Application of Machine Learning to Identify Risk Factors for Anaemia in Pregnancy

Moses Wabwile Mukhanya

169118

A dissertation Submitted in Partial Fulfilment of the Requirements for the Degree of Master of Science in Data Science and Analytics at Strathmore University

Strathmore Institute of Mathematical Sciences

Strathmore University

Nairobi, Kenya

June 2025

Declaration and Approvals

Declaration

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the dissertation contains no material previously published or written by another person except where due reference is made in the dissertation itself.

Name: **Moses Wabwile Mukhanya**

Signature: 

Date: May 29, 2025

Approval

The dissertation of Moses Mukhanya was reviewed and approved by the following:

Dr. Betsy Muriithi,
Lecturer iLAB Africa
Strathmore University.

Dr. Godfrey Achono,
Strathmore Institute of Mathematical Sciences,
Strathmore University.

Prof. Bernard Shwibwabo Kasamani,
Director of Graduate Studies,
Strathmore University.

Abstract

WHO defines a pregnant woman to be anaemic when the Haemoglobin Concentration value is less than 11g/dl. Reports published by WHO in 2019, 37 % (32 million) of pregnant women globally were Anaemic. These reports also indicate that pregnant women from low and middle-income countries carry the highest burden of prevalence of anaemia.

Anaemia negatively affects the mother and both the unborn and born babies. There are research studies that show maternal mortality at delivery, preterm birth, stillbirth, low birth weight, neonatal death, poor development milestones for the babies and postpartum haemorrhage are caused by Anaemia.

This research project will utilise data collected from Mariakani and Rabai Hospitals to identify the risk factors of Anaemia and also build a predictive model for early screening of Anaemia. Results from this research could be used to improve the existing maternal health policies and contribute towards attaining sustainable development goals.

KEY WORDS: Anaemia, risk factor, Haemoglobin.

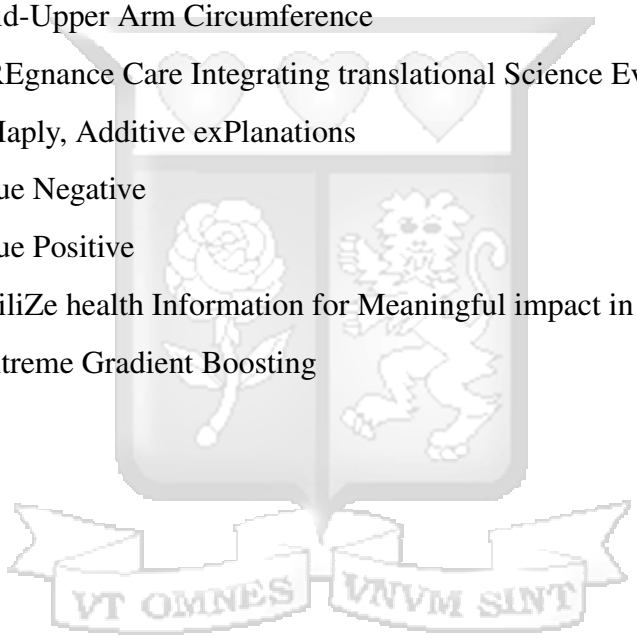
Table of Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Meaning
AUC-ROC	Area Under the Receiver Operating Characteristic Curve
BMI	Body Mass Index
FN	False Negative
FP	False Positive
MUAC	Mid-Upper Arm Circumference
PRECISE	PREgnance Care Integrating translational Science Everywhere
SHAP	SHaply, Additive exPlanations
TN	True Negative
TP	True Positive
UZIMA	UtiliZe health Information for Meaningful impact in East Africa through Data Science
XGBoost	Extreme Gradient Boosting



Acknowledgement

First and foremost, I am deeply grateful to the Almighty God for His provision, grace, and the gift of good health throughout this journey.

I am deeply grateful to my wife and daughter for their unwavering support. Their encouragement and motivation kept me going, especially during moments when I felt overwhelmed with analysis and writing. I also extend my heartfelt appreciation to my parents.

My deepest gratitude goes to my supervisor, Dr. Betsy Muriithi, for their guidance, availability, and invaluable support. Her mentorship has been instrumental in the successful completion of this dissertation.

I am sincerely indebted to Dr. Angela Koech, whose encouragement inspired me to pursue further academic growth when I had not considered it. Her expertise has been invaluable, providing critical medical and domain-specific guidance throughout the analysis process. A heartfelt thank you to Onesmus and Marie-Volvert, whose unwavering support and commitment in forming a dedicated support group played a significant role in ensuring success.

I sincerely appreciate my employer for the support and flexibility given to me to work on my master's dissertation. Your understanding and encouragement have been invaluable in helping me balance my academic and professional responsibilities, and I am truly grateful.

My sincere appreciation to UZIMA/NIH for funding my master's project. Your support has been instrumental in advancing my research, and I am truly grateful.

Thank you so much to PRECISE study for providing the data used in this study. Your contribution has been invaluable in advancing our understanding of the subject matter, and I am truly grateful for your support.

Dedication

This dissertation is dedicated to God Almighty for giving me wisdom and good health.

Chapter 1

Introduction

1.1 Background

According to WHO anaemia in Pregnancy is defined as a Haemoglobin concentration less than 11g/dl WHO (2011). Data reported by WHO in 2019 showed that 37% (32 million) of pregnant women globally had their Haemoglobin levels less than 11g/dl WHO (2023a). These reports indicated that anaemia is a public health issue. Preventing and controlling anaemia through primary health care policies have been in existence since 1987 WHO (2011). But anaemia is yet to be addressed to date. Pregnant women from low and middle-income countries are the most affected by anaemia globally Y. Balarajan (2011).

Mapping of anaemia among pregnant women in Kenya (2016 – 2019) indicated coastal counties of Kilifi, Kwale, Mombasa, Lamu and Taita Taveta had a high prevalence of elevated Anaemia Odhiambo (2020). Even with this high prevalence of anaemia among pregnant women in these regions, little has been done in terms of identifying the leading causes anaemia so that interventions are developed based on evidence to control Anaemia.

There is enough evidence indicating the negative impacts both on the mother and the child for the mothers who have been anaemic while pregnant. Anaemia is one of the main causes of maternal death in pregnancy Daru (2018). It is also the main risk factor for postpartum haemorrhage after delivery Mehrnoush (2023). Anaemia has also been identified as the main cause of preterm birth, stillbirth, neonatal death, low birth weight and poor development among children Levy (2023) Rahman (2016).

Sustainable Development Goals require that every country should put into action strategies that improve the health of the citizens UN (2015). This research study will help towards the attainment of SDGs by studying and analysing anaemia in pregnancy.

Preliminary findings from the PRECISE research study indicated that more than half of the enrolled participants were Anaemic. The PRECISE stands for PREgnancy Care Integrating translational Science Everywhere and it is a network of research scientists and health advocates based in Europe, Africa, America and other continents. This network was established to investigate the cause and risk factors of pregnancy complications like Hypertension, fetal growth restriction, stillbirth and preterm. PRECISE research study was conducted in Gambia, Mozambique and Kenya and the research study we are proposing will use data from Kenya von Dadelszen (2020).

1.2 Statement of the Problem

Antenatal data collected by PRECISE research study from women visiting Mariakani and Rabai hospitals indicate that more than half of the women have anaemia. Also, reports and data from anaemia mappings of coastal regions show a high prevalence of severe Anaemia in the coastal counties.

Conventional statistical methods have been used to identify the potential risk factors of Anaemia but have not developed an intelligent predictive model for early detection of potential Anaemia cases among pregnant women.

Rules and logic have been derived from conventional statistical methods that have been prebuilt into electronic healthcare systems that can be used to predict anaemia among pregnant women. As far as these electronic systems based on conventional statistical approaches are good, they are not just static but also not efficient and effective in detecting the change in data patterns.

1.3 Objectives

1.3.1 General Objective

To use machine learning algorithms to identify predisposing factors of anaemia among pregnant women attending Antenatal Services at Mariakani and Rabai Hospitals.

1.3.2 Specific Objectives

1. Identify the prevalence of anaemia among pregnant women attending antenatal services at Mariakani and Rabai Sub Counties.
2. Identify risk factors of anaemia among pregnant women.
3. Develop a machine learning screening tool.

1.4 Scope and Limitations

This research project used demographic, socio-economic, environmental, nutritional and wash data as predictors while anaemic status data based on haemoglobin concentration was used as the dependent variable.

This project used secondary data from the PRECISE research study. This study used cross-sectional data with no case and control comparison groups. Risk factors were identified based on the patterns identified from the data and further work about establishing a longitudinal cohort of both case and control groups might be needed.

Also, the results of this research project might not be generalizable to other settings outside the coastal region because of regional differences. Studies from different regions tend to show different risk factors for anaemia among pregnant women.

1.5 Research Relevance

The results from this research project could be used to prospectively predict women at risk of developing anaemia in pregnancy. This could help in the early detection of potential anaemia patients so that medical interventions are initiated as early as possible.

Results from this analysis could also inform policymakers and decision-makers to implement community-level strategies and policies to prevent and control anaemia. These policies could not only help to cut down the expenditure on healthcare, especially in the management of anaemia among pregnant women but also inform the data-informed allocation of the limited resources in Kilifi County.

Chapter 2

Literature Review

2.1 Introduction

A lot of research studies and analyses carried out in the area of anaemia among pregnant women have employed conventional statistical approaches. There also exists some research analyses that have used machine learning approaches but not in the Kenyan context. The literature review of this research will address the gaps in existing studies done before and also demonstrate why machine learning methods outshine conventional statistical methods.

2.2 Anaemia Risk Factors

A secondary data analysis was conducted in China to understand the prevalence and risk factors of anaemia among Chinese pregnant women Lin (2018). The overall prevalence of anaemia was identified to be 23.5% among the enrolled pregnant women. Maternal age greater than or equals to 35 years, family per capita monthly income less than 1000 CNY, rural residence and pregnancy BMI of less than 18.5kg/m2 were identified to be significantly associated with anaemia.

A total of 424 pregnant women were enrolled in a cross-sectional study Azhar (2021) conducted in public and private hospitals in Bangladesh. Binary logistic regression analysis was used in the analysis and identification of the risk factors of anaemia among pregnant women. This study showed that maternal age 20–25 and 26–30 years, monthly family income of less than US 300, attending ANC services in government hospitals, having given birth two

or more times, being pregnant two or more times, not using contraception, and not using iron supplements were the risk factors statistically associated with anaemia.

Another study to show the prevalence and identify the risk factors of anaemia was conducted in Bangladesh with a sample size of 384 pregnant women Ahmed (2019). The analysis results of this study showed a prevalence of 58.9%. Also, the study identified monthly family income, family size, gestational age, level of birth spacing, excessive blood loss in the previous surgery, food group eaten 24 hours and breakfast regularly were the main risk factors of anaemia among pregnant women. Some of the risk factors of anaemia in pregnancy enlisted by this study were consistent with another project carried out in Bangladesh [20].

