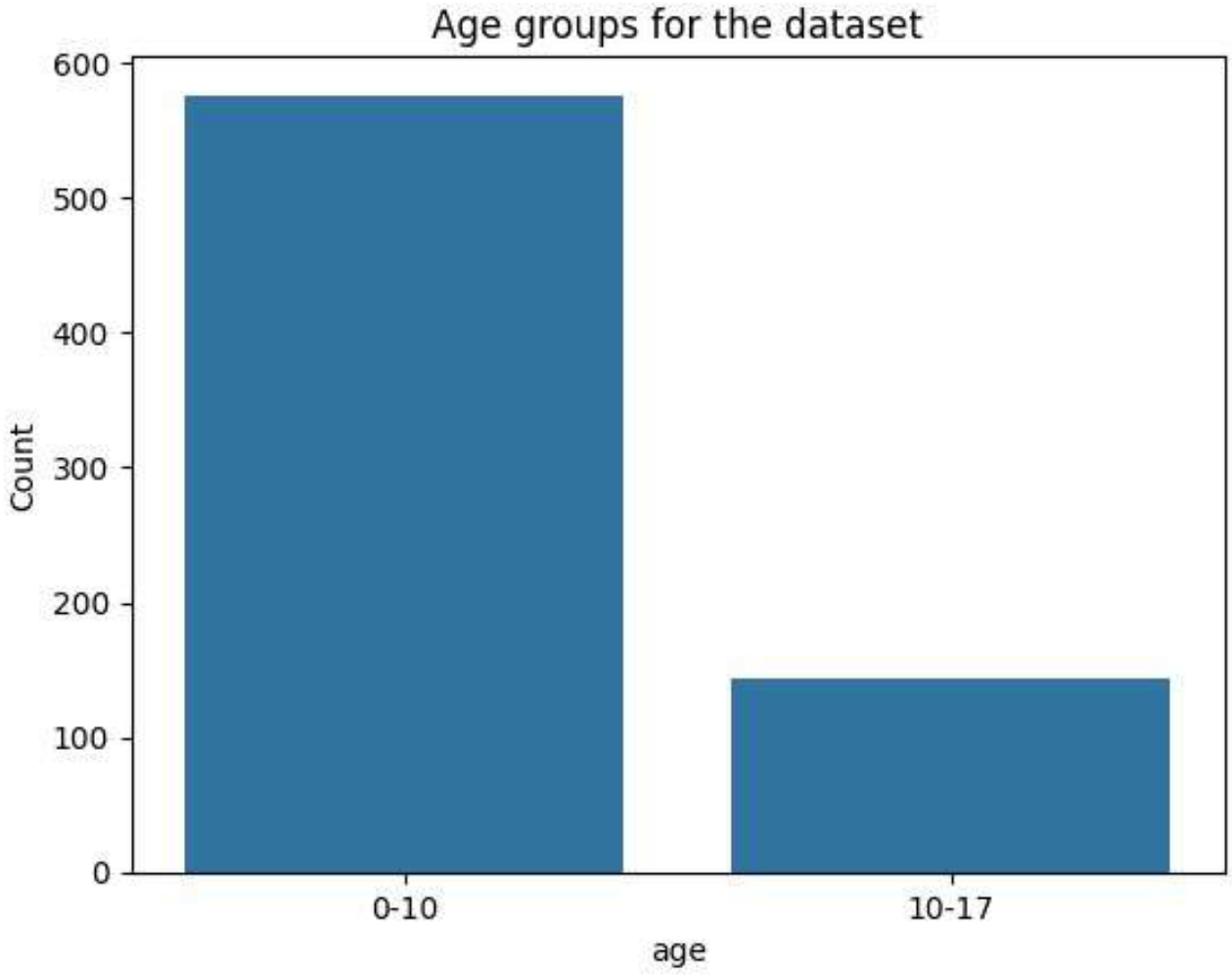


Figure 4.1: Age bins for the children

Figure 4.1 showed that most of the children admitted with severe malaria were between 0.1 to 10 years (576 counts) followed by those that were between 10 to 17 years by (144 counts) That is, malaria was seen to be more in younger children than in older ones.

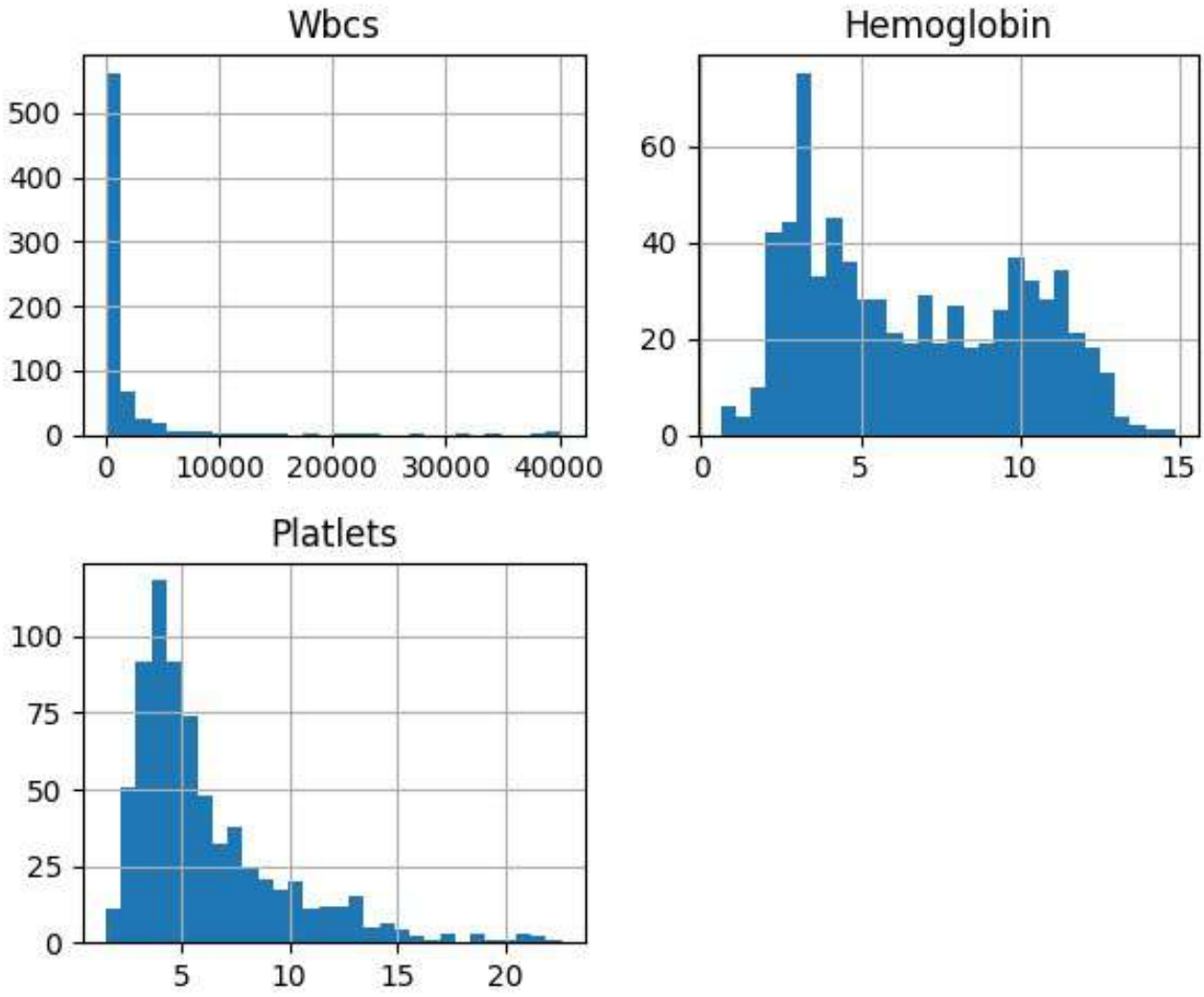


Figure 4.2: Distribution of the numeric variables

Figure 4.2 showed the distribution of the study variables that are numeric and the findings showed that most of the children had lower white blood cells and hemoglobin, which meant that they were anemic and more so they had low platelet levels, which implied that they had thrombocytopenia.

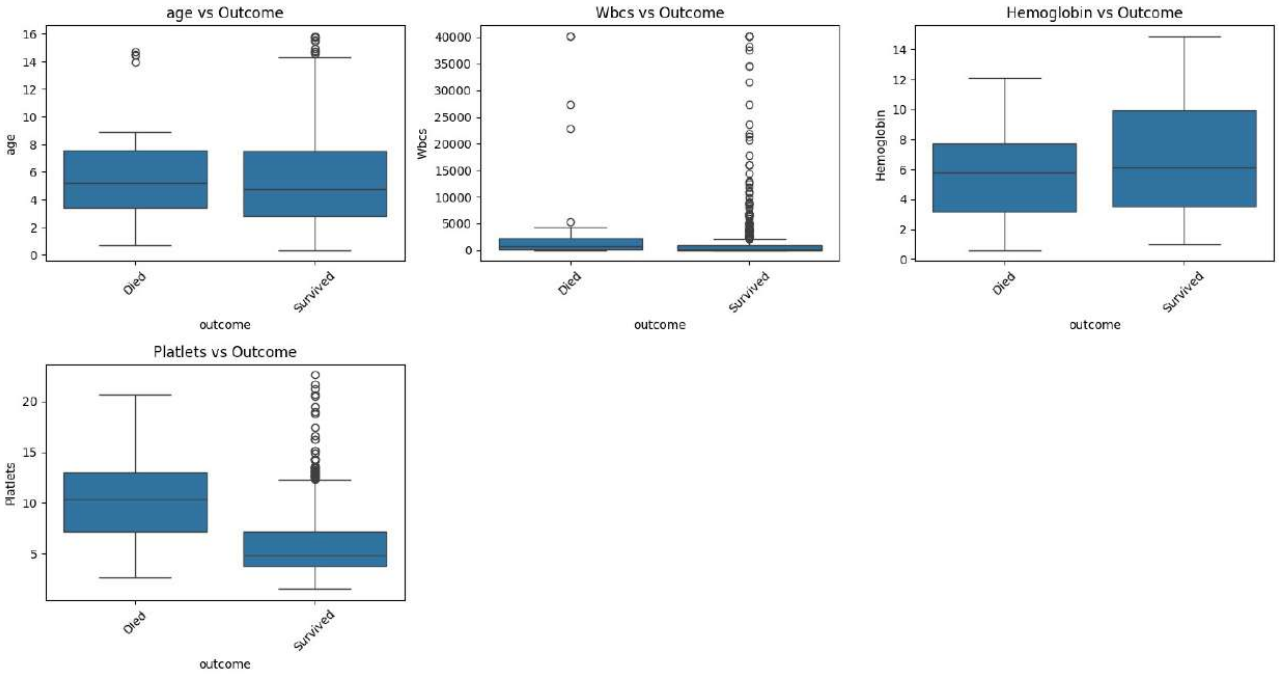


Figure 4.3: Box plots for numeric against categorical variables

From figure 4.3,death was seen to be more in children that had lower hemoglobin levels. The children that survived had higher hemoglobin levels. Due to the immune system detecting malaria parasites, the children who died had more white blood cells produced in the body compared to those who survived. The platelet levels for children that survived had outliers in the upper bound and those that died the platelets were skewed to the lower bound. The age of the children did not have any significant impact on the outcome of the patients as the distribution does not show any much difference.