Data collected from a tertiary referral hospital in Northern Ghana showed that 50.8% of the enrolled pregnant women were anaemic Wemakor (2018). Bivariate and multivariate analyses were done and noticed that pregnant women's knowledge about anaemia and pregnancy trimester at the interview were the only risk factors for anaemia.

A study was conducted in Egypt at the Minia University Hospital Mostafa (2022) and adjusted logistic regression analyses showed that rural residence, low education level, low family income, multi-para (given birth to two or more offsprings), low pregnancy interval, insufficient meals per day, insufficient meat intake, insufficient vegetable intake, insufficient egg intake, insufficient milk intake and parasitic infestation were the most statistically significant risk factors of anaemia among pregnant women in Egypt. Whether a pregnant woman was working or not working was found to be not statistically significant with anaemia in pregnancy.

Berhe (2022) looked at the risk factors of Anaemia among pregnant women attending antenatal services in health facilities in some parts of Ethiopia. Bivariate and multivariate logistic regression analyses were done. The findings of this research study reported that intestinal parasites, farmer occupation, unprotected sources of drinking water, drinking coffee/tea with or immediately after meal daily and a diet diversity score (a score that is used to assess the variety of foods or food groups consumed by an individual or household over a certain period) of less than 3 were found to be statistically significant for anaemia among pregnant women.

The prevalence of anaemia and its risk factors were also investigated in a study conducted in Gondar, Northwest Ethiopia Melku (2014). This was a cross-sectional study that enrolled pregnant women attending their antenatal services at Gondar University Teaching Hospital. The results from this study indicated a prevalence of 16.6% from the sampled participants. Additionally, the bivariate analysis showed that being illiterate, low monthly family income, large family size, being underweight, being pregnant two or more times, hookworm infection, and being HIV positive were significantly associated with anaemia in pregnancy. Also, multivariate logistic regression was carried out and showed that low monthly family income , large family size, hookworm infection, and being HIV positive remained independent predictors of pregnancy anaemia.

A study conducted in Northern Tanzania to identify the risk factors of anaemia showed very different results Stephen (2018). In this study, the clinic where the pregnant women were attending their ANC services and low education level were identified as the only factors associated with anaemia. Conversely, food security and household characteristics like the number of meals per day, intake of meat, history of food security and water source sanitation were not associated with anaemia in pregnancy. Also, the study indicated that anaemia is not a risk factor for poor pregnancy outcomes like low birth weight, preterm birth and stillbirth.

A study conducted in the southern part of Tanzania Kara (2020) showed that being illiterate or having a primary level education, either public or privately employed, and average monthly family income less than TZS SH.100,000 were highly associated with anaemia. This cross-sectional study indicated an overall prevalence of anaemia at 46.3%. Lastly, the risk factors identified in this research study revealed that the high prevalence of anaemia was caused by poverty and lack of knowledge on the prevention of anaemia. Chi-square tests and logistic regression were used to identify the risk factors of anaemia in pregnancy. Using data from a single ANC clinic and not considering haematological, hereditary and dietary factors were identified as the limitations of this research study.

A systematic review was done to understand the prevalence and association between anaemia and pregnancy outcomes in low- and middle-income countries Rahman (2016). The results indicated that 42.7% of women experienced anaemia during pregnancy in low- and middle-

income countries. Also, anaemia was identified as a risk factor for low birth weight, preterm birth, perinatal mortality, and neonatal mortality in pregnant women with anaemia. The results further indicated that 12% of low birth weight, 19% of preterm births and 18% of perinatal mortality were caused by anaemia. This systematic review finally concluded by stating that anaemia in pregnancy in low- and middle-income countries remains a big burden that needs to be addressed.

Otieno (2024) looked at the prevalence and risk factors of Anaemia among pregnant women attending their Antenatal Clinics at Kilifi County Referral Hospital. This cross-sectional research study enrolled 191 pregnant women using a structured questionnaire and extracted haemoglobin concentration data from antenatal clinic books. Bivariate and multivariate logistic regression was used to identify the association between independent variables and anaemia. This study found that the prevalence of anaemia among the enrolled sample was 54%. This study finally found that increased household size, history of malaise (the general feeling of unease and physical and emotional discomfort without a specific cause) and fever were positively associated with anaemia while increased intake of vegetables, fruits and red meat was associated with reduced likelihood of developing anaemia in pregnancy.

Okube (2022) analysed data from pregnant women attending antenatal clinics in their pregnancy's second and third trimesters at Pumwani Maternity Hospital, Kenya. This study employed a cross-sectional design and enrolled 258 participants. The findings of this analysis identified that mid-upper arm circumference of less than 23 cm, advanced maternal age (>31 years), not taking iron/folic acid supplementation and government/private employed and self-employed women as predictors of Anaemia. One interesting finding of this study indicated that employed pregnant mothers (government/private/self-employed) were at a higher risk of developing when compared to housewives because they do not have enough resting time and also enough time to attend ANC clinics compared to housewives. This analysis used Pearson's chi-square test and odds ratio (OR) with corresponding 95% confidence intervals (CI) to find the association between independent and dependent variables.

Another research study to identify the risk factors of severe anaemia among pregnant women was conducted in Kisumu District, Kenya Ondimu (2000). Multivariate logistic regression

was used to analyse the data collected from 1455 participants. The results from this analysis showed that low education levels, lack of formal employment and economic autonomy, poor quality of housing, lack of clean piped water within the household, pregnancy at an early age and rural residence where the mode of transport is commonly non-motorised were the main risk factors of anaemia. The author continues to state that these factors lead to poor and delayed health care during pregnancy which predisposes pregnant mothers to severe conditions of anaemia. The research recommended that the government develop interventions geared towards addressing these risk factors in Kisumu District.

2.3 Modelling the Risk of Anaemia in Pregnant Women

Machine learning methods have been used to predict the risk of developing cardiovascular diseases and the findings compared to the established models developed based on conventional statistical methods Weng (2020). Random Forest, logistic regression, gradient boosting and neural networks performed better than the statistical approaches. AUC metric was used to compare the performance of both approaches. On this metric, the currently established algorithm achieved 72.6% while Random Forest, logistic regression, gradient boosting and neural networks achieved 74.5%, 76.0%, 76.66 and 76.90% respectively. The prediction of patients at risk of developing cardiovascular problems increased when machine learning approaches were used compared to the standard risk prediction algorithm based on the statistical approaches.

Another study demonstrating the power and robustness of machine-learning algorithms was conducted by Rajula (2020). This study compared the performance of machine learning against conventional statistical approaches. Machine learning approaches performed better than statistical methods in predicting clinical deterioration in the ward, mortality in acute coronary syndrome, mortality in ICU and readmission in patients hospitalised for heart failure.

Deep Neural networks and machine learning approaches performed better in accurately identifying diabetic patients Tripathy (2024). Deep Neural networks achieved the highest

accuracy score of 93.0% while LR, KNN, CART, RF, SVM, XGB and LightGBM achieved accuracy of 84.0%, 84.0%, 85.0%, 88.0%, 85.0%, 89.0% and 88.0% respectively. This study showed that machine learning algorithms and deep learning neural networks can be applied to predict anaemia and other diseases in the medical field correctly.

Machine learning algorithms remain unexplored in low-middle-income countries yet have proved to be more reliably robust than statistical techniques in the medical field. Data Science, Machine learning and artificial intelligence approaches are the recent technologies in low-middle-income countries yet most of the developed countries have been using these robust techniques to drive their economies and improve the health sector. A study conducted in Ethiopia [14] used Random Forest, extreme gradient boosting, Cat Boost and Decision Trees machine learning algorithms to predict the level of Anaemia Among Ethiopian Pregnant Women. This study used data from the Ethiopia Demographic Health Survey and this not only included all the regions in Ethiopia but also provided enough sample size (over 115,000 records) that is required by machine learning algorithms. Decision Tree had a score of less than 80% on Accuracy, Precision, Recall, F1_Score and Cross-validation metrics. Random Forest, Extreme Boost and Cat Boost algorithms performed well with a score of above 90% on the evaluation metrics. Strong predictors of Anaemia were identified and then rules based on the predictors were prebuilt into the logic of predicting Anaemia. Feature importance was used to identify current pregnancy, age, source of drinking water, respondent's (pregnant women) occupation, number of household members, wealth index, husband/partner's education level, and birth history as the risk factors of anaemia in pregnancy.

2.4 Conclusion

Prevalence and risk factors of anaemia among pregnant women have not been extensively studied in coastal regions and Kenya. The studies so far conducted in Kenya focused on the urban population Otieno (2024) Dejene (2022a) . The study conducted at Kilifi County Referral Hospital Otieno (2024) discriminated against the rural population because this facility is located in an urban setting. The majority of the pregnant women attending the

antenatal facilities in this facility come from the urban setting. Also, a study conducted in Nairobi at Pumwani Maternity Hospital enrolled pregnant women from an urban setting. The research project we are proposing will use data from both rural and urban populations and the results will be generalizable to both the urban and rural setting.

Research projects from other countries show that the risk factors of anaemia among pregnant women vary in different settings this shows that the results from other countries might not fully represent the risk factors of pregnant women in the Kenyan context. The proposed research project will identify the risk factors of anaemia among pregnant women in the Kenyan context and show how these factors are comparable to the findings from other countries.

Though traditional statistical methods are easy to understand and their results are interpretable, machine learning approaches have the advantage of learning complex non-linear relationships between variables to arrive at well-informed inferences Weng (2020) Rajula (2020) Tripathy (2024) . Machine learning algorithms also can learn and analyse all types of data including images, and laboratory data among other factors that cause illness or recommend treatments. This indicates the need to test and validate machine learning algorithms in the medical field. The research project we are proposing will utilise the robustness of machine learning algorithms which have proved to perform better than the traditional conventional statistical methods.

Chapter 3

Methodology

3.1 Study Setting

The data used in this project was collected by PRECISE study from pregnant women attending their Antenatal Clinics at Mariakani and Rabai Hospitals in Kilifi County.

While Rabai is a level 3 hospital located in Rabai Sub County and serves the rural population, Mariakani is a level 4 hospital located in an urban setting. Mariakani Sub County Hospital is located along the Nairobi – Mombasa highway at Mariakani town and it is 34.3 kilometres from Mombasa city. Rabai Health Centre is located along Mazera – Kaloleni.

Patients with pregnancy complications from Rabai Hospital are referred to Mariakani Hospital for care and further management.

Figure 3.1 below shows the location of Mariakani and Rabai Hospitals in the map of Kilifi County

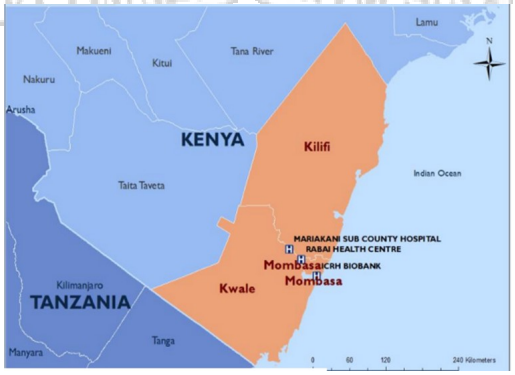


Figure 3.1: Kilifi County Map

3.2 Study Participants

Pregnant women aged between 16-49 years who attended their antenatal clinic visits at the two hospitals were enrolled into PRECISE Study. Also, PRECISE enrolled non-pregnant women of reproductive age and pregnant women with their pregnancy outcomes already known, for example, women who delivered stillbirths, women with foetal growth restriction and gestational hypertension. Non-pregnant women of reproductive age and women with known pregnancy outcomes will not be included in this analysis.

3.3 Study Design

This project used a predictive modelling study design. Supervised classification machine learning algorithms were used to find the association between anaemia as an outcome with demographic, medical history, socioeconomic, environmental, nutritional and clinical factors.

3.4 Data Source

This project utilised secondary data collected by the PRECISE study. This data was collected between June 2019 to December 2022. The PRECISE study was ethically approved by Aga Khan University and other ethical approval bodies.

3.5 Data Preprocessing

3.5.1 Handling Missing Data

Mode was used for imputation of missing categorical variables while mean was used for imputation of missing continuous variables.

3.6 Feature Scaling and Normalisation

Feature Scaling is the process of normalising the independent variables in a machine learning project before modelling. The main purpose of feature scaling is to ensure all the independent variables have the same scale to optimise the performance of machine learning algorithms.

The proposed research project used MinMax scaler to transform numerical variables because it is immune to outliers than the other feature scaling techniques proposed in the literature.

3.7 Fearure Encoding

Most machine learning algorithms perform well when the independent and dependent features are converted into numerical values. This project used both Label encoder and one-hot encoder to carryout encoding of categorical variables.

3.8 Feature Engineering

This is the process of creating new features based on existing features. It can include creating categorical variables from discrete variables.

The age of the participants was transformed into categories based on the domain knowledge. Pregnant women below 20 years were put into one group since they seem to be teenagers and will be interesting to understand the association of pregnant teenagers to anaemia. Participants aged between 20 – 34 also seem to have the same characteristics while those above 35 years were classified together.

Weight and Height was used in the calculation of body mass index (BMI). Most of the studies in the literature have used BMI rather than the individual variables of height and weight. The literature is also consistent with the domain knowledge provided by the expert I am working with.

Nutritional variables, like frequency of use of meat, vegetables etc have been used to calculate the dietary diversity score. This research project used the dietary diversity score. This standard score is calculated based on the individual nutritional variables.

3.9 Feature Selection

This is the process of identifying and using variables that are significantly associated with the outcome variable. Feature Selection helps remove redundant and irrelevant features from the dataset. Irrelevant features are not associated with the outcome variable but negatively affect the learning process of the algorithms. Redundant features do not have any effect on the learning process and also are not related to outcome variables. The literature has suggested the following algorithms that will be used to identify the best features of this analysis.

3.9.1 Filter Method

These algorithms do not take into account any classification modelling algorithm but use either Fisher Score, mutual information or relief to assign scores to the features. Filter methods measure the dependence between features and labels and calculate information gain between a feature and a label. High information gain implies the feature is relevant. The threshold and reference points for information gain will be defined in this study.

Mutual Information

It determines how the joint distribution of two variables, (A,B) is from the marginal product of the marginal distributions of A and B.

$$I(A;B) = \sum_{a \in A} \sum_{b \in B} P(a,b) \log \frac{P(a,b)}{P(a)P(b)} \tag{1}$$

where $P(a,b)$ is the joint probability of A and B, $P(a)$ is the marginal probability of A and $P(b)$ is the marginal probability of B.

The mutual information measures the amount of information that knowing the value of one variable provides about the other variable. It quantifies the reduction in uncertainty about one variable given knowledge of another variable.

The mutual information is zero if A and B are independent, meaning that knowing the value of one variable does not provide any information about the other variable. The mutual information is positive if A and B are dependent, meaning that knowing the value of one variable provides some information about the other variable.

The mutual information is symmetric, meaning that $I(A;B) = I(B;A)$. This means that the amount of information that knowing the value of A provides about B is the same as the amount of information that knowing the value of B provides about A. Alternatively, mutual information measures uncertainty reduction between variables A and B. This uncertainty is usually referred to as entropy and therefore mutual information can as well be defined in terms of entropy.

$$I(A;B) = H(A) - H(A|B) \tag{2}$$

where

$$H(A) = -\sum_{a \in A} P(a) \log P(a) \tag{3}$$

And conditional property of A given B is defined as;

$$H(A|B) = -\sum_{b \in B} P(b) \sum_{a \in A} P(a|b) \log P(a|b) \tag{4}$$

3.9.2 Wrapper Algorithm

These algorithms usually use one of the classification algorithms to select the best features from the feature set. For example, logistic regression can be used to identify the relevant features. One of the weaknesses of these approaches is they are biased towards the classification algorithm that was used.

3.9.3 Embedded Algorithms

Ensemble machine learning algorithms have the prebuilt ability to identify and rank the importance of independent variables. They use both filter wrapper approaches to identify the relevant independent variables for machine learning.

3.10 Machine learning Modelling

3.10.1 Splitting Data into Train Test Data.

The data was split into 75 % for training and 25% for testing. The skit-learn Train-Test method was used to partition the data randomly.

3.10.2 Possible Machine Learning Algorithms.

This research study is a supervised classification project, so it will use both logistic regression and ensemble machine-learning algorithms. Other possible machine learning algorithms include logistic regression, random forest, decision tree, Gradient boosting, and deep XG Boost algorithms.

The models were selected based on their complementary strengths in addressing different aspects of the classification problem. Logistic regression was included for its ability to identify linear relationships and its ease of interpretability. Tree-based algorithms, such as decision trees and random forests, were selected for their capacity to capture non-linear relationships and to provide insights into feature importance. Additionally, boosting algorithms, including XGBoost and Gradient Boosting, were chosen for their ability to iteratively correct the errors of previous models, thereby achieving higher predictive accuracy. These diverse yet complementary capabilities make the selected algorithms well-suited for the analysis of this machine learning problem.

Logistic regression

This transformed linear regression is used for supervised machine-learning tasks that aim to classify dependent variables into classes. It is suitable for binary outcomes but can also be used to model multi-class labels. Logistic regression predicts the probability of belonging to a particular class, 0 or 1 for binary and other specified classes for multi-class tasks.

Logistic regression is based on sigmoid function to accurately classify the labels into their respective classes. The sigmoid function transforms all real values to a probability between 0 and 1. Below is the formula for the sigmoid function.

$$P(Y = 1|X) = \sigma(z) = \frac{1}{1 + e^{-z}} \tag{5}$$

Logistic regression uses logit function below to represent the log-odds of the positive class. Because of the right-handside formula of the logit function, logistic regression works with the assumption that the features(independent variables) are linear.

$$\log\left(\frac{P}{1-P}\right) = w_0 + w_1 x_1 + \cdots + w_n x_n \tag{6}$$

Logistic regression uses the log-loss function as a performance measure between the actual and predicted classes.

$$\ell = -\sum_{i=1}^{N} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \tag{7}$$

where

$$y_i = \text{actual class label (0 or 1)} \tag{8}$$

$$\hat{y}_i = \text{predicted probability} \tag{9}$$

Decision Tree Classifier

This is a supervised machine learning algorithm used for classification tasks and it is an appropriate algorithm for prediction of anaemia among pregnant women. This project is a binomial classification supervised machine learning and a decision tree is suitable for identifying the predictors of anaemia. Decision trees do not assume any specific function in data and therefore they are best for identifying non-linear relationships and interactions between features.

Decision Tree classifier employs a tree data structure approach in classifying the classes based on the features. A tree data structure has nodes, branches(edges) and leaves. The starting node (topmost node) is called the root node and it represents the entire dataset. The data is then split into internal nodes based on a feature in the dataset. The leaf nodes are the final nodes at the lowest depth of the tree that represent the classified class labels. The root node, internal nodes and leaf nodes are connected through the branches(edges). The branches show the outcome of a test that leads to the next node.

The decision tree algorithm employs three impurity measures to determine nodes that best classify the outcome variable.

Gini Impurity – This metric is used to measure the likelihood of an incorrect classification of the selected sample. This is calculated by the following equation;

$$\text{Gini}(D) = 1 - \sum_{i=1}^{C} p_i^2 \tag{10}$$

where

p_i is the probability of class i in the dataset D.

C is the number of classes.

The value of gini impurity ranges from 0 to 1. When the value of gini impurity is 0, it means that all the classes are classified correctly. When the value of gini impurity is 1, it means that all the classes are classified incorrectly.

When the value of gini impurity is low then the classes are classified correctly. The objective is to minimize this metric so that classes at a particular node are correctly labelled.

Entropy – This is used to measure the uncertainty in the data. A larger value of entropy means there is high uncertainty in the dataset and therefore more prone to incorrectly classify the outcome variable. The objective is to minimize the value of entropy for the node to have correct labels. The following equation is used to classify entropy;

$$H(D) = -\sum_{i=1}^{C} p_i \log_2 p_i \tag{11}$$

where

p_i is the proportion of samples belonging to class i. C is the number of classes.

Information Gain – This metric is used to indicate the reduction in entropy after a split. A higher information gain indicates a better split. The following formula is used to determine information gain based on the entropy.

$$\text{IG} = H(D) - \sum_{j} \frac{|D_j|}{|D|} H(D_j) \tag{12}$$

where

$H(D)$ is the entropy of the dataset before the split.

$H(D_j)$ is the entropy of the dataset after the split.

$|D|$ is the total number of samples in the dataset.

$$\frac{|D_j|}{|D|} = \text{proportion of samples in subset } D_j$$

Random Forest Classifier

This classifier combines multiple decision trees to improve predictive performance and reduce overfitting. This classifier employs bagging(bootstrap aggregating), a method that randomly samples the data with replacement, each tree is trained on the randomly sampled data and their predictions are combined by majority voting. Unlike the decision tree classifier,

the Random Forest classifier samples the data and features to be trained at each tree. The number of features is determined by the formula m = $\sqrt{M}$

Random Forest's key parameters are:

- The parameter Number of Trees: The total number of trees in the forest.
- Number of Features: The number of features to consider at each split.
- Maximum Depth: The maximum depth of each tree.
- Minimum Samples per Leaf: The minimum number of samples required to be at a leaf node

Random Forest classifier uses the splitting metrics similar to the decision tree.

Gradient Boosting Classifiers.

Gradient and XGBoost boosting classifiers utilize the principle of boosting to combine weak decision trees into a strong predictive model. Each tree corrects the errors of the previous one during the tree-building process.

Gradient boosting algorithms usually minimize the objective function using the loss function and regularization term.

The predicted output for a given input is the sum of all predictions from the previous trees.