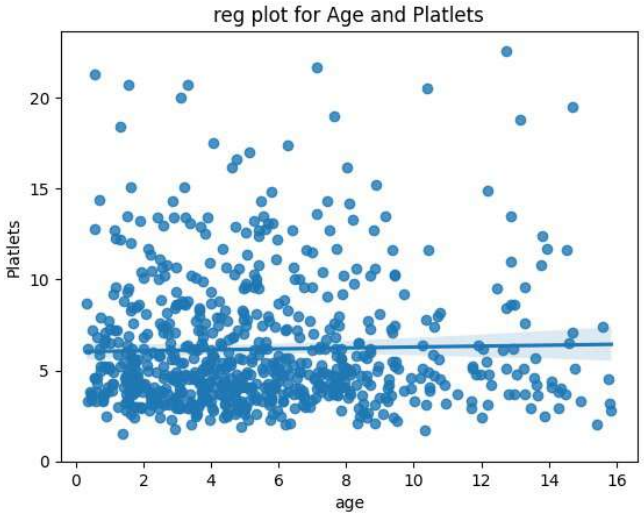


Figure 4.4: Regression plot of age against platelets

The Figure 4.4 showed that there was a slight positive relationship between the platelet levels of children with age though majority of the points were not along the line of best fit. Platelet levels slightly increased as the child grew though it wasn't so much significant.

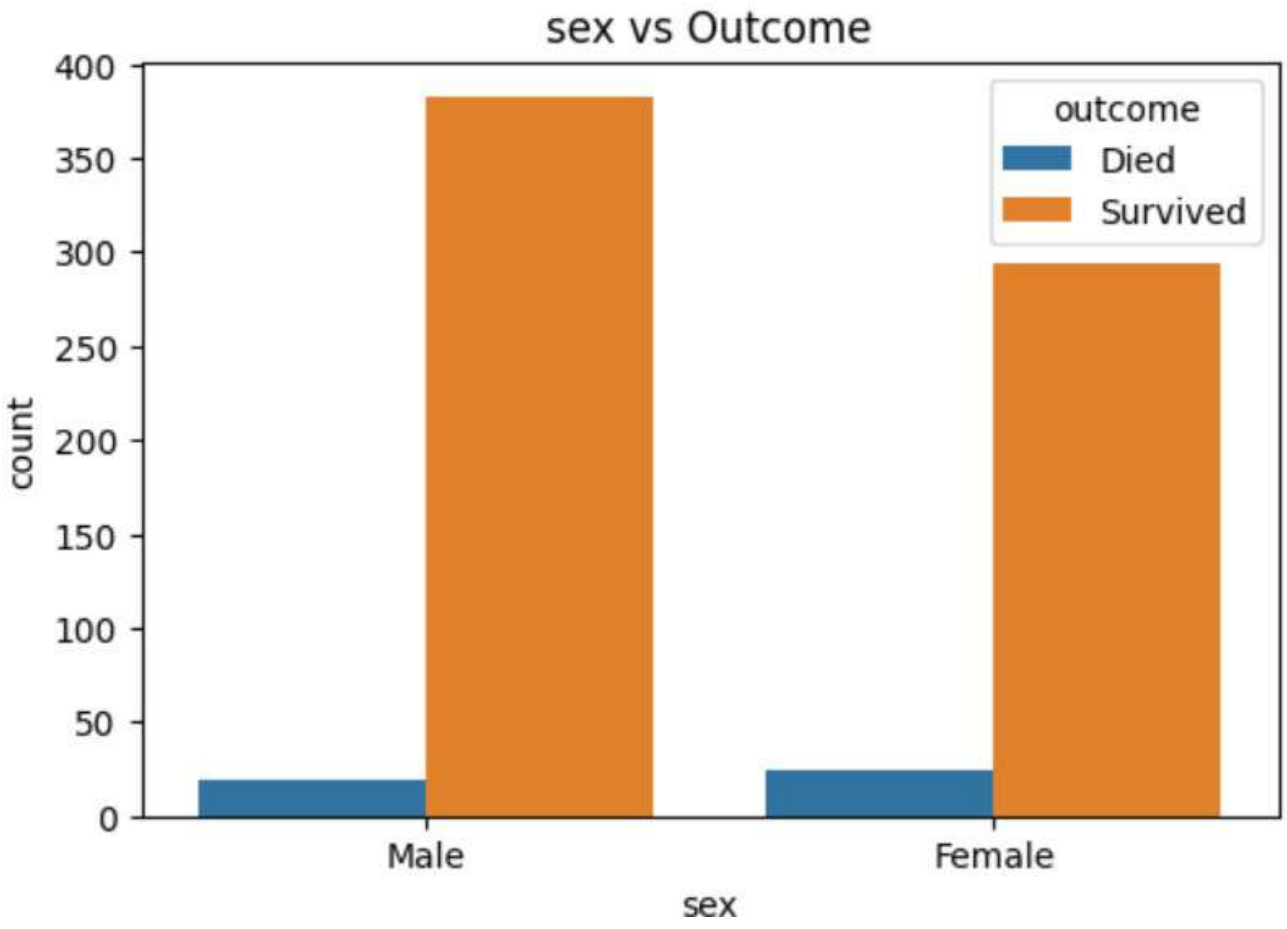


Figure 4.5: Bar plot of sex against the outcome

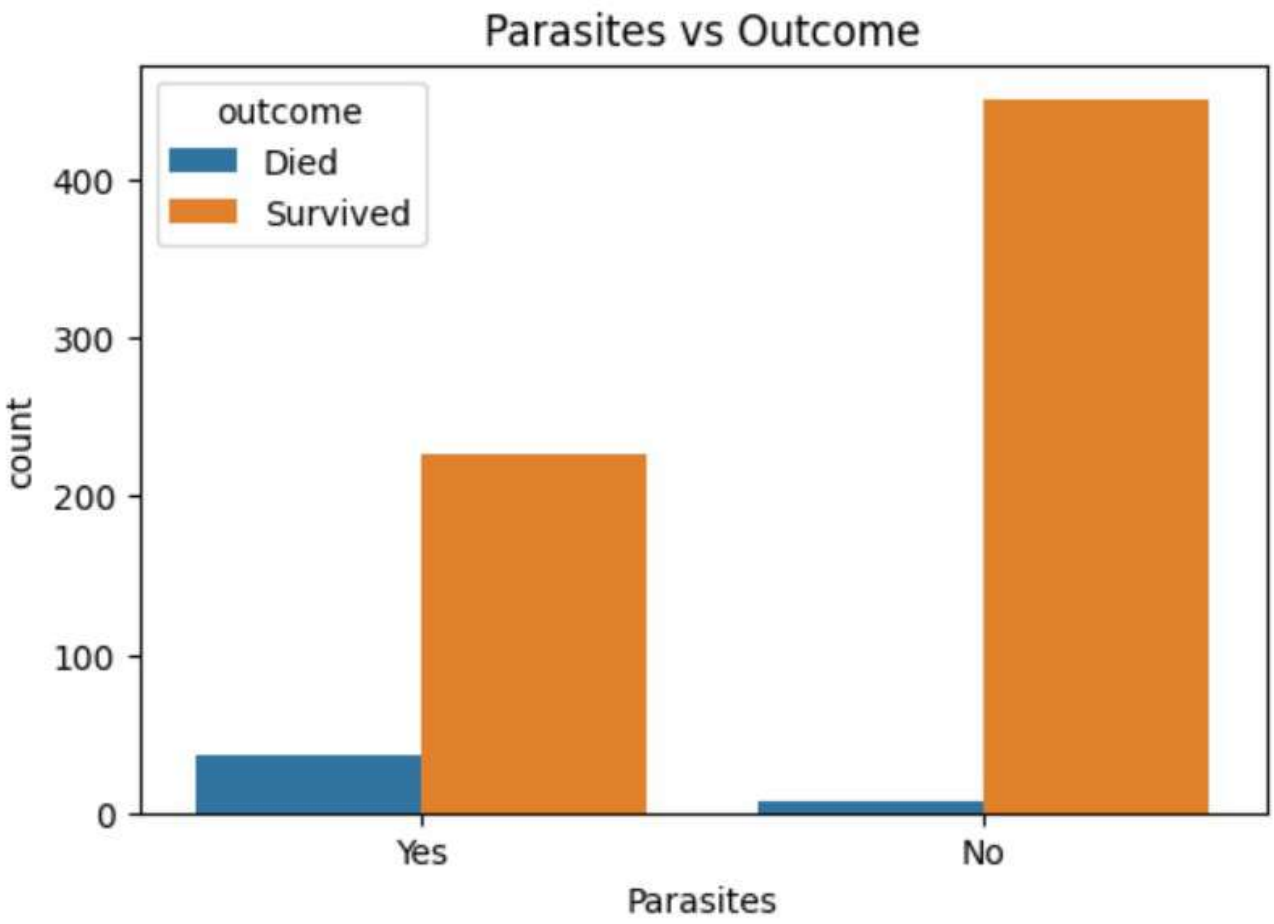


Figure 4.6: Bar plot of parasites against the outcome

Figure 4.5 showed that mortality was observed to be more among the females (25 counts) than the males (19 counts) and the males (380 counts) who survived were more than the females (296 counts) who survived. Figure 4.6 showed that mortality was observed to be more in children that had more malaria parasites present in their bodies unlike those that died without parasites present in their bodies. Survival was observed to be more in children without malaria parasites in their bodies with 485 counts. The summary observed was that the females were more vulnerable to malaria mortality and malaria was seen to have a real impact to mortality of children.

4.1 Feature engineering

Feature engineering was used to create a variable hemoglobin category and this was based on the World Health Organization definition of anemia where for children with hemoglobin levels greater than or equal to 11 g/dl is referred to as normal, then for ranges between 8 to 11 g/dl is referred to as mild anemia, 5 to 8 g/dl is referred to as moderate anemia and hemoglobin levels less than 5 g/dl is referred to as severe anemia.