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathscr{F} \tag{13}$$

where

K is the number of trees.

f_k is the k-th tree.

$\mathscr{F}$ is the space of all possible trees.

x_i is the input feature vector.

$\hat{y}_i$ is the predicted output.

The objective function equation is given by;

$$\text{Obj} = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{14}$$

The regularization term penalizes the complexity of the tree and is given by the equation;

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{15}$$

where

γ is the complexity parameter. T is the number of leaves in the tree. λ is the regularization parameter. w_j is the weight of the j-th leaf.

Gradient boosting algorithm uses the second-order Taylor expansion to approximate the objective function as shown in the equation below;

$$\text{Obj}^{(t)} \approx \sum_{i=1}^{n} \left[L(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \tag{16}$$

$$g_i : \text{First-order gradient of the loss function w.r.t. } \hat{y}_i^{(t-1)} : \quad g_i = \frac{\partial L(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)}} \tag{17}$$

$$h_i : \text{Second-order gradient (Hessian) of the loss function w.r.t. } \hat{y}_i^{(t-1)} : \quad h_i = \frac{\partial^2 L(y_i, \hat{y}_i^{(t-1)})}{\partial (\hat{y}_i^{(t-1)})^2} \tag{18}$$

The key parameters of gradient boosting classifiers are;

- Learning Rate: It is used to control the contribution of each tree to the predictions. Lower values of learning rate make stable and robust predictions but it requires many trees.

- Subsample: his is the fraction of samples used to train each tree.
- Number of Trees $n\_estimators$: This is the number of trees.
- Maximum Depth: The depth of trees.
- Regularizarion: lambda$L2$,alpha$L1$ used for regularization of leaf weights while gamma used for minimum loss reduction required to make a split.

3.10.3 Model Evaluation

The following supervised classification algorithms evaluation metrics were used for the evaluation of the best performing algorithm.

Confusion Matrix

This matric table is sometimes called an error matrix or contingency table and it provides a detailed analysis of the classification performance of a model by showing the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

A confusion Matrix is a 2X2 matrix table that is used to represent the results of a binary classification model. The binary combination for the research project we are proposing will be anaemic or not anaemic.

The following are the definitions of the components of a confusion matrix.

True Positive (TP): The number of positive instances that are correctly classified as positive by the model. In this project, these are anaemic patients in the test data correctly predicted as anaemic.

False Positive (FP): The number of negative instances that are incorrectly classified as positive by the model. These will be the number of non-anaemic participants predicted as anaemic.

True Negative (TN): The number of negative instances that are correctly classified as negative by the model. These will be non-anaemic participants correctly predicted as non-anaemic.

False Negative (FN): The number of positive instances that are incorrectly classified as negative by the model. These will be non-anaemic participants predicted as anaemic.

Figure 3.2 shows an example of Confusion Matrix.

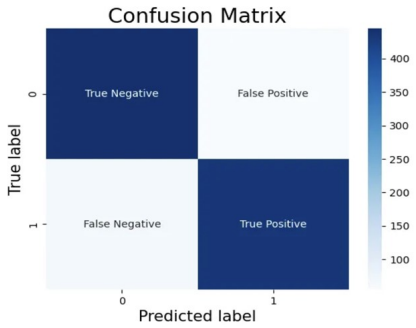


Figure 3.2: Confusion Matrix

Accuracy

This shows the ratio of correctly predicted records to the total number of records in the data. Confusion matric will be used to indicate the predicted labels against the true classes graphically. This metric is most suitable for balanced data and is one of the most used metrics for classification machine learning projects. A high accuracy score indicates that the model is making a large proportion of correct predictions, while a low accuracy score indicates that the model is making too many incorrect predictions.

$$Accuracy = \frac{TruePositives + TrueNegatives}{TotalSample} \tag{19}$$

Precision

This indicator is defined as the proportion of correct positive predictions. A high precision score indicates that the model can accurately identify positive instances, while a low precision score indicates that the model is making too many false positive (FP) predictions.

$$precision = \frac{TruePositives}{TruePositives + FalsePositives} \tag{20}$$

Recall

This metric is used to show the actual positive records that have been correctly predicted. This research project will utilize this variable because correctly predicting true positives in the medical field is very crucial. The number of incorrectly predicted positives should be reduced since we do not want to predict true positives as negatives. Incorrectly predicted positives will make patients miss interventions which is a big crisis in the medical file.

A high recall score indicates that the model can identify a large proportion of positive instances, while a low recall score indicates that the model is missing many positive instances.

$$Recall = \frac{TruPositives}{TruePositives + FalseNegatives} \tag{21}$$

F1-Score

This metric combines precision and recall to provide a comprehensive evaluation of a binary classification model. It measures the harmonic mean of precision and recall, giving equal importance to both metrics. A high F1-score indicates that the model is performing well in both precision and recall, while a low F1-score indicates that the model is not performing well in either precision or recall.

$$F1 - Score = 2 * \frac{Precision * Recall}{precision + Recall} \tag{22}$$

AUC-ROC

AUC denotes the Area Under the Curve while ROC denotes Receiver Operating Characteristic. The ROC is a graphical representation of a binary classifier that indicates the trade-off between true positive rate (TPR) and false positive rate (FPR) at different thresholds.

$$TruePositiveRate(TPR) = \frac{TruePositives}{TruePositives + FalseNegatives} \tag{23}$$

$$FalsePositiveRate(FPR) = \frac{FalsePositives}{TruePositives + TrueNegatives} \tag{24}$$

The ROC curve is obtained by plotting TPR against FPR at different thresholds. A perfect classifier would have a ROC curve that passes through the top-left corner (TPR = 1 and FPR = 0), while a random classifier would have a ROC curve that passes through the diagonal (TPR = FPR)

3.10.4 Model Deployment

The best-performing model was deployed to help in the early screening and identification of pregnant mothers who are at risk of developing anaemia. Python's Dash framework was be employed during the deployment phase.

3.10.5 Model Explainability

Model explainability is all about ensuring machine learning algorithms are not black boxes. This process ensures that results from machine learning algorithms can be interpreted and understood to enhance transparency, fairness, robustness, privacy, and interpretability and build trust among domain experts and the users of the machine learning models.

Model explainability is a key component in research especially in medical fields because health is a sensitive area. The healthcare personnel, end users and consumers of machine learning outputs not only want to trust the results from models but are also concerned about how the algorithms arrive at the clinical decisions.

The following techniques was used to achieve this purpose;

Using Interpretable Algorithms

Supervised Machine learning algorithms like logistic regression and decision trees have simple-to-follow steps and rules that show how the results are derived. It is possible to print a

tree structure of the entire process of the decision tree classification algorithm. This offers the opportunity to evaluate and audit the process of arriving at the best-performing independent variables against the target variable. This project used a decision tree as one of the supervised classification algorithms to identify the risk factors of anaemia in pregnancy and therefore, the tree structure was available to explain and be audited by the domain experts. But decision tree was not the best performing algorithm and therefore other methods were used to explain how the identified risk factors predict anaemia in pregnancy.

SHAP Method (SHaply, Additive exPlanations)

This method helps to explain model predictions by computing the contribution of each independent variable to the predictions. The contribution of each variable (SHAP values) is ranked and can be represented in a graph to show the feature importance for every variable. Domain experts and other medical practitioners can understand and interpret the results from machine learning algorithms based on the graphical chart. A Python library called SHAP has been developed to support in implementation of explainable AI in most of the machine learning algorithms.

SHAP values can be presented in terms of variable importance with summary plot, summary plot of a specific target and dependence plots.

Chapter 4

Results and Interpretation

4.1 Introduction

This chapter discusses the descriptive statistics, modeling, and interpretation of modeling results.

This is a binomial supervised classification project that involved the classification of patients into groups of whether anaemic or not anaemic. Anaemia status outcome variable is a binary variable derived from the continuous variable called haemoglobin concentration. Haemoglobin was collected by a trained research laboratory technologist and analyzed using a Sysmex device, which is the gold standard device for measuring haemoglobin.

The independent variables for this analysis are derived from the main groups of demographic, socio-economic, environmental, medical history and clinical anthropometries taken at the first antenatal contact with the pregnant women.

4.2 Data Understanding

Table 4.1 below represents the description of the variables used in this analysis. The description tries to provide a brief meaning to every variable that was collected and used in this analysis.

Table 4.1: Description of Dataset Variables

Variable	Description
f2a_participant_id	Unique Participant ID
f2_visit_date	The date participant visited facility
marital_status	Marital Status
age_enrolment	Maternal Age
occupation	Occupation
highest_school_level	Highest Education Level
maternal_bmi	Body Mass Index
wealth_quintile	Socioeconomic Wealth Quintile
PPI_score	Score of Poverty Index
extreme_poverty_line	Measures poverty level
poverty_line	Measure of poverty level
PPI_scoreX	Measure of Poverty Level
given_birth_before	If participant has given birth before
parity	Number of deliveries before current visit
previous_csection	History of c-section
f3_number_of_people_live_with	Number of people living with participant
f3_source_drinking_water	Source of drinking water
f3_source_water_cooking	Source of water for cooking
f3_water_source_location	Location of source of water
f4_contraception_last_year	If used contraception
f4_high_bp_or_hypertension	History of high blood pressure or hypertension
f4_high_blood_sugar_history	History of high blood sugar
f4_chronic_heart_condition_history	History of chronic heart condition
f4_pregnancy_unrelated_seizures	History of seizures not related to pregnancy
f6a_vitamins	Taking Vitamins
f6a_currently_taking_bp_medication	Taking blood pressure medication
f6a_currently_taking_diabetes_medication	Taking diabetes medication
f6a_tobacco	Using tobacco
f6a_antibiotics	Taking any antibiotics
f3_toilet_facility	Type of toilet facility at home
f3_toilet_facility_location	Location of toilet facility
f3_neighbor_help_pregnancy_problem	Neighbours who can offer help during pregnancy
f3_community_help_pregnancy_problem	Community help during pregnancy
f3_decision_maker_money	Main decision maker on household money
f3_decision_maker_pregnancy	Decision maker on pregnancy care spending
maternal_height	Maternal Height
maternal_weight	Maternal Weight
average_muac	Maternal mid-upper arm circumference
f6a_haemoglobin_concentration	Haemoglobin concentration
Anaemia_Status	Anaemia Status

4.3 Participants' Characteristics Table

The data in Table 4.2 below shows the table characteristics of for the dataset used in this analysis. It shows the distribution of data across key variables used in this analysis. The table provides a basic understanding of the dataset for this project.