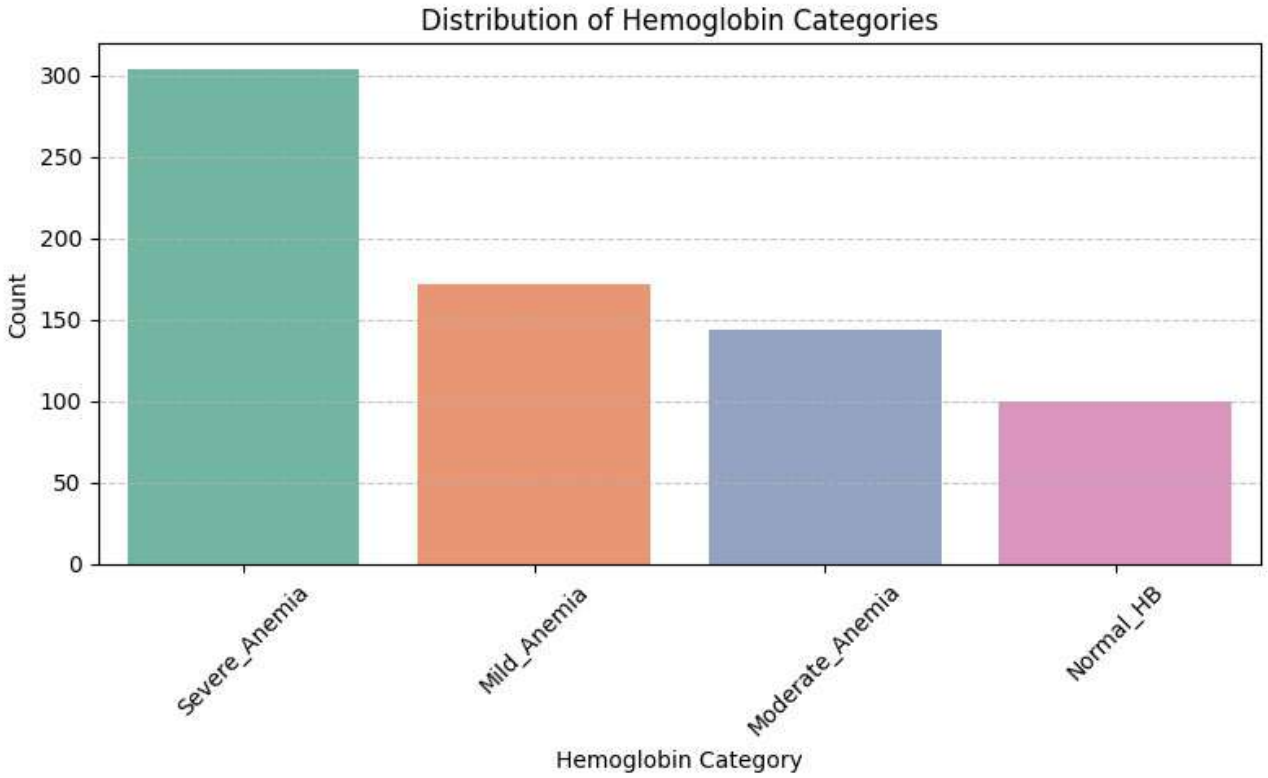


Figure 4.7: Hemoglobin categories

The figure 4.7 showed that most of the children admitted had severe anemia with 310 counts, mild anemia had 170 counts, moderate anemia had 140 counts and the children who had normal hemoglobin levels were 100. Malaria was observed to have a significant impact on the hemoglobin levels of the children and this backs up the literature and science of malaria parasites attacking the red blood cells that produce the hemoglobin in the blood.

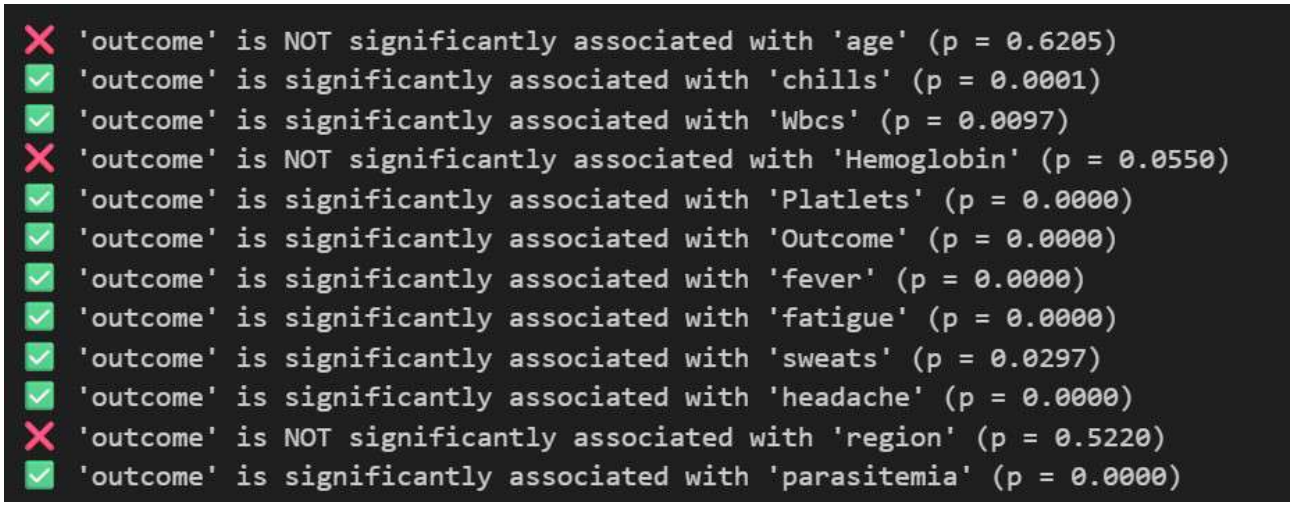


Figure 4.8: Analysis of variance

The figure 4.8 showed that the variables correlated to the target variable were chills, white blood cells, platelet levels, fever, fatigue, sweats, headache, malaria parasitemia since they had a probability value of less than 0.05. The variables that were not statistically significant were age, hemoglobin, region and gender. For feature selection, the features that were correlated to the target variable were used in the machine learning model.

4.2 Machine learning modeling

For prediction purposes, the machine learning models were trained in order to assess which one gets us the best predictions without being biased, over fitted on training data and this was assessed through evaluation metrics like precision, recall, F1 score, confusion matrix at k fold cross validation in assurance of consistency of prediction per fold, receiver operating characteristic area under curve, learning curves for assessment of model over fitting or not. During model training at k fold cross validation, the optimal features of grid search used in getting the best model were also captured in the table.

Table 4.2: Optimal Tuning Parameters for Models

Models	**Optimal Tuning Parameters**
Decision Tree classifier	Criterion = entropy, max_depth = 10, min_samples_leaf = 4, min_samples_split = 10
Naïve Bayes classifier	10^{-9}
Random Forest Model	Bootstrap = True, max_depth = None, min_samples_leaf = 1, min_samples_split = 5, n_estimators = 100
Logistic Regression	C = 10, max_iter = 200, penalty = l2, solver = lbfgs
XGBoost	Colsample_bytree = 0.8, learning_rate = 0.1, max_depth = 5, n_estimators = 200, subsample = 1.0

The table 4.2 showed that despite of the fact that some hyper parameters had variation across the folds, there were relatively minor differences and there was not any substantial effect on the average model performance where the final model for each machine learning algorithm was retrained by the help of the best overall parameter set on the full training dataset. tabularx The decision tree was evaluated on 5 folds and the overall accuracy of

Table 4.3: Decision Tree Model Performance Evaluation using K-Fold Cross Validation

Fold	**Precision (Died/Survived)**	**Recall (Died/Survived)**	**F1 Score (Died/Survived)**	**False Predictions**
1	0.92 / 0.96	0.96 / 0.92	0.94 / 0.94	13
2	0.90 / 0.92	0.93 / 0.90	0.91 / 0.91	19
3	0.85 / 0.97	0.97 / 0.83	0.91 / 0.90	21
4	0.89 / 0.91	0.91 / 0.89	0.90 / 0.90	22
5	0.85 / 0.95	0.95 / 0.83	0.90 / 0.89	23
Test Set Accuracy: 0.8681				

the model based on the test set which was 86.81% which showed that the model was good at its prediction. The precision of the model across all the folds for class died ranged from 0.85 to 0.92 and the precision for class survived ranged from 0.91 to 0.97 meaning that when the model predicted a patient outcome it was often correct in a range of 85 to

97%. The recall values for the class died ranged from 0.91 to 0.97 meaning that among that class the model could accurately predict the children who died in a range of 91 to 97%. For the class survived the model accurately predicted the children who survived in a range of 83 to 92% meaning that a model was able to accurately predict the true cases for an outcome in a range of 83 to 97%. The balance between precision and recall was assessed by the F1 score and for class died it had a range of 0.90 to 0.94 then for class survived the range was 0.89 to 0.94. The confusion matrix showed a range of false prediction between 13 to 23 counts which was relatively a small proportion as compared to number of correct predictions across the folds thus implying that the model was reliable across the data splits.

Table 4.4: Random Forest Model Performance Evaluation using K-Fold Cross Validation

Fold	**Precision (Died/Survived)**	**Recall (Died/Survived)**	**F1 Score (Died/Survived)**	**False Predictions**
1	0.92 / 0.99	0.99 / 0.91	0.95 / 0.95	11
2	0.97 / 0.96	0.96 / 0.97	0.97 / 0.97	7
3	0.86 / 0.99	0.99 / 0.84	0.92 / 0.91	18
4	0.88 / 0.91	0.92 / 0.88	0.90 / 0.90	22
5	0.89 / 0.94	0.94 / 0.88	0.91 / 0.91	19
Test Set Accuracy: 0.9028				

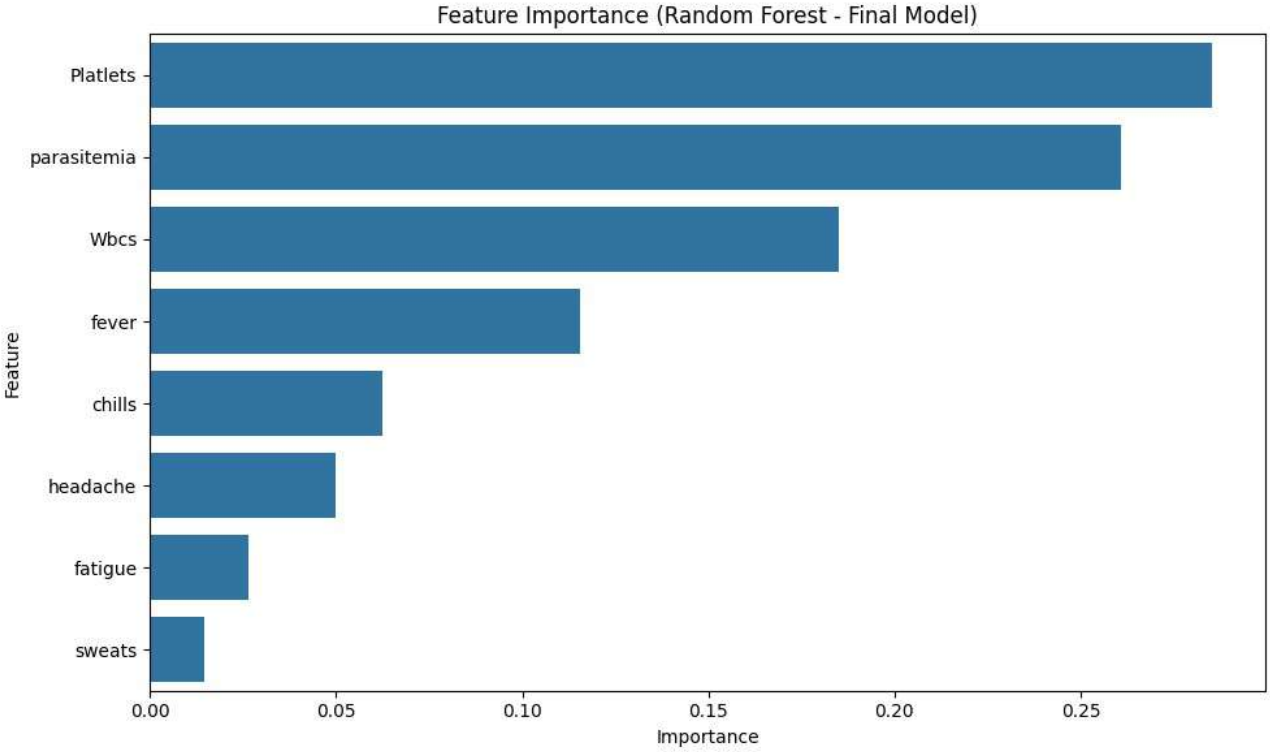


Figure 4.9: Random Forest Model feature importance

The random forest model had an overall test accuracy of 90.28% meaning that its performance at prediction was excellent and through its evaluations across the folds, it had precision of 0.86 to 0.97 for class died meaning that the model accurately predicted class died between 86 to 97% and for class survived the model's precision was ranging from 0.91 to 0.99 meaning that it accurately predicted class survived in a range between 91 to 99% which meant that it was excellent at identifying the right predictions in a range of 86 to 99%. Recall was also used in evaluation which was ranging from 0.92 to 0.99 for class died implying that at a scale 92 to 99% the model accurately predicted the true children who died among that class and then for class survived recall was 0.84 to 0.97 meaning that of that class the true prediction for children who survived were ranging between 84 to 97% thus making the model a good one. The balance between precision and recall was assessed by the F1 score and for class died it was ranging from 0.90 to 0.97 which implied that the model was good at identifying class died correctly in a range of 90 to 97% and for class survived its F1 score was ranging between 0.91 to 0.97. The false predictions were assessed by the confusion matrix but they had a range of 7 to 22 observations seen across the folds which number is minor as compared to the true predictions. Feature importance was assessed for the model to find variables that play a very significant role in the model performance and the figure showed that platelets counts do play a key role in mortality of children especially due to thrombocytopenia followed by parasite presence, white blood cell count, fever, chills, headache, fatigue and lastly sweats.

Table 4.5: Naïve Bayes Model Performance Evaluation using K-Fold Cross Validation

Fold	**Precision (Died/Survived)**	**Recall (Died/Survived)**	**F1 Score (Died/Survived)**	**False Predictions**
1	0.87 / 0.90	0.91 / 0.86	0.89 / 0.88	25
2	0.82 / 0.87	0.88 / 0.81	0.85 / 0.84	34
3	0.78 / 0.89	0.91 / 0.75	0.84 / 0.81	37
4	0.85 / 0.85	0.85 / 0.85	0.85 / 0.85	32
5	0.82 / 0.91	0.92 / 0.81	0.87 / 0.85	30
Test Set Accuracy: 0.8125				

The Naïve Bayes classification had an average testing accuracy of 81.25% meaning that it was good at predicting the mortality of children with malaria. The model's precision for class died was between 0.82 to 0.87 meaning that it could accurately predict class died on a range of 82 to 87% across the 5 folds then for class survived it could accurately predict 85 to 91% children that survived in the hospital thus making it a good model. Recall was also used where class died had a performance range of 0.85 to 0.92 implying that of the correct predictions for class died made by the model, 85 to 92% of them were actually correct and then for class survived 75 to 86% of the children predicted to have survived actually accurately survived thus making the model a good one for predictive analysis. The F1 score was also used in the evaluation and it showed that there was a balance between precision and recall for class died by 85 to 89% then for class survived the balance was between 81 to 88%. After analysis of the confusion matrices across the folds, the false predictions ranged from 25 to 37 which was minor as compared to the correct predictions therefore the model's performance was good. The logistic regression model achieved an overall test accuracy of 77.78% meaning that its performance at predicting the outcome was fairly good and from the assessment of precision, it had a range of 0.80 to 0.85 for class died meaning that it could accurately predict class died across the folds

Table 4.6: Logistic Regression Model Performance Evaluation using K-Fold Cross Validation

Fold	Precision (Died/Survived)	Recall (Died/Survived)	F1 Score (Died/Survived)	False Predictions
1	0.82 / 0.99	0.99 / 0.79	0.90 / 0.88	24
2	0.85 / 0.91	0.92 / 0.84	0.88 / 0.88	26
3	0.80 / 0.98	0.98 / 0.76	0.88 / 0.85	28
4	0.84 / 0.96	0.96 / 0.81	0.90 / 0.88	24
5	0.80 / 0.96	0.97 / 0.76	0.88 / 0.85	29
Test Set Accuracy: 0.7778				

in a range of 81 to 85% then for class survived the range was 91 to 98%. The recall of the model for class died had a range of 0.92 to 0.99 meaning that of the predictions of class died it correctly identified 92 to 99% of the children that died then for class survived of the true predictions made for that class, the model correctly identified 76 to 84% of the children that survived. The model's balance between precision and recall for class died was ranging from 88 to 90% and then for class survived it was ranging from 85 to 88% which performance wasn't bad. Through the confusion matrix analysis, the model had a range of 24 to 28 of the false observations seen but since they were minor as compared to the total number of true predictions the model was acceptable. The model's average

Table 4.7: Gradient Boosting Model Performance Evaluation using K-Fold Cross Validation

Fold	Precision (Died/Survived)	Recall (Died/Survived)	F1 Score (Died/Survived)	False Predictions
1	0.92 / 0.98	0.98 / 0.92	0.95 / 0.95	11
2	0.95 / 0.94	0.94 / 0.95	0.94 / 0.95	12
3	0.86 / 0.99	0.99 / 0.84	0.92 / 0.91	18
4	0.86 / 0.94	0.94 / 0.85	0.90 / 0.89	22
5	0.86 / 0.95	0.95 / 0.84	0.90 / 0.89	22
Test Set Accuracy: 0.8681				

test accuracy was 86.81% meaning that the model was good at predicting the outcome variable. The precision of the model for the class died was ranging from 0.86 to 0.95 meaning it accurately predicted class died across all the 5 folds at a rate of 86 to 95% and for class survived it ranged from 94 to 99%. The recall of the model showed that for class died 94 to 99% was the model's capacity to correctly predict children that died across that class and for class survived it had the capacity to predict 84 to 95% of the children that correctly survived under that class. Then the F1 score for class died was between 90 to 95% implying that it balanced precision and recall highly for class died and for class survived it was 85 to 95% thus meaning it was a good model. The false predictions for the model ranged between 11 to 22 and since the numbers were minor as compared to the correct true predictions, the model had a good performance.