Table 4.2: Demographic and Health Characteristics of the Study Population

Indicator	Categories	N = 3450
Age		
	< 19	117 (3.39%)
	19 – 24	1197 (34.70%)
	25 – 29	1012 (29.33%)
	30 – 34	710 (20.58%)
	35 – 39	337 (9.77%)
	≥ 40	76 (2.20%)
	Missing	1 (0.03%)
Parity		
	0 – 1	1912 (55.42%)
	2 – 4	1278 (37.04%)
	≥ 5	260 (7.54%)
BMI		
	Underweight	143 (4.14%)
	Normal	1827 (52.96%)
	Overweight	920 (26.67%)
	Obese	536 (15.54%)
	Missing	24 (0.70%)
Marital Status		
	Married/Co-habiting	3184 (92.29%)
	Never married (or single)	212 (6.14%)
	Separated/Divorced	54 (1.57%)
Occupation		
	Business	325 (9.42%)
	Construction	2 (0.06%)
	Factory	83 (2.41%)
	Housewife	1892 (54.84%)
	Informal – Employment	441 (12.78%)
	Large-scale agriculture	4 (0.12%)
	Market trader	340 (9.86%)
	Other (specify)	31 (0.90%)
	Professional	248 (7.19%)
	Student	69 (2.00%)
	Missing	15 (0.43%)
Education		
	Did not attend school	322 (9.33%)
	Higher	408 (1.83%)
	Primary	1812 (52.52%)
	Secondary	895 (25.94%)
	Missing	13 (0.38%)
Anaemia Status		
	Anaemic	1464 (42.43%)
	Non-Anaemic	966 (28.00%)
	Missing	1020 (29.57%)

The study population consisted of 3,450 participants, with their demographic and health characteristics summarized in Table 4.2. The majority of participants were young adults, with 34.7% (n = 1197) aged 19–24 years and 29.3% (n = 1012) aged 25–29 years. Smaller proportions were observed in older age groups, with 9.8% (n = 337) aged 35–39 years and only 2.2% (n = 76) aged 40 years or older. A minority (3.4%, n = 117) were under 19 years, and one participant had missing age data.

Regarding parity, more than half of the participants (55.4%, n = 1912) had 0–1 previous births, while 37.0% (n = 1278) had 2–4 births. A smaller group (7.5%, n = 260) had 5 or more births, indicating that most participants had relatively low parity. Body Mass Index (BMI) distribution showed that 52.9% (n = 1827) of participants had a normal BMI, while

26.7% (n = 920) were overweight and 15.5% (n = 536) were obese. Underweight participants accounted for 4.1% (n = 143), and 0.7% (n = 24) had missing BMI data. This suggests a significant prevalence of overweight and obesity in the study population.

Marital status was predominantly married or cohabiting, with 92.3% (n = 3184) falling into this category. Only 6.1% (n = 212) were never married or single, and 1.6% (n = 54) were separated or divorced, indicating a high rate of partnership among participants.

Occupationally, the majority (54.8%, n = 1892) identified as housewives, reflecting a traditional role for many participants. Other notable occupations included informal employment (12.8%, n = 441), market trading (9.9%, n = 340), and business (9.4%, n = 325). Professional occupations were reported by 7.2% (n = 248), while smaller groups were engaged in factory work (2.4%, n = 83), student roles (2.0%, n = 69), and other occupations such as construction (0.1%, n = 2) and large-scale agriculture (0.1%, n = 4). A small fraction (0.9%, n = 31) reported other unspecified occupations, and 0.4% (n = 15) had missing data.

Education levels varied, with 52.5% (n = 1812) having primary education and 25.9% (n = 895) having secondary education. A smaller proportion (11.8%, n = 408) had higher education, while 9.3% (n = 322) did not attend school. Missing education data were minimal (0.4%, n = 13). This distribution highlights the predominance of basic education among participants.

Anaemia status was assessed for a subset of the population, with 42.4% (n = 1464) classified as anaemic and 28.0% (n = 966) as non-anaemic. However, a substantial proportion (29.6%, n = 1020) had missing data for this indicator, suggesting potential challenges in data collection or testing for anaemia in this study.

The study population was characterised by a young with a mean age of 27.10, predominantly married group with low to moderate parity, a high prevalence of normal to overweight BMI, and a majority engaged in traditional roles such as housework. Education levels were mostly primary, and anaemia was common among those with available data, though missing data for this indicator warrants caution in interpretation.

4.4 Missing Data

Table 4.3: Percentage of Missing Data for Study Variables

Variable	% Missing
f2a_participant_id	0
f2_visit_date	0
marital_status	0
age_enrolment	0.03
occupation	0.43
highest_school_level	0.38
maternal_bmi	0.7
wealth_quintile	0
PPI_score	0.41
extreme_poverty_line	0.41
poverty_line	0.41
PPI_scoreX	0.41
given_birth_before	0.58
parity	0
previous_csection	0
f3_number_of_people_live_with	0.43
f3_source_drinking_water	0.41
f3_source_water_cooking	0.41
f3_water_source_location	0.41
f4_contraception_last_year	0.58
f4_high_bp_or_hypertension	0.58
f4_high_blood_sugar_history	0.58
f4_chronic_heart_condition_history	0.61
f4_pregnancy_unrelated_seizures	0.58
f6a_vitamins	0.61
f6a_currently_taking_bp_medication	0.64
f6a_currently_taking_diabetes_medication	0.61
f6a_tobacco	0.61
f6a_antibiotics	0.61
f3_toilet_facility	0.41
f3_toilet_facility_location	0.46
f3_neighbor_help_pregnancy_problem	0.41
f3_community_help_pregnancy_problem	0.55
f3_decision_maker_money	0.43
f3_decision_maker_pregnancy	0.52
maternal_height	0.49
maternal_weight	0
average_muac	0.38
f6a_haemoglobin_concentration	3.86
Anaemia_Status	29.57

Table 4.3 summarizes the extent of missing data for 40 variables in a study with 3,450 participants. It includes variables related to demographic, socio-economic, health, and pregnancy-related factors, such as marital_status, maternal_bmi, PPI_score, and Anaemia_Status. The percentage of missing data ranges from 0% (e.g., f2a_participant_id, marital_status) to 29.57% (Anaemia_Status), with most variables having less than 1% missing data (e.g., age_enrolment at 0.03%, occupation at 0.43%)

4.5 Data Preprocessing

4.5.1 Feature Engineering

The outcome variable(Anaemia Status) was derived from the continuous variable called hemoglobin concentration. Haemoglobin concentrations less than 11 grams/dl were recorded as anaemic while concentrations above or equal to 11 were recorded as non-anaemic.

The number of people living with the research study participant was collected as a continuous variable. Previous studies categorized this into a size of less than or equal to 5 and above 5 people, which represent a small or large household size, respectively. A categorical variable called household size was created.

The age variable was converted into a categorical variable based on similar characteristics of each group.

The body mass index variable was converted from continuous to categorical according to WHO recommendations for Underweight, Normal, Overweight, and Obese.

Literature has also converted parity variables into groups of 0 to 1, 2 to 4, and above or equal to 5 deliveries.

4.5.2 Missing Values Imputation

The records without missing data in the outcome variable were selected for further preprocessing and the following table indicates the new level of missing in the dataset.

Table 4.4: Number of Rows with Missing Data for Study Variables

Variable	Number of rows missing data
marital_status	0
occupation	4
highest_school_level	2
wealth_quintile	0
given_birth_before	7
previous_csection	0
f3_water_source_location	3
f4_contraception_last_year	7
f4_high_bp_or_hypertension	7
f4_high_blood_sugar_history	7
f4_chronic_heart_condition_history	7
f4_pregnancy_unrelated_seizures	7
f6a_vitamins	6
f6a_currently_taking_bp_medication	6
f6a_currently_taking_diabetes_medication	6
f6a_tobacco	6
f6a_antibiotics	6
f3_toilet_facility_location	4
f3_neighbor_help_pregnancy_problem	3
f3_community_help_pregnancy_problem	5
Anaemia_Status	0
age_group	0
parity_group	0
bmi_group	16
HH_size	4
household_mone_decisionmaker	3
pregnancy_money_decisionmaker	5
drinking_water_source	3
cooking_water_source	3
toilet_facility	3

The number of missing rows ranges from 0 (e.g., marital_status, wealth_quintile, Anaemia_Status) to 16 (bmi_group). Several variables, particularly those related to health history (e.g., f4_high_bp_or_hypertension, f4_contraception_last_year), have 7 missing rows, while others like f6a_vitamins and f6a_tobacco have 6. Variables such as drinking_water_source and f3_water_source_location show 3 missing rows each. The table indicates that most variables have minimal missing data, with bmi_group having the highest number of missing entries.

Median was used to impute continuous variables, while mode was used for categorical ones.

4.6 Feature Selection

This stage involves selecting features that are relevant for modeling. Filter, wrapper, and embedded feature selection techniques have been recommended by the literature. The following are the main features selection methods used in this analysis.

4.6.1 Filter Method

Mutual information was selected filter feature selection technique because of its ability to identify non-linear relationships and works well for both continuous and categorical data.

Mutual Information

Mutual information values were computed between the outcome variable, Anaemia Status, and each feature. Table 4.4 below shows the relationship between the outcome variable and features based on the mutual information.

Table 4.5: Mutual Information Scores for Features

Feature	Mutual Information
pregnancy_money_decisionmaker	0.018556351
highest_school_level	0.016912191
PPI_score	0.016266035
f4_pregnancy_unrelated_seizures	0.013083184
f4_high_blood_sugar_history	0.012394122
f3_neighbor_help_pregnancy_problem	0.009991033
maternal_bmi	0.006841207
toilet_facility	0.005732024
household_mone_decisionmaker	0.004805715
occupation	0.004634656
f3_water_source_location	0.004509684
age_enrolment	0.004450088
f6a_tobacco	0.003797815
previous_csection	0.002905825
f4_chronic_heart_condition_history	0.002356293
HH_size	9.29E-05
marital_status	0
f4_high_bp_or_hypertension	0
parity	0
f6a_currently_taking_bp_medication	0
f6a_currently_taking_diabetes_medication	0
f6a_antibiotics	0
f3_community_help_pregnancy_problem	0
drinking_water_source	0
cooking_water_source	0

Table 4.5 reveals that certain variables exhibit no association with the outcome variable at the individual level. Variables with a mutual information score of zero, such as `marital_status`, `f4_high_bp_or_hypertension`, and `drinking_water_source`, are independent of the outcome variable and thus do not contribute to predicting anaemia. Conversely, the table highlights that variables like `pregnancy_money_decisionmaker`, `highest_school_level`, and `PPI_score` (a proxy for wealth status) demonstrate a stronger association with anaemia, with mutual information scores of 0.0186, 0.0169, and 0.0163, respectively. However, mutual information, as a filter method, evaluates the relationship between each feature and the outcome variable independently, without accounting for interactions or interdependencies

among the independent variables. This limitation may overlook the combined effects of features in predicting the outcome, potentially underestimating their predictive power.

4.6.2 Embedded Method

This technique uses one of the machine learning algorithms to identify the main features that predict the outcome.

The decision tree machine learning algorithm was used to identify the best features that can predict anaemia. Table 4.6 below shows the scores per variable according to the decision tree machine learning algorithm.