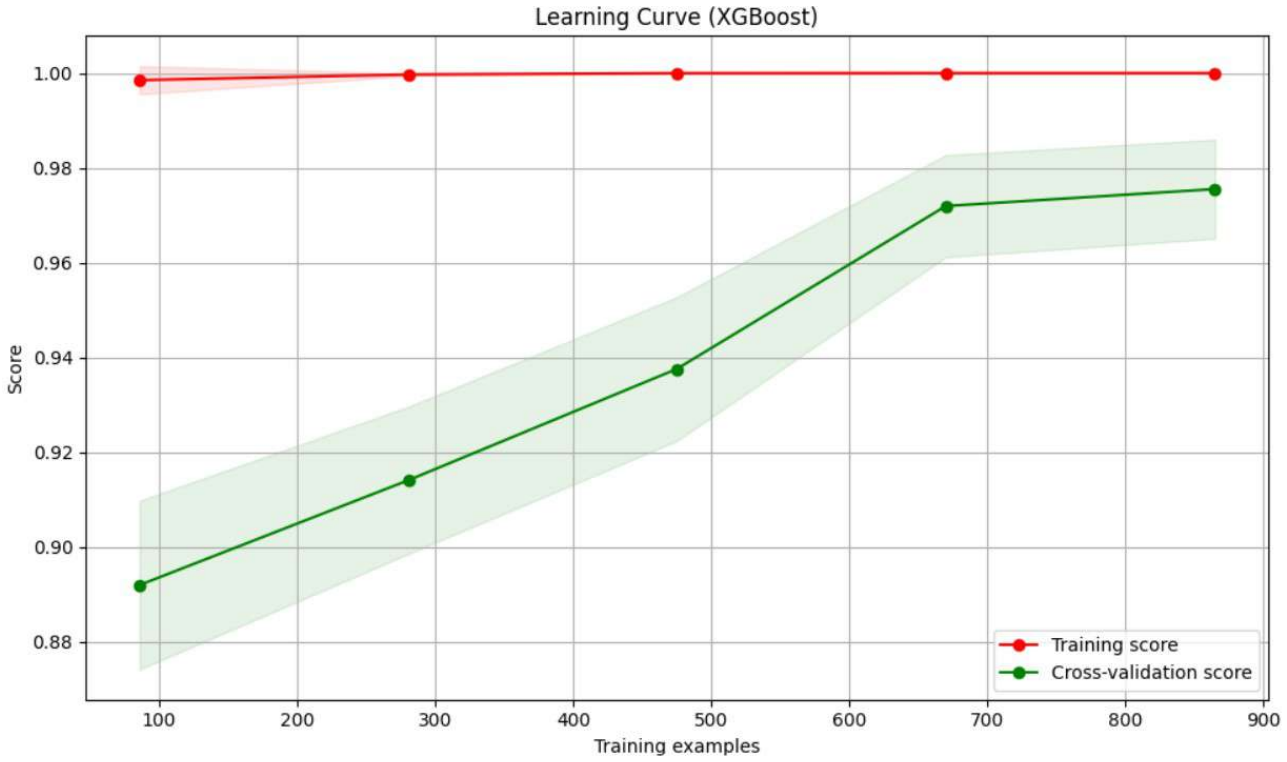


Figure 4.10: XG Boost learning curve

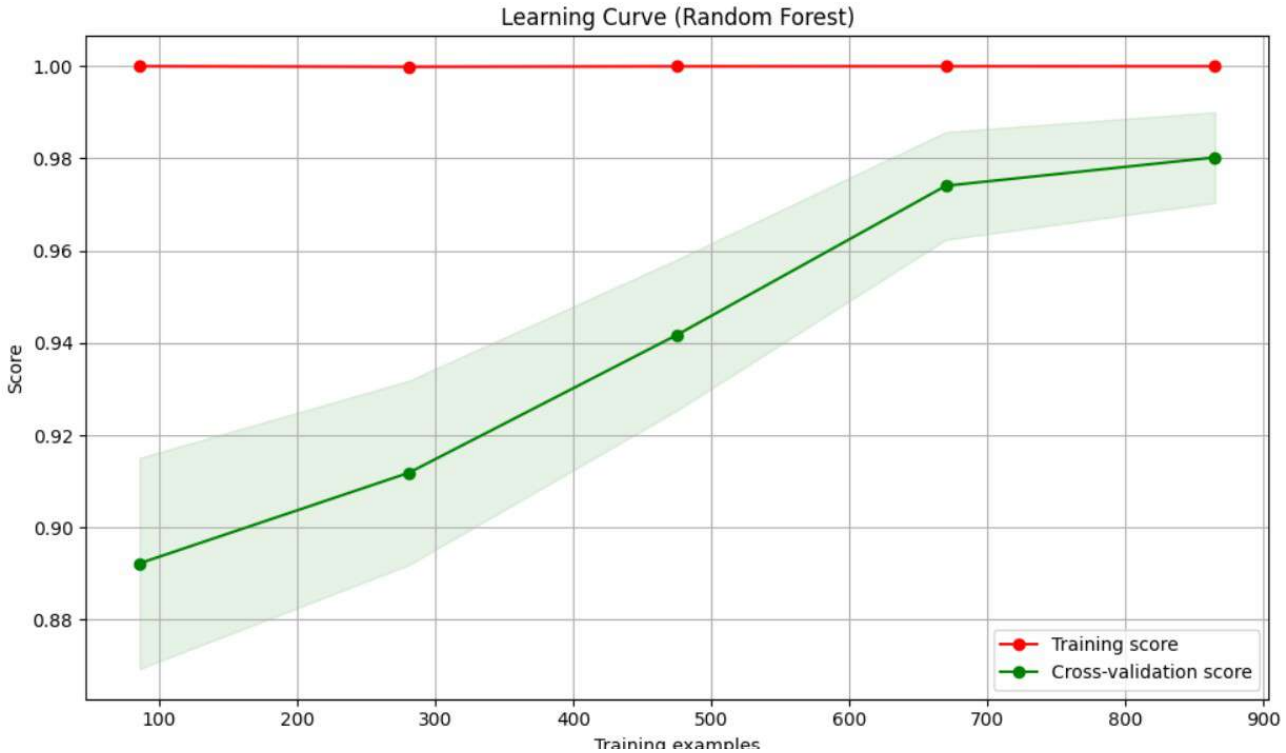


Figure 4.11: Random Forest Model learning curve

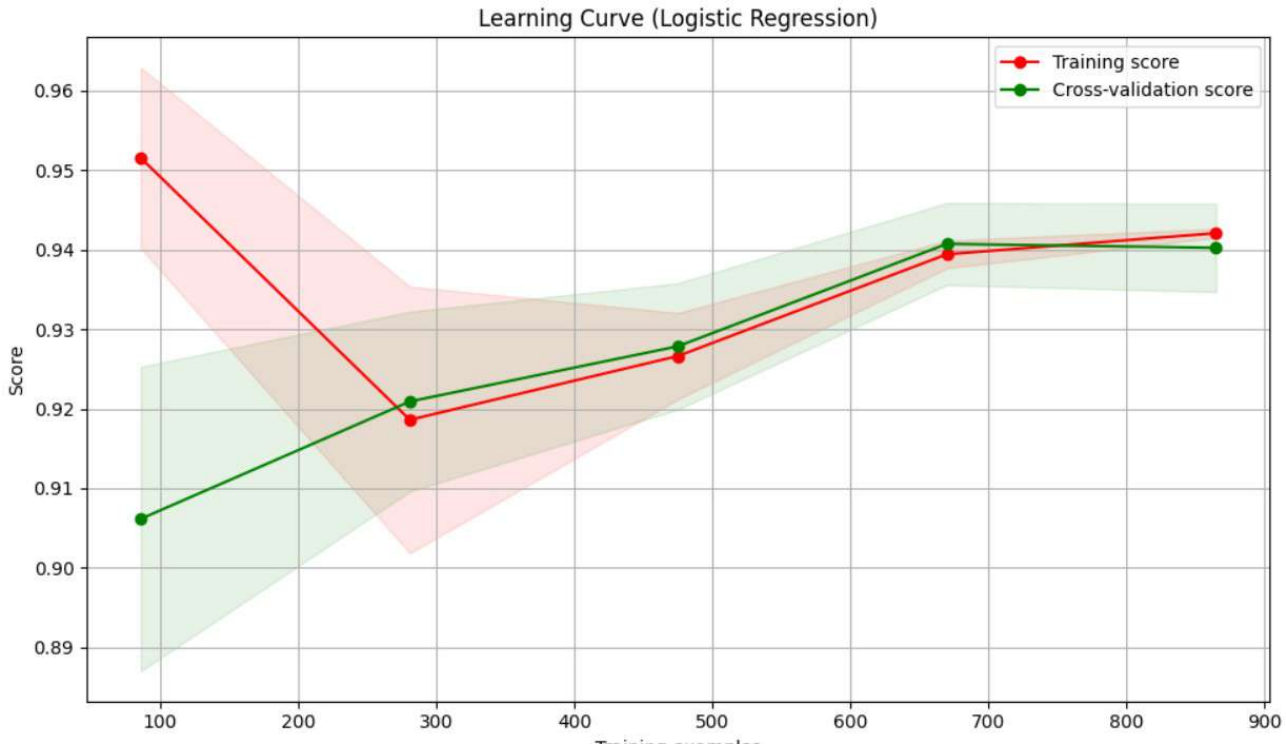


Figure 4.12: Logistic Regression learning curve

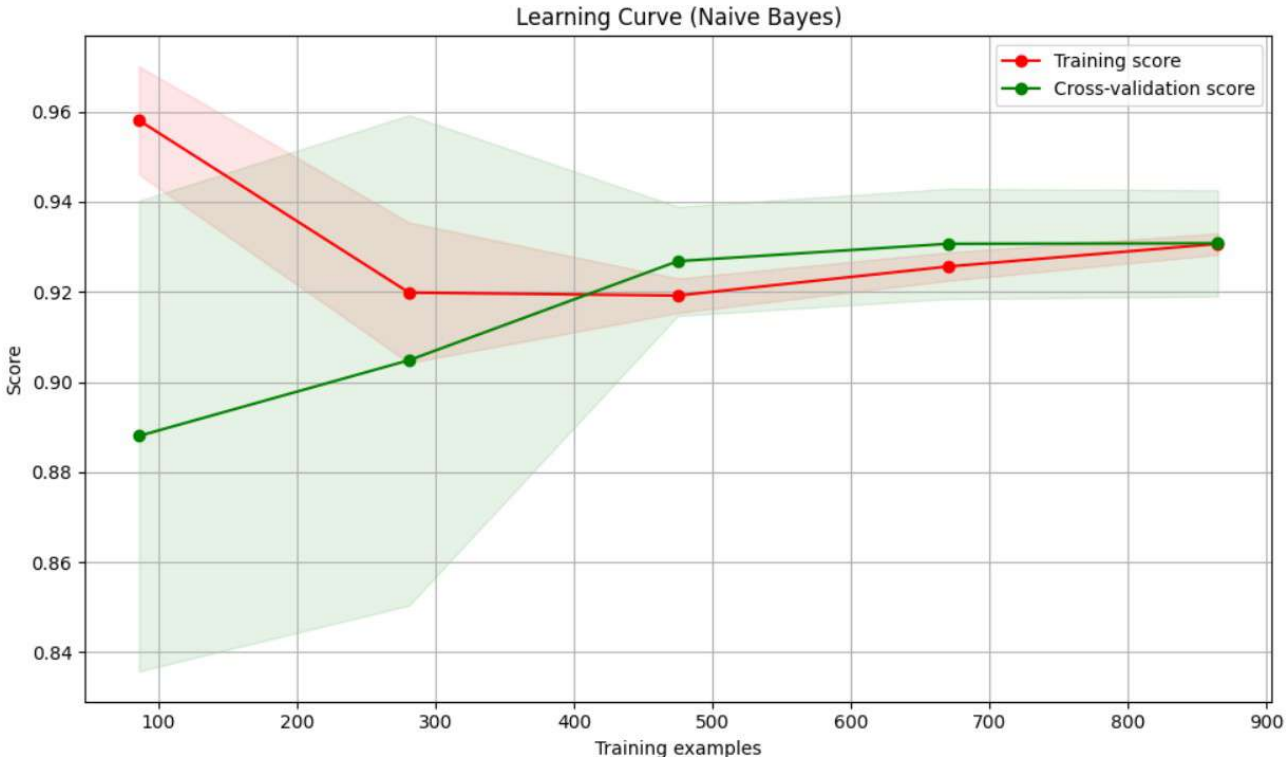


Figure 4.13: Naive Bayes learning curve

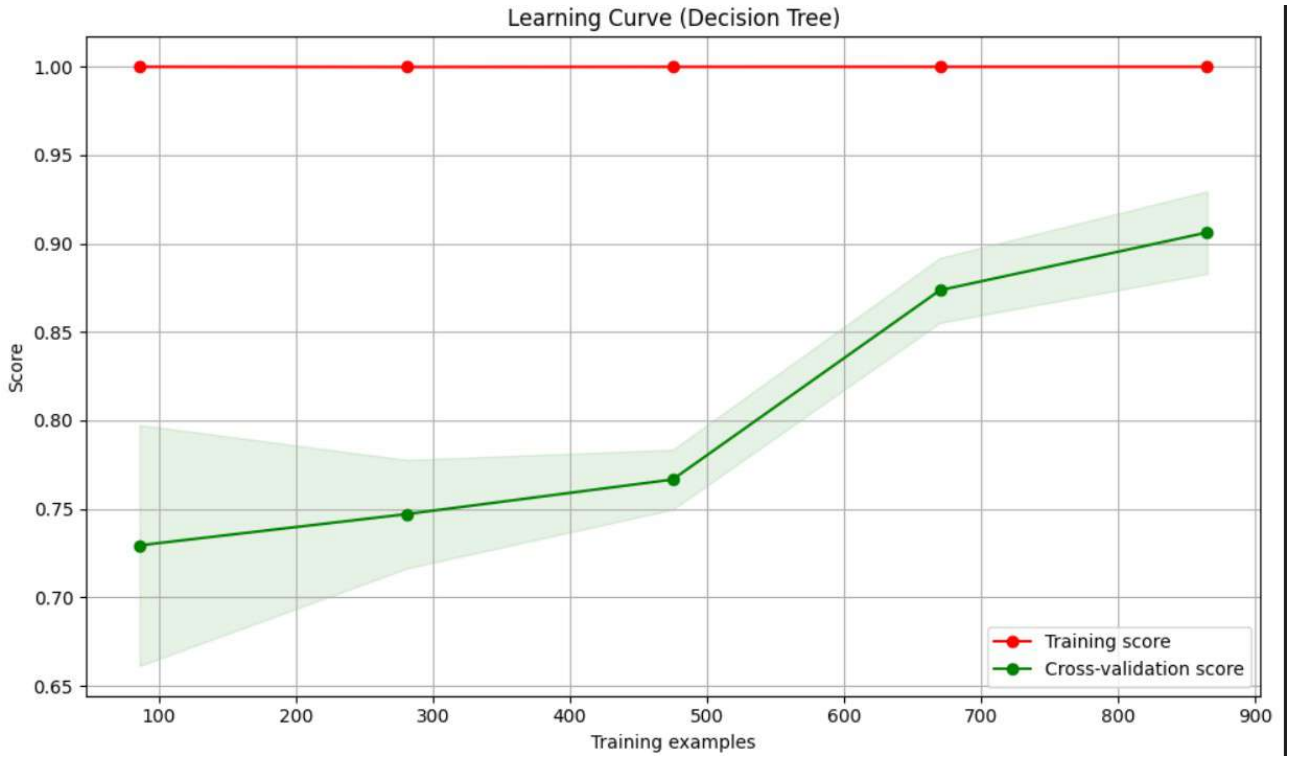


Figure 4.14: Decision Tree classifier learning curve

Predictive models for malaria mortality should be extremely accurate if they are to have any hope of being useful in clinical decision making therefore learning curves are very important in making sure that the model selected is not over fitting on the training data. The logistic regression model was a good fit since it had a low bias and low variance therefore it was complex enough to capture underlying data patterns and also it was not over fitting on the training data though it was unstable during it predictions as evidenced by the fluctuations in its performance, the random forest model performed well because of its stable increase in the cross validation curve though it needed more data to continue performing well so that the testing and training curve can meet at some point, decision tree classifier was over fitting because of the big distance seen between the training and cross validation curve therefore it was not the best model to use, naive bayes started as over fitting and unstable but when it kept on learning more data it became stable and the training and cross validation curves were able to meet at some point meaning that the model was able to generalize well on unseen data, XG boost model performance was good and the model had started when it was over fitting but later on after being trained on more data it kept on improving though it showed that it needed more data to keep capturing the data patterns and learning un seen data for better performance.

Chapter 5

Discussion

Discussions In this study, we trained machine learning models to determine which factors affect the in-hospital mortality among malaria children where the analysis of variance showed factors that affect patient outcomes. The performance of each classification model was assessed through precision, recall, F1 score and confusion matrix. Irrespective of the fact that each machine learning models incurred various strengths and weaknesses, all the models apart from the decision tree classifier produced promising results in terms of performance as well as the Receiver Operating Characteristic Area under Curve values as shown in the section of results.

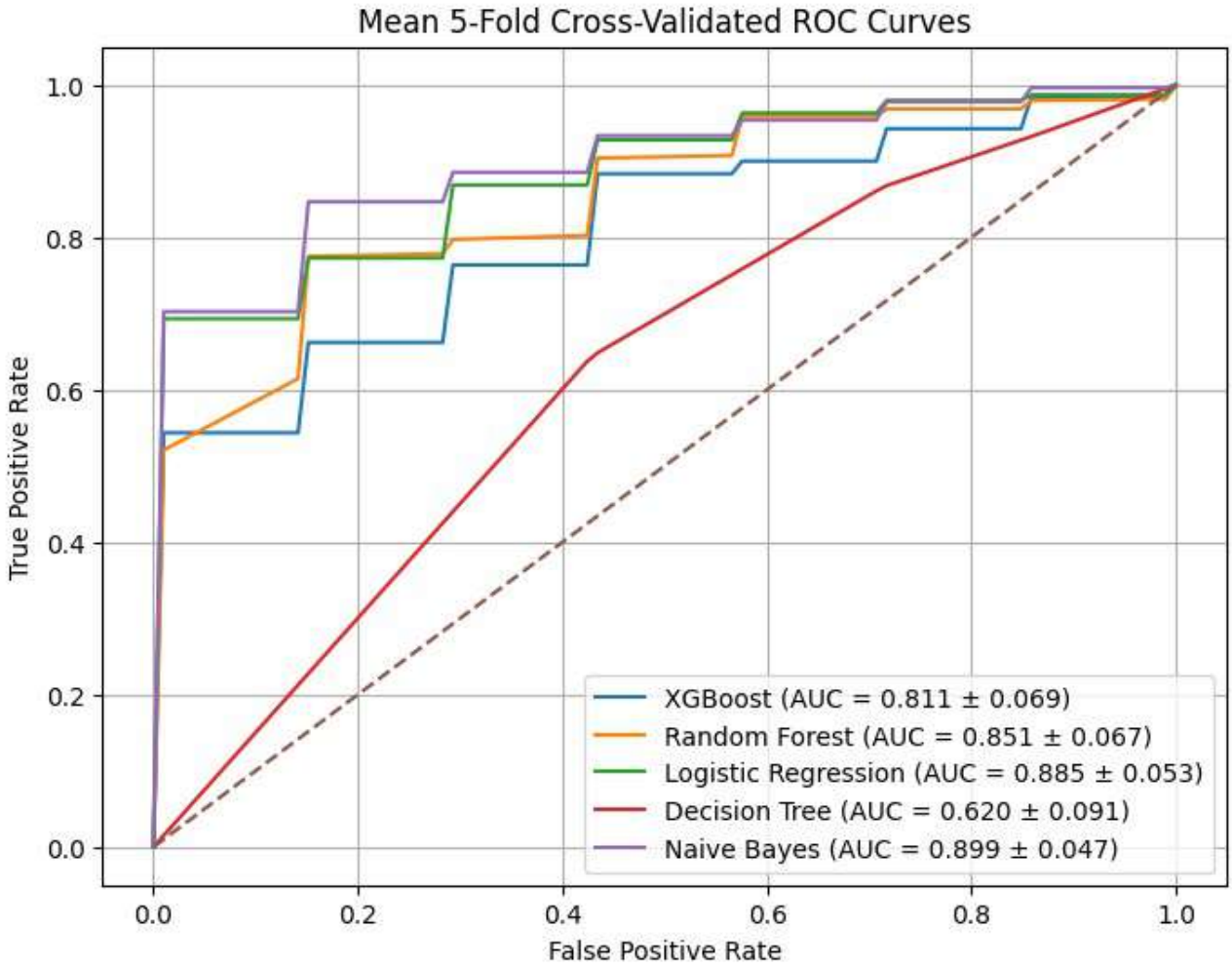


Figure 5.1: Receiver Operating Characteristic Area Under Curve

The figure 5.1 brought out clearly that the model with the highest receiver operating characteristic area under the curve was the best model to be used for machine learning which was the naïve Bayes classifier at 0.899 followed by the logistic regression model at 0.885, then the random forest model at 0.851, XG Boost model at 0.811 and lastly the decision tree classifier at 0.620. The study used the receiver operating characteristic area under the curve in determining the best model because using accuracy as the baseline indicator for the best model is misleading under imbalanced datasets. The analysis showed that the Naïve Bayes classifier was the best model since it had the highest receiver operating characteristic area under the curve 0.899 followed by the logistic regression with an area under curve of 0.885, Random Forest model with an area under curve of 0.8513, Gradient Boost model with an area under curve of 0.811 and lastly the decision tree classifier with an area under curve of 0.620.

5.1 Model choice rationale

The machine learning model evaluation conducted in this study demonstrated varying strengths across the classification algorithms. The Random Forest model achieved the highest performance in terms of precision, recall, F1-score, and overall accuracy, indicating strong capability in correctly classifying mortality outcomes within the study dataset. The confusion matrix further showed that the model produced fewer misclassifications compared to the other evaluated models. These findings suggest that the Random Forest model was highly effective in capturing complex relationships among the clinical and laboratory variables associated with mortality prediction. However, despite the superior classification metrics achieved by the Random Forest model, the Naive Bayes classifier demonstrated stronger generalization characteristics based on the learning curves and Receiver Operating Characteristic–Area Under the Curve (ROC-AUC) analysis. The Naïve Bayes classifier achieved the highest ROC-AUC score of 0.899, indicating a stronger overall ability to discriminate between mortality and survival cases across multiple classification thresholds. This is particularly important in clinical prediction problems where maintaining a balance between sensitivity and specificity is critical. The findings further revealed that the learning curve for the Naïve Bayes model showed relatively stable convergence between training and validation performance, suggesting improved consistency when exposed to unseen data. In contrast, although the Random Forest model achieved superior predictive metrics, its stronger performance on the training dataset compared to validation performance showed a tendency toward overfitting. This implied that the model may have learned dataset specific patterns that could reduce its ability to generalize effectively in different clinical environments. In this study, ROC-AUC was considered a particularly important evaluation metric because the data set was unbalanced, with survival cases exceeding mortality cases. Under such conditions, relying solely on accuracy produced misleading conclusions as the random forest model achieved high accuracy while performing poorly on minority mortality cases thus greater emphasis was placed on discriminative ability, sensitivity to mortality prediction and generalization performance. Additionally, the probabilistic nature of the Na ïve Bayes classifier made it clinically valuable for the classification of mortality risk because it estimated the probability of patient outcomes rather than relying entirely on rigid classification boundaries, thus supporting early clinical intervention and prioritization of high-risk pediatric malaria patients. Although the Random Forest model demonstrated stronger direct classification performance at a fixed threshold, the Naïve Bayes classifier showed more stable threshold independent discrimination and generalization capability. Finally, to improve interpretability and reduce black-box decision-making challenges commonly associated with machine learning

models, global feature importance analysis and probability based relationships between clinical and laboratory variables were examined to identify the most significant predictors contributing to mortality outcomes.

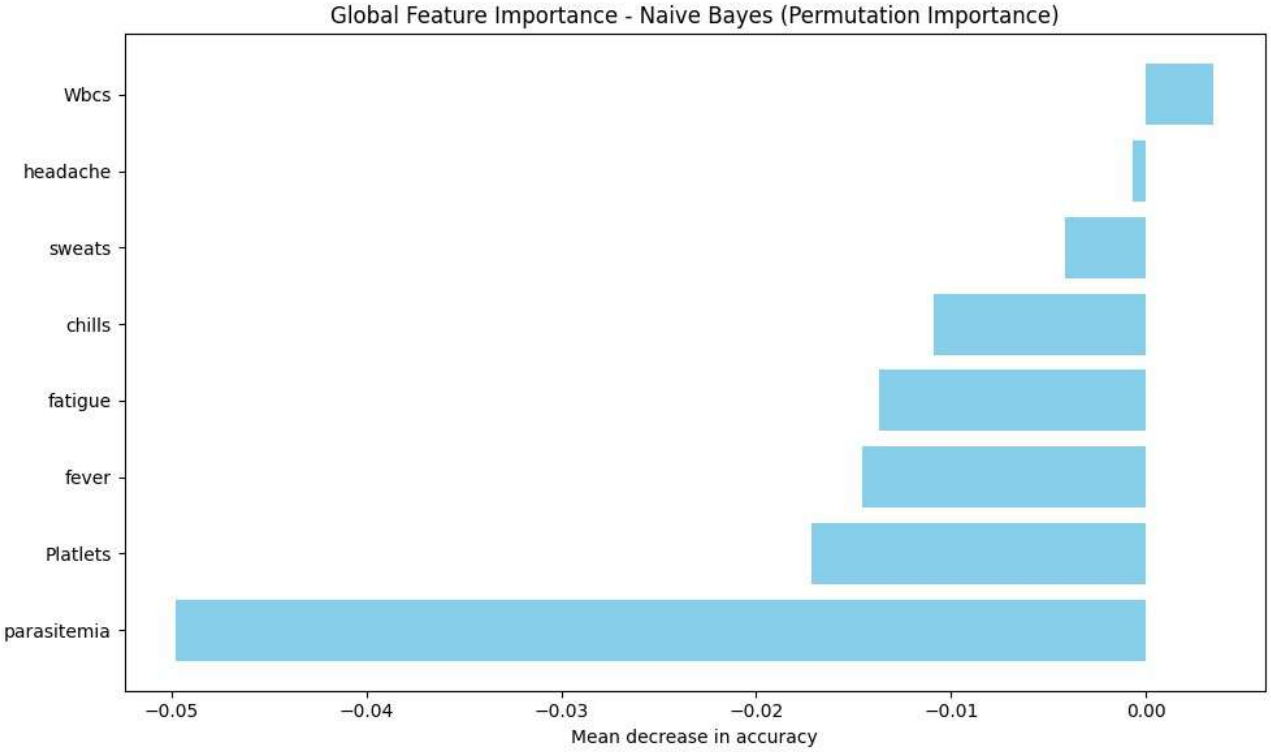


Figure 5.2: The global feature importance for feature evaluation.