Table 4.6: Feature Importance Scores

Feature	Importance
maternal_bmi	0.3035
PPI_score	0.1660
age_enrolment	0.1304
parity	0.0646
highest_school_level	0.0390
household_mone_decisionmaker	0.0310
drinking_water_source	0.0305
occupation	0.0303
cooking_water_source	0.0300
f3_water_source_location	0.0290
pregnancy_money_decisionmaker	0.0268
HH_size	0.0220
f3_neighbor_help_pregnancy_problem	0.0200
toilet_facility	0.0142
f3_community_help_pregnancy_problem	0.0118
f4_pregnancy_unrelated_seizures	0.0114
previous_csection	0.0101
marital_status	0.0099
f4_high_bp_or_hypertension	0.0079
f6a_antibiotics	0.0052
f6a_tobacco	0.0029
f4_high_blood_sugar_history	0.0027
f4_chronic_heart_condition_history	0.0015
f6a_currently_taking_diabetes_medication	0.0000
f6a_currently_taking_bp_medication	0.0000

Embedded methods are often preferred due to their ability to account for feature interactions when predicting the outcome variable.

Table 4.6 indicates that `maternal_bmi`, `PPI_score` (a proxy for wealth status), `age_enrolment`, `parity`, `highest_school_level`, and `pregnancy_money_decisionmaker` are key predictors of anaemia, with importance scores of 0.3035, 0.1660, 0.1304, 0.0646, 0.0390, and 0.0268, respectively. However, some variables that showed no association with anaemia according to mutual information (e.g., `f4_high_bp_or_hypertension` exhibits a modest

relationship with the outcome variable based on decision tree feature importance, with scores such as 0.0079 and 0.0000, respectively. This discrepancy highlights the advantage of embedded methods in capturing subtle feature contributions that filter methods like mutual information may overlook.

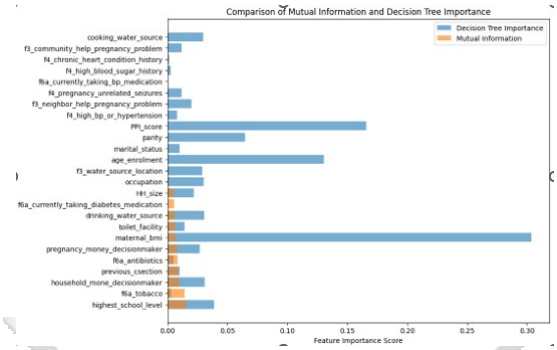


Figure 4.1: Comparison of Mutual Information and Decision Tree Feature Selection

Figure 4.1 illustrates that `maternal_bmi` is the most significant predictor of anaemia according to the decision tree algorithm, with the highest feature importance score. Furthermore, the decision tree demonstrates robustness in identifying key predictors of anaemia, as variables such as cooking_water_source, f3_community_help_pregnancy_problem, f4_chronic_heart_condition_history, f4_pregnancy_unrelated_seizures, and marital_status which exhibit no association with anaemia based on mutual information—align with expectations of independence from the outcome variable.

4.7 Data Visualization

The following visualizations show the distribution of features according to anaemia status outcome variable.

Marital Status Variable

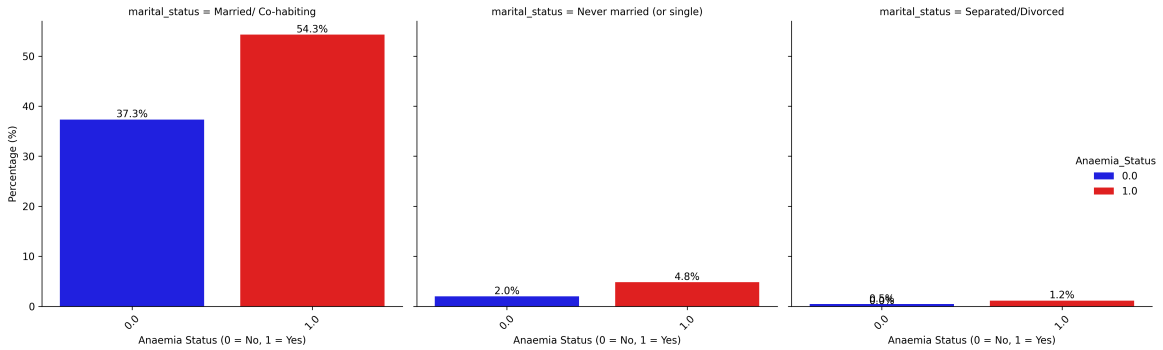


Figure 4.2: Prevalence of Anaemia in marital status variable

The figure 4.2 analysis examined the prevalence of anemia across three marital status groups: Married/Co-habiting, Never Married (or Single), and Separated/Divorced. The results are presented as percentages of individuals with anemia (Anemia Status = 1, Yes) and without anemia (Anemia Status = 0, No) within each group.

Married/Co-habiting Group

Among individuals who are married or co-habiting, 54.3% were found to have anemia, while 37.3% did not have anemia. This group exhibited the highest prevalence of anemia among the three marital status categories, with a notable difference of 17.0 percentage points between those with and without anemia. The high percentage of anemia in this group suggests that married or co-habiting individuals may be more susceptible to anemia compared to other marital status groups in this dataset.

Never Married (or Single) Group

In the never married or single group, 4.5% of individuals were identified as having anemia, compared to 2.0% who did not have anemia. Although the overall prevalence of anemia in this group is lower than in the married/co-habiting group, the percentage of individuals with anemia is more than double that of those without anemia, indicating a relatively higher burden of anemia within this subgroup.

Separated/Divorced Group

For individuals in the separated or divorced group, 1.2% were found to have anemia, while 0.5% did not have anemia. This group showed the lowest overall percentages for both anemia and non-anemia statuses. Despite the small percentages, the proportion of individuals with anemia is more than double that of those without anemia, consistent with the trend observed in the other groups.

Comparative Analysis

Across all marital status groups, the prevalence of anemia was consistently higher than the percentage of individuals without anemia. The married/co-habiting group had the highest prevalence of anemia at 54.3%, followed by the never married/single group at 4.5%, and the separated/divorced group at 1.2%. A similar pattern was observed for those without anemia, with 37.3% in the married/co-habiting group, 2.0% in the never married/single group, and 0.5% in the separated/divorced group. The largest disparity between anemia and non-anemia statuses was observed in the married/co-habiting group, while the smallest relative difference was in the separated/divorced group.

Household Size Variable

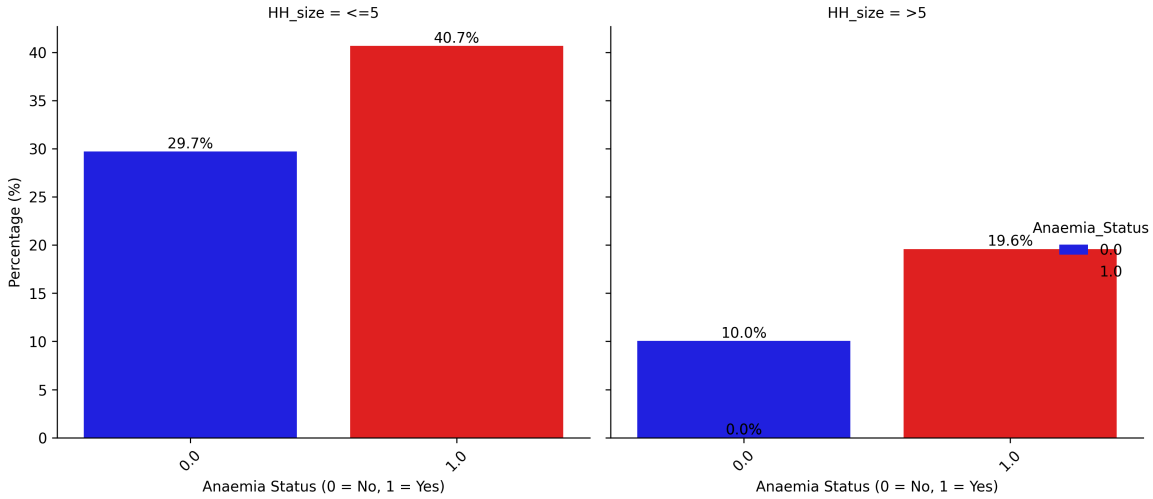


Figure 4.3: Prevalence of Anaemia in Accross the different household sizes

The analysis examined the prevalence of anemia across households of different sizes, specifically comparing households with 5 or fewer members (HH_size <= 5) to those with more than 5 members (HH_size > 5). The results are summarized in the bar chart.

For households with 5 or fewer members (HH_size <= 5), 29.7% of individuals were found to have no anemia (Anemia_Status = 0), while 40.7% were identified as having anemia (Anemia_Status = 1). This indicates a higher prevalence of anemia in smaller households, with a difference of 11.0 percentage points between those with and without anemia.

Education Level

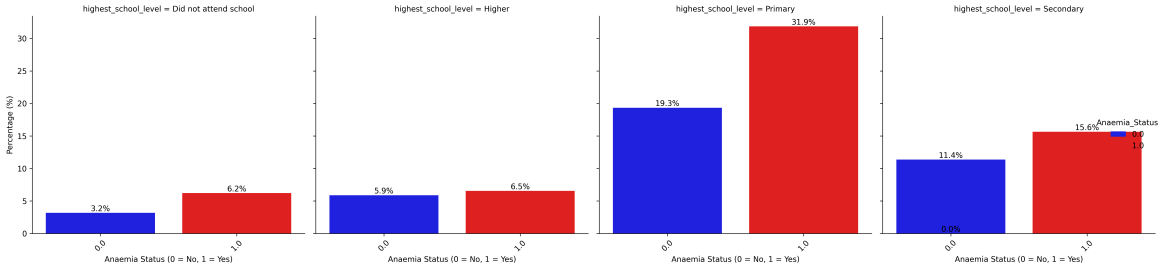


Figure 4.4: Prevalence of Anaemia in Accross the different Education Levels

The analysis further explored the prevalence of anemia based on the highest level of school attended, categorized into four groups: Did not attend school, Higher, Primary, and Secondary. The results are presented in the bar chart.

For individuals who did not attend school (highest_school_level = Did not attend school), 3.2% had no anemia (Anemia_Status = 0), while 6.2% were identified as having anemia (Anemia_Status = 1). This indicates a higher prevalence of anemia among those who did not attend school, with a difference of 3.0 percentage points between those with and without anemia.

Among individuals with a higher education level (highest_school_level = Higher), the prevalence of anemia was similar to the "Did not attend school" group. Specifically, 5.9% had no anemia (Anemia_Status = 0), and 6.5% had anemia (Anemia_Status = 1), a difference of 0.6 percentage points, suggesting a relatively balanced distribution.

For those with a primary education (highest_school_level = Primary), the prevalence of anemia was notably higher. Only 19.3% of individuals had no anemia (Anemia_Status = 0), while 31.9% were identified as having anemia (Anemia_Status = 1). This represents a

substantial difference of 12.6 percentage points, indicating that anemia is more common among individuals with a primary education compared to those with no anemia.

Finally, for individuals with a secondary education (highest_school_level = Secondary), 11.4% had no anemia (Anemia_Status = 0), while 15.6% had anemia (Anemia_Status = 1), a difference of 4.2 percentage points. Additionally, the chart includes small proportions of individuals in this group with an anemia status of 0.0 (0.0%) and 1.0 (1.0%), which may suggest a data categorization or labeling issue that requires further investigation.

Overall, the data indicates that anemia prevalence varies across education levels, with the highest prevalence observed among individuals with a primary education (31.9%) and the lowest among those with a higher education (6.5%). Individuals who did not attend school and those with a higher education showed similar anemia prevalence rates (6.2% and 6.5%, respectively), while those with a secondary education had a moderate prevalence of 15.6%.

Wealth Status

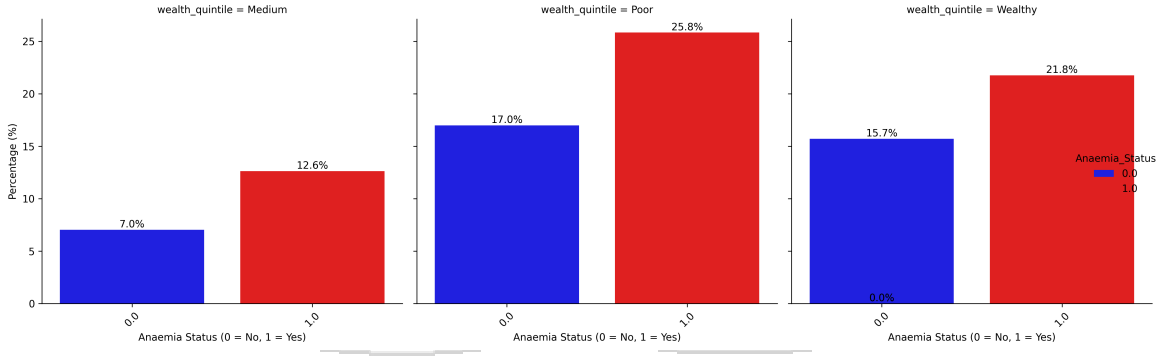


Figure 4.5: Prevalence of Anaemia in Accross the different Wealth Status

The prevalence of anemia was analyzed across three wealth quintiles: Medium, Poor, and Wealthy. In the Medium wealth quintile, 7.0% of individuals had no anemia (Anemia%Status = 0), while 12.6% had anemia (Anemia%Status = 1), a difference of 5.6 percentage points. In the Poor wealth quintile, 17.0% had no anemia, and 25.8% had anemia, a difference of 8.8 percentage points. For the Wealthy quintile, 15.7% had no anemia, and 21.8% had anemia, a difference of 6.1 percentage points, with small proportions (0.0% for Anemia%Status = 0 and 1.0% for Anemia%Status = 1) suggesting a potential data issue. The Poor wealth

quintile exhibited the highest anemia prevalence (25.8%), followed by the Wealthy (21.8%) and Medium (12.6%) quintiles.

Body Mass Index

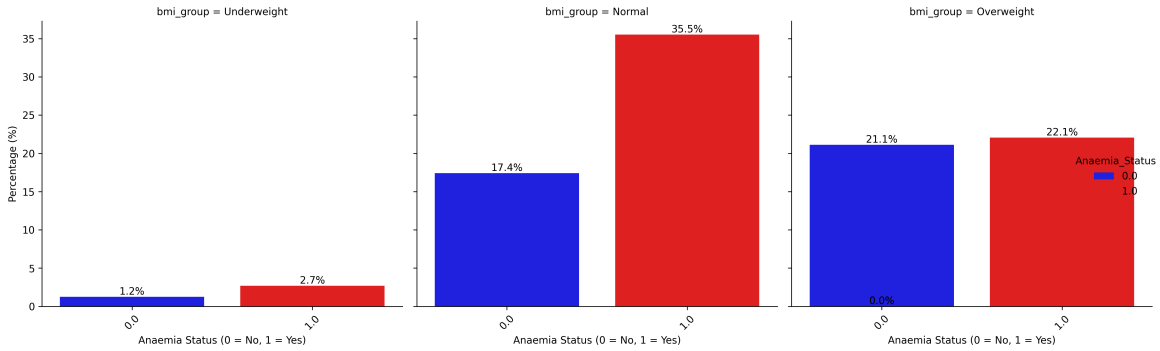


Figure 4.6: Prevalence of Anaemia in Accross the different Body Mass Index

The prevalence of anemia was analyzed across three BMI groups: Underweight, Normal, and Overweight. In the Underweight group (bmi_group = Underweight), 1.2% of individuals had no anemia (Anemia_Status = 0), while 2.7% had anemia (Anemia_Status = 1), a difference of 1.5 percentage points. In the Normal BMI group (bmi_group = Normal), 17.4% had no anemia, and 35.5% had anemia, a substantial difference of 18.1 percentage points, indicating a high burden of anemia in this group. For the Overweight group (bmi_group = Overweight), 21.1% had no anemia, and 22.1% had anemia, a difference of 1.0 percentage points, with small proportions (0.0% for Anemia_Status = 0 and 1.0% for Anemia_Status = 1) suggesting a potential data categorization issue. The Normal BMI group exhibited the highest anemia prevalence (35.5%), followed by the Overweight (22.1%) and Underweight (2.7%) groups.

4.8 Modelling

This section covers train test dataset splitting, machine learning algorithms used, evaluation metrics, hyperparameter tuning, and model explainability.

4.8.1 Train Test Splitting

The data was split into two groups: one for training and one for testing. 75% of the data was used for training, which means it was used to teach or build a model. The remaining 25% was set aside for testing, to check how well the model works on new data it hasn't seen before.

4.8.2 Machine Learning Models

Logistic regression, decision tree, random forest, XGBoost, and gradient boosting algorithms were selected for data analysis and model development.

4.8.3 Model Evaluation

The performance of machine learning models must be measured against the established metrics and indicators to determine whether they are performing well. Different metrics have been developed to assess the robustness of machine learning algorithms depending on their type. Recall, Accuracy, Precision, F-Score, and AUC-ROC are the main metrics used to evaluate supervised classification machine learning algorithms.

Table 4.7: Performance Metrics of Machine Learning Algorithms

ML Algorithm	Accuracy	Precision	Recall	F1_Score	ROC-AUC Score
Logistic Regression	0.6168	0.6235	0.8676	0.7256	0.6129
Decision Tree	0.5905	0.6456	0.6620	0.6537	0.5760
Random Forest	0.5806	0.6142	0.7577	0.6784	0.5902
XGBoost	0.5855	0.6272	0.7155	0.6684	0.5912
Gradient Boosting	0.6250	0.6315	0.8592	0.7279	0.6263

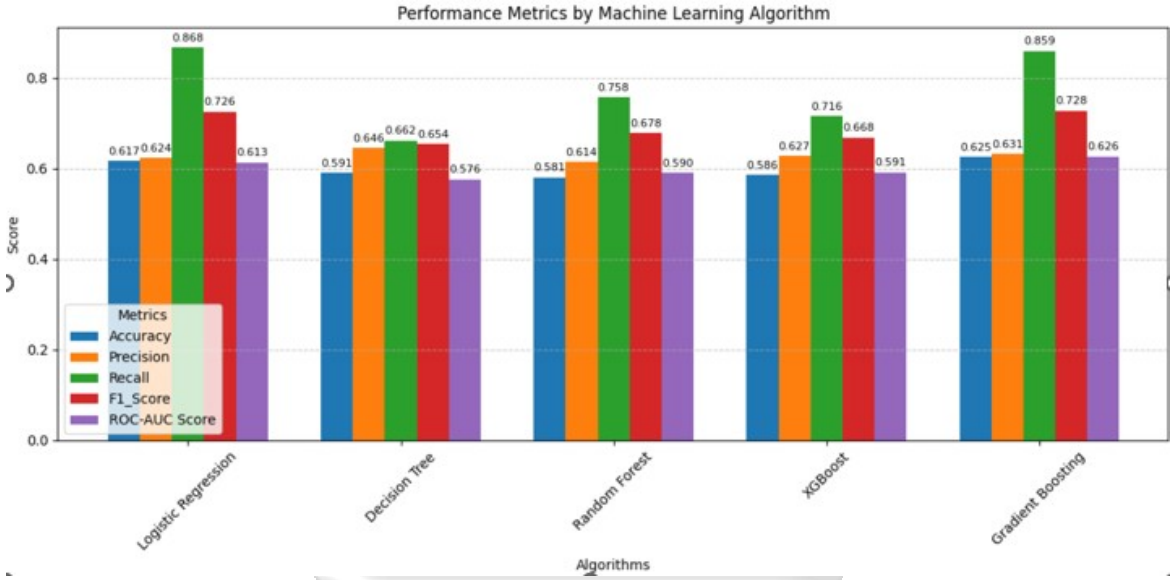


Figure 4.7: Performance of Machine Learning Models

Logistic Regression:

- Accuracy: 0.617 (about 62% correct overall)
- Precision: 0.624 (62% of its "yes" guesses were correct)
- Recall: 0.868 (it found 87% of the people who actually have anemia)
- F1 Score: 0.726 (a good balance between precision and recall)
- ROC-AUC: 0.613 (61% good at separating "yes" and "no")
- This method is good at finding people with anemia because of a recall of 87%

Decision Tree:

- Accuracy: 0.591 (59% correct)
- Precision: 0.646 (65% of its "yes" guesses were correct)
- Recall: 0.662 (it found 66% of the people with anemia)
- F1 Score: 0.654 (a decent balance)
- ROC-AUC: 0.576 (58% good at separating "yes" and "no")

- This method is okay but not great. It's better at being careful with its "yes" guesses (precision) than finding all anemia cases (recall). It's the weakest at telling "yes" and "no" apart because of the ROC-AUC of 58%

Random Forest:

- Accuracy: 0.581 (58% correct)
- Precision: 0.614 (61% of its "yes" guesses were correct)
- Recall: 0.758 (it found 76% of the people with anemia)
- F1 Score: 0.678 (a decent balance)
- ROC-AUC: 0.590 (59% good at separating "yes" and "no")
- This method is not the best overall (lowest accuracy), but it's good at finding people with anemia (recall). It's not as careful with its guesses (precision is lower), and it's average at telling "yes" and "no" apart because of recall of 59%

XGBoost:

- Accuracy: 0.586 (59% correct)
- Precision: 0.627 (63% of its "yes" guesses were correct)
- Recall: 0.716 (it found 72% of the people with anemia)
- F1 Score: 0.668 (a decent balance)
- ROC-AUC: 0.591 (59% good at separating "yes" and "no")
- This method is similar to Random Forest but a bit better in some areas. It's okay at finding anemia cases (recall) and being careful with its guesses (precision).