The figure 5.2 for global feature importance by use of permutation importance showed that the factors that derived the model to make the predictions it made were because of malaria parasite density in the children, platelet levels, fevers, fatigue, chills, sweats and headache. The white blood cells levels didn't have any impact on the model performance.

Chapter 6

Conclusion and Recommendation

Conclusions

6.1 The exploratory analysis in identification of features associated with mortality among pediatrics

Initially the dataset was wrangled and the study found out that some variables had outliers but since this was a clinical dataset where by every single value had significant impact on the model performance, the outliers were not dropped. The data was then analyzed through visualizations and the findings showed that all the numeric variables were skewed to the right. Various box plots were used to check for relationships between the categorical variables and the numeric variables and the findings showed children that survived had more hemoglobin levels as compared to those that died meaning that malaria parasites reduce blood levels in the body because they attack the red blood cells, those that died had more white blood cells than those that survived meaning that when the body detected malaria, it produced more white blood cells to defend the body. The children with malaria parasites were seen to have higher mortality rates than the ones who did not have meaning malaria has an impact on mortality. The variables that were most significant to the outcome were parasitemia, platelet levels, headache, fatigue, fever.

6.2 Machine learning models that can predict mortality among malaria children

The findings confirmed that machine learning techniques hold significant promise in predictive health care by automating detection of mortality among malaria children with precision and reliability. Models like the decision tree classifier, random forest model, logistic regression, naïve bayes classifier, xg boost model were assessed on precision, recall, F1 score, confusion matrix, receiver operating characteristic area under curve and due to the class imbalance in the dataset the receiver operating characteristic area under curve was considered more than the accuracy measurements got from the model performance. The study found out that the Naïve Bayes classification model had the best performance and was the best to be used for machine learning since it had the highest receiver operating characteristic area under the curve, meaning that the model could easily distinguish between the 2 classes. Although some models had slightly higher precision, recall and F1

score values, the difference was not so big and more so the Naïve Bayes model had good recall values meaning that it could highly detect risky patients.

6.3 Recommendations

Health facilities and digital health platforms should consider integration of the Naïve Bayes classifier into the electronic medical records to enhance the early detection of malaria children admitted and have high mortality levels so as to enable the medical personnel distribute the necessary resources required for these children effectively. Although the performance of the Naïve Bayes model was good, further feature engineering, hyper parameter tuning and inclusion of more clinical data variables directly related to mortality like renal functional tests, acidosis, glass glow coma scales, cardiovascular features like blood pressures and pulse rates, respiratory rates and finally metabolic features like glucose levels and acidosis could boost the model's performance further more plus training the model on more diverse datasets that have inclusion for broader demographic and geographic characteristics.

6.4 Random forest and Gradient boost performance generalisability

Since the random forest and Gradient Boost model had a good performance via F1 score and there learning curves required more data for better generalization on unseen data, my study recommended that they should be trained on more diverse datasets to enhance their ability of learning broader clinical patterns to reduce on their bias towards a specific outcome. For a random forest, limiting tree depth plus controlling the number of features at each split is key to generalization enhancement and for the gradient boost using a smaller learning rate, early stopping and regularization techniques can prevent the model from over fitting thus better performance on receiver operating characteristic area under curve and learning curves.

Bibliography

[1] G. V. C. Sekhar and C. Alemu, "Comparative analysis of machine learning models for malaria detection using validated synthetic data: a cost-sensitive approach with clinical domain knowledge integration," *Scientific Reports*, vol. 15, no. 1, p. 27108, 2025.

[2] A. Olalekan Kehinde, "Leveraging machine learning for predictive models in healthcare to enhance patient outcome management," *Int Res J Mod Eng Technol Sci*, vol. 7, no. 1, p. 1465, 2025.

[3] Y. W. Lee, J. W. Choi, and E.-H. Shin, "Machine learning model for predicting malaria using clinical information," *Computers in biology and medicine*, vol. 129, p. 104151, 2021.

[4] S. M. Williamson and V. Prybutok, "Balancing privacy and progress: a review of privacy challenges, systemic oversight, and patient perceptions in ai-driven healthcare," *Applied Sciences*, vol. 14, no. 2, p. 675, 2024.

[5] T. Ciecierski-Holmes, R. Singh, M. Axt, S. Brenner, and S. Barteit, "Artificial intelligence for strengthening healthcare systems in low-and middle-income countries: a systematic scoping review," *NPJ digital medicine*, vol. 5, no. 1, p. 162, 2022.

[6] M. Badawy, N. Ramadan, and H. A. Hefny, "Healthcare predictive analytics using machine learning and deep learning techniques: a survey," *Journal of Electrical Systems and Information Technology*, vol. 10, no. 1, p. 40, 2023.

[7] G. H. Khodaee, G. Khademi, and M. Saeidi, "Under-five mortality in the world (1900-2015)," *Journal of Pediatric Perspectives*, vol. 3, no. 6.1, pp. 1093–1095, 2015.

[8] A. Dhumale, S. Kakade, and V. Patil, "Predictive analytics in precision medicine leveraging machine learning algorithms for personalized diagnosis, treatment planning, and patient outcome prediction," *Afr J Biol Sci*, vol. 6, pp. 2413–23, 2024.

[9] R. B. Nyasa, E. L. Fotabe, and R. N. Ndip, "Trends in malaria prevalence and risk factors associated with the disease in nkongho-mbeng; a typical rural setting in the equatorial rainforest of the south west region of cameroon," *Plos one*, vol. 16, no. 5, p. e0251380, 2021.

[10] U. N. C. Fund, "Levels and trends in child mortality," *Estimates developed by the UN Inter-agency Group for Child Mortality Estimation report*, 2019.

[11] J. Smith and J. Doe, "Machine learning applications in healthcare," *Journal of Medical AI*, vol. 15, no. 3, pp. 45–60, 2022.

[12] S. Fatima, "Improving healthcare outcomes through machine learning: Applications and challenges in big data analytics," *International Journal of Advanced Research in Engineering Technology & Science*, vol. 11, 2024.

[13] M. Sarker, "Revolutionizing healthcare: the role of machine learning in the health sector," *Journal of Artificial Intelligence General science (JAIGS) ISSN: 3006-4023*, vol. 2, no. 1, pp. 36–61, 2024.

[14] A. W. Demsash, "Using best performance machine learning algorithm to predict child death before celebrating their fifth birthday," *Informatics in medicine unlocked*, vol. 40, p. 101298, 2023.

[15] S. Das, L. Erdman, D. Brals, B. Boczek, S. T. Hasan, P. Massara, M. A. Alam, S. M. Fahim, M. Mahfuz, M. Hoogendoorn, *et al.*, "Development of machine learning models predicting mortality using routinely collected observational health data from 0-59 months old children admitted to an intensive care unit in bangladesh: critical role of biochemistry and haematology data," *BMJ Paediatrics Open*, vol. 8, no. 1, p. e002365, 2024.

[16] D. Uzun Ozsahin, B. B. Duwa, I. Ozsahin, and B. Uzun, "Quantitative forecasting of malaria parasite using machine learning models: Mlr, ann, anfis and random forest," *Diagnostics*, vol. 14, no. 4, p. 385, 2024.

[17] M. Khosravi, Z. Zare, S. M. Mojtabaeian, and R. Izadi, "Artificial intelligence and decision-making in healthcare: a thematic analysis of a systematic review of reviews," *Health services research and managerial epidemiology*, vol. 11, p. 23333928241234863, 2024.

[18] A. Adeyinka, Y. Lamina, O. Tawo, Y. Adeyeye, and A. Minkah, "Machine learning efforts that enhance personalized patient care and chronic disease management," *International Journal of Scientific Research in Science and Technology*, pp. 647–671, 2022.

[19] S. Khazaei, E. Ayubi, S. Nematollahi, and S. Khazaei, "Variations of infant and under-five child mortality rates around the world, the role of human development index (hdi)," *Int J Pediatr*, vol. 4, no. 5, pp. 1671–1677, 2016.

[20] A. L. Conroy, R. O. Opoka, P. Bangirana, R. Namazzi, A. E. Okullo, M. K. Georgieff, S. Cusick, R. Idro, J. M. Ssenkusu, and C. C. John, "Parenteral artemisinins are associated with reduced mortality and neurologic deficits and improved long-term behavioral outcomes in children with severe malaria," *BMC medicine*, vol. 19, no. 1, p. 168, 2021.

[21] D. Singh and B. Singh, "Investigating the impact of data normalization on classification performance," *Applied Soft Computing*, vol. 97, p. 105524, 2020.

[22] N. S. Mansur Huang, Z. Ibrahim, and N. Mat Diah, "Machine learning techniques for early heart failure prediction," *Malaysian Journal of Computing (MJoC)*, vol. 6, no. 2, pp. 872–884, 2021.

[23] W. Zhang, Y. Zhu, L. Tong, G. Wei, and H. Zhang, "Leverage machine learning to identify key measures in hospital operations management: a retrospective study to explore feasibility and performance of four common algorithms," *BMC Medical Informatics and Decision Making*, vol. 24, no. 1, p. 286, 2024.

[24] W. H. Organization, *The World Health Report 2005: Make every mother and child count*. World Health Organization, 2005.

Appendix A

Budget

[I used a budget of 440000 ugandan shillings in my research journey]

Table A.1: Project Tasks, Budget and Timeline

Task	**Amount (UGX)**	**Status**	**Months**					
			1	2	3	4	5	6
Proposal submission	50,000	Done						
Proposal approval	20,000	Done						
Data collection	150,000	Done						
Data analysis and modelling	80,000	Done						
Draft report submission	40,000	Done						
Final report submission	100,000	Pending						
Total	**440,000**							

Appendix B

Research Project timelines

The research was done in a period of six months and the different tasks, deliverables, dependencies and duration for each task are refelcted on a gant chart.

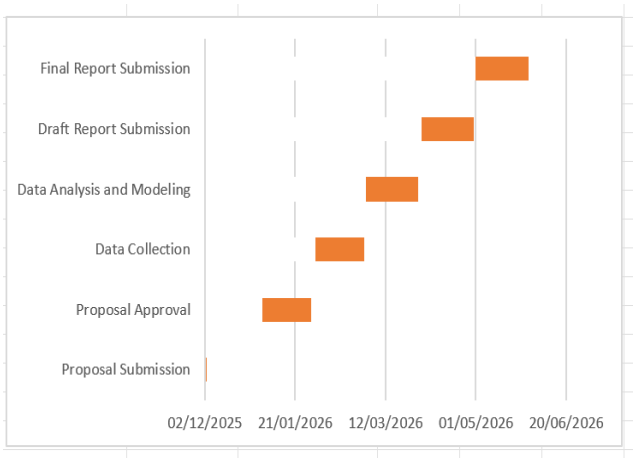


Figure B.1: Gantt Chart showing the project deliverables

Appendix C

Declaration of AI use

Declaration Statement:

I, *Nyanzi Dennis Arthur*, declare that I used

i. ***CoPilot*** to ***To troubleshoot code*** and ***evaluated the error cause***. I modified the outputs in the anova code ***by correcting the wrong variable name.***.

Signature: _____ Date: 28-04-2026

Appendix D

List of 10 Major Journal Publications reviewed

(a) O. Kehinde, Leveraging machine learning for predictive models in health care to enhance patient outcome management, International Research Journal of modernisation in engineering technogy and science, 2025.

(b) J. W. C. You Won Lee, "Machine learning model for predicting malaria using clinical information," in Computers in biology and medicine, 2021.

(c) N. R. Mohammed Badawy, Health care predictive analytics using machine learning and deep learning teachniques, Journal of electrical systems and inf technol, 2023

(d) S. P. Dhumale Kakade, Predictive analytics in precision medicine leveraging machine learning algorithms for personalise diagnosis, treatment planning and patient outcome prediction, 2024.

(e) F. E. Nyasa RB, . Trends in malaria prevalence and risk factors associated with the disease in Nkongho-mbeng; a typical rural setting in the equatorial rainforest of the South West Region of, International Journal of information technology and computer engineering, 2021.

(f) J. DOe, Machine learning applications in health care. A review of IEEE transactions on biomedical engineering, 2022 Vol 69, no 2 pp 300-312.

(g) S. Fatima, Improving health care outcomes through machine learning, International Journal of Research in Engineering Technology, 2024.

(h) A. W. Demsash, "Using best performance machine learning models to predict child death before celebrating their fifith birthday," in Informatics in medicine unlocked, Mettu university college of health sciences department of health informatics, Ethiopia, 2023.

(i) S. D. D. B. P. M. Lauren Erdman, "Development of machine learning models predicting mortality using routinely collected observational health data from 0-59 months old children admitted to an ntensive care unit in Bangladesh," in BMJ Paediatrics Open, Bangladesh, 2024.

(j) V. W. L. O. O. A. Dennis Kariuki Muriithi, "Explainable Artificial Intelligence Models for predicting Malaria Risk in Kenya," in European Journal of Artificial Intelligence and Machine Learning, 2025.

Appendix E

Research alignment with national (NDPIV) and Global (SDGs) agenda

E.1 Alignment with the National Development Plan IV (NDPIV)

E.1.1 Alignment with NDPIV

The research study of predictive analytics of malaria in hospital mortality aligns with human capital development programme pillar for the NDPIV through strengthening and maintaining supportive infrastructure and facilities for health, increasing adoption of preventive health measures plus reduction in the cost of provision and access to health care services to ensure decent and productive work environment for everyone in the community.

E.1.2 Alignment with Sustainable Development Goals (SDG)

The research study aligns with the sustainable development goal of promoting good health and well being for all through fighting communicable diseases by 2030 which include malaria, AIDS, tuberculosis to ensure that everyone has a healthy life.

E.1.3 Alignment with an African Union agenda

The research study aligns with the African Union goal 3 of promoting healthy and well nourished citizens.

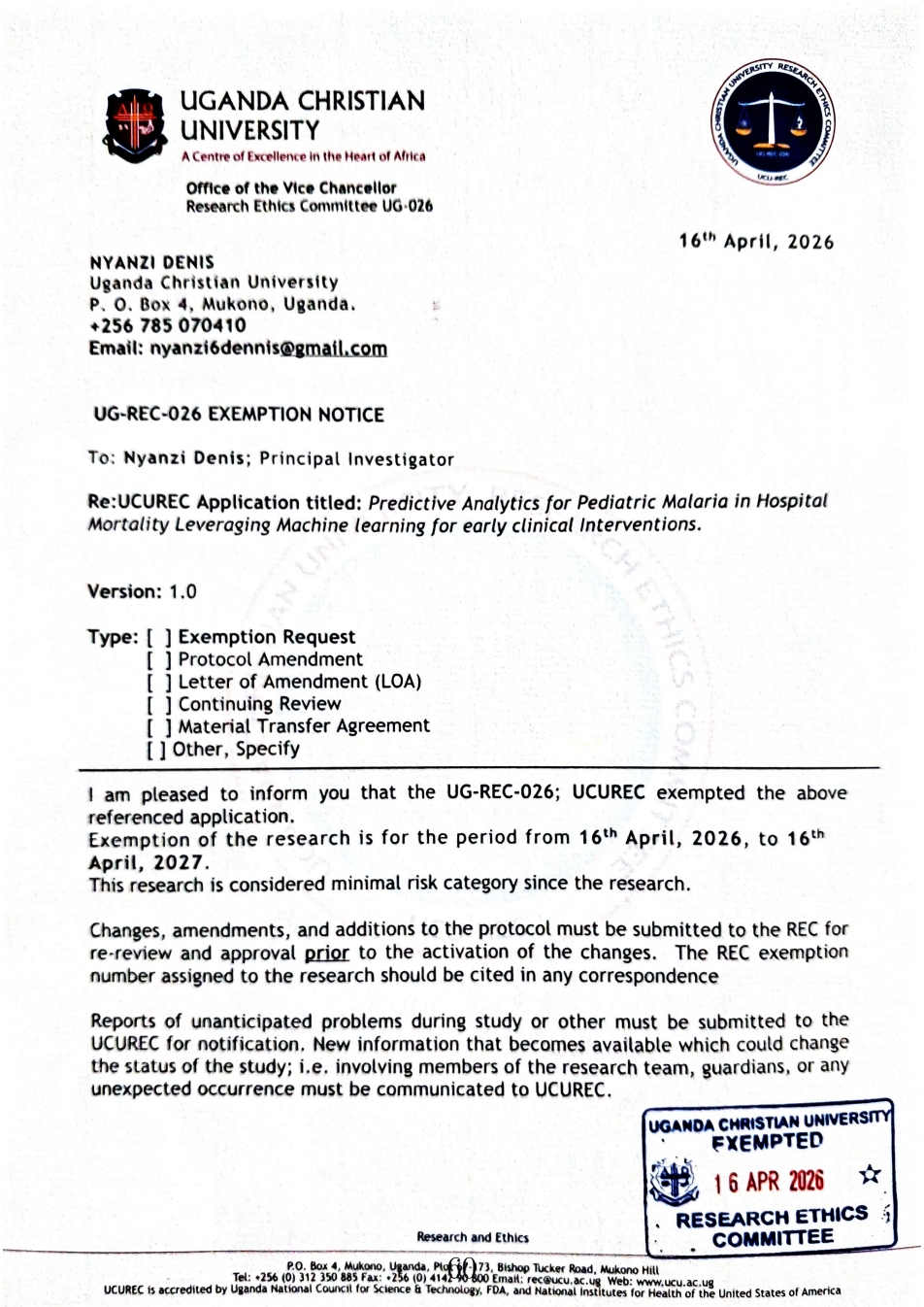
Appendix F

Algorithms, Code of interest

https://github.com/NyanziArthur/Malaria-Mortality-Thesis-code/blob/main/Github$_{t}hesis_{c}ode.IP$

Appendix G

Research Ethics Committee (REC) Approval



UGANDA CHRISTIAN UNIVERSITY
A Centre of Excellence in the Heart of Africa

Office of the Vice Chancellor
Research Ethics Committee UG-026

16th April, 2026

NYANZI DENIS
Uganda Christian University
P. O. Box 4, Mukono, Uganda.
+256 785 070410
Email: nyanzi6dennis@gmail.com

UG-REC-026 EXEMPTION NOTICE

To: Nyanzi Denis; Principal Investigator

Re:UCUREC Application titled: *Predictive Analytics for Pediatric Malaria in Hospital Mortality Leveraging Machine learning for early clinical Interventions.*

Version: 1.0

Type: [ ] Exemption Request
[ ] Protocol Amendment
[ ] Letter of Amendment (LOA)
[ ] Continuing Review
[ ] Material Transfer Agreement
[ ] Other, Specify

I am pleased to inform you that the UG-REC-026; UCUREC exempted the above referenced application.
Exemption of the research is for the period from 16th April, 2026, to 16th April, 2027.
This research is considered minimal risk category since the research.

Changes, amendments, and additions to the protocol must be submitted to the REC for re-review and approval prior to the activation of the changes. The REC exemption number assigned to the research should be cited in any correspondence

Reports of unanticipated problems during study or other must be submitted to the UCUREC for notification. New information that becomes available which could change the status of the study; i.e. involving members of the research team, guardians, or any unexpected occurrence must be communicated to UCUREC.

UGANDA CHRISTIAN UNIVERSITY
EXEMPTED
16 APR 2026
RESEARCH ETHICS COMMITTEE

Research and Ethics

P.O. Box 4, Mukono, Uganda, Plot 173, Bishop Tucker Road, Mukono Hill
Tel: +256 (0) 312 350 885 Fax: +256 (0) 4142-90-600 Email: rec@ucu.ac.ug Web: www.ucu.ac.ug
UCUREC is accredited by Uganda National Council for Science & Technology, FDA, and National Institutes for Health of the United States of America

Figure G.1: Exemption Letter

Appendix H

Student - supervisor meeting/supervision plan

Table H.1: Research plan

Task / Activity	Description	Expected Deliverable	Timeline
1	Protocol review	Review of proposal	Month 1
2	Data collection	Clean dataset ready	Month 2
3	Data Analysis	Data analysis done	Month 3
4	Draft report	Draft report ready	Month 4
5	Final report	Final report ready	Month 5
6	Research poster	Research poster ready	Month 6

Supervisor: _____ **Date:** 28-04-2026

Student: _____ **Date:** 28-04-2026

Appendix I

Research poster layout

[25pt, a0paper, landscape]tikzposter

graphicx amsmath

Default Default