Gradient Boosting:

- Accuracy: 0.625 (63% correct)
- Precision: 0.632 (63% of its "yes" guesses were correct)

- Prevalence: 0.859 (it found 86% of the people with anemia)
- F1 Score: 0.728 (the best balance between precision and recall)
- ROC-AUC: 0.626 (63% good at separating "yes" and "no")
- This method is the best overall. It has the highest accuracy, the best balance (F1 Score), and the best ability to tell "yes" and "no" apart with recall of 63%. It's also great at finding people with anemia (high recall)

4.8.4 Model Explainability

SHAP Summary Plot The chart below is a SHAP (SHapley Additive exPlanations) summary plot, which helps us understand how different factors (or "features") influence a machine learning model's predictions—in this case, predicting whether someone has anemia. The chart lists the most important factors on the left side, and the bars show how much each factor impacts the model's predictions on average (measured by the "mean SHAP value"). The longer the bar, the more important that factor is in deciding whether someone has anemia.

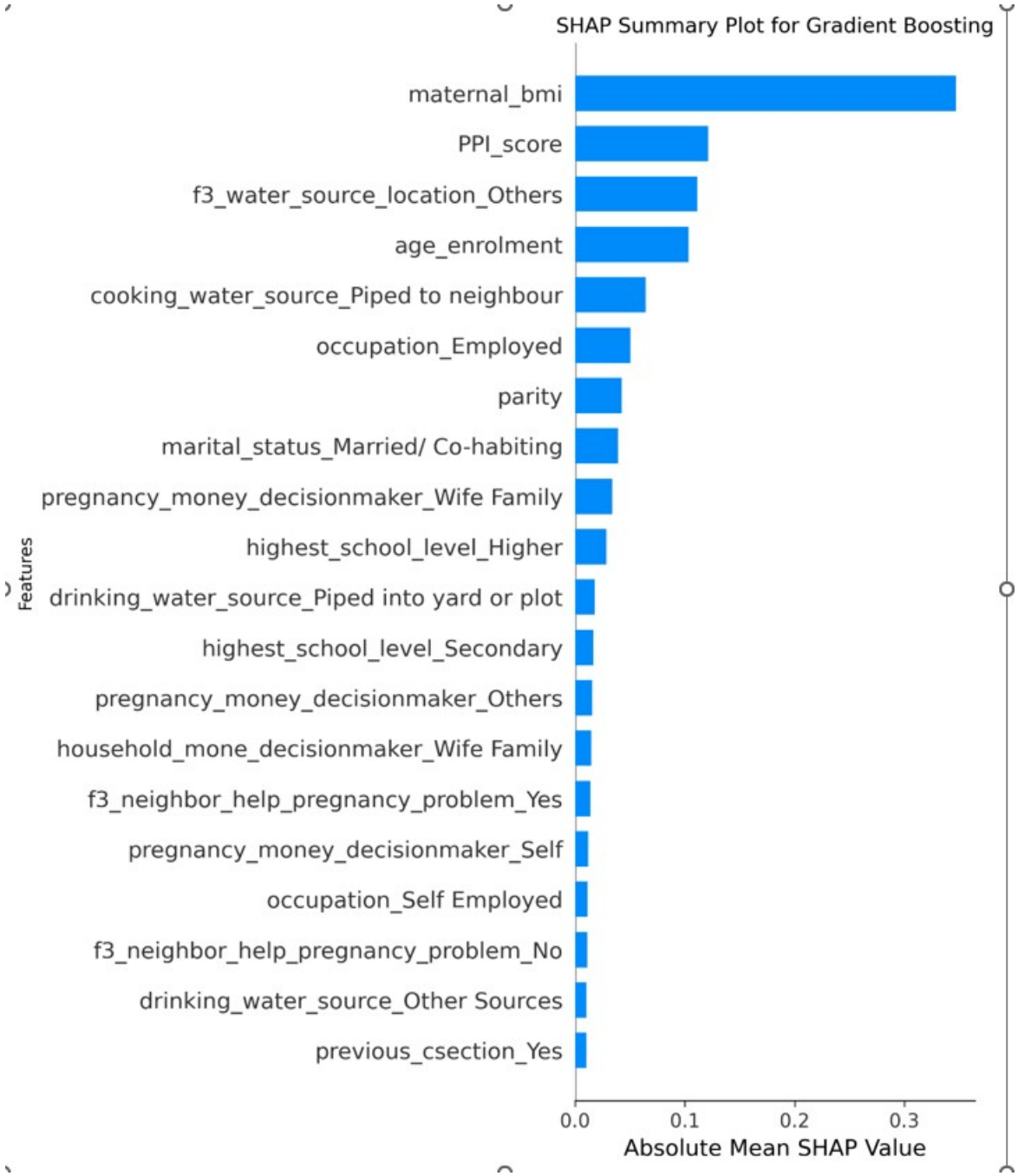


Figure 4.8: SHAP Summary Plot to show Feature Importance

- The Y-axis has features which are the factors the model uses to make predictions.
- The X-axis has the mean SHAP values hows how much each factor influences the model's prediction on average. A higher value means the factor has a bigger impact on predicting anemia.

- The features are ranked from most important (top) to least important (bottom) based on their SHAP values.

SHAP Beeswarm Plot

This chart is a SHAP (SHapley Additive exPlanations) Beeswarm plot for a Gradient Boosting model, which helps us understand how different features influence the model's predictions about whether someone has anemia. Unlike the previous SHAP plot, which showed only the average impact of each feature (mean SHAP value), this plot provides the direction to which the feature impact anaemia.

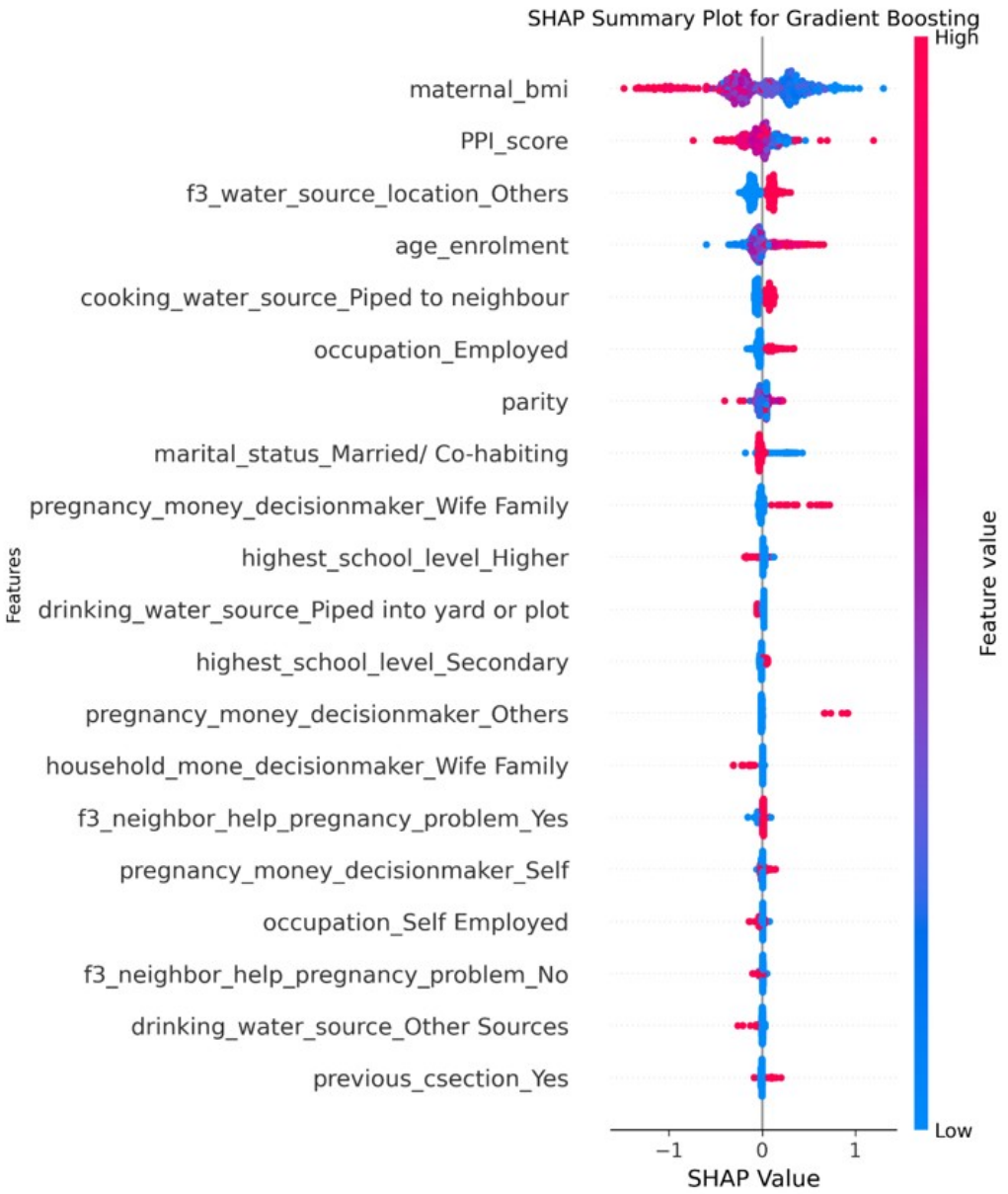


Figure 4.9: SHAP Beeswarm Plot to show direction of each Feature

Beeswarm Plot has the following Components

- Features (Y-Axis) which represent the variables used by the model to make anaemia predictions.
- SHAP Value (X-Axis): The horizontal axis shows the SHAP value, ranging from -1 to 1. A positive SHAP value (to the right) means the feature value increases the

likelihood of predicting anemia (pushes the prediction toward "Yes, the person has anemia") while A negative SHAP value (to the left) means the feature value decreases the likelihood of predicting anemia (pushes the prediction toward "No, the person does not have anemia").

- Feature Value (Color): The color of each dot represents the value of the feature. Red means a high value for that feature while Blue means a low value for that feature.
- Dots: Each dot represents an individual in the dataset. The spread of dots shows how much that feature's value impacts the prediction for different people.

Chapter 5

Discussions, Conclusions and Recommendations

5.1 Introduction

This study employed Gradient Boosting, a machine learning model, to predict the likelihood of anemia among pregnant women based on various socioeconomic, demographic, and health-related factors. The model analyzed key predictors such as maternal body mass index (BMI), socioeconomic status (measured through the Poverty Probability Index or PPI score), water source, age at enrollment, and educational attainment. To interpret the model's decision-making process, we used SHapley Additive exPlanations (SHAP), which allowed us to assess the influence of each predictor on anemia risk. Below, we discuss our key findings in detail, comparing them to existing literature to provide further context.

5.2 Discussions

This section will discuss the performance of machine learning algorithms and the risk factors of anaemia identified by Gradient boosting.

5.2.1 Machine Learning Performance

The Gradient Boosting algorithm performed well in correctly distinguishing between anaemic and non-anaemic classes, achieving an AUC-ROC score of 0.63%. This indicates that the model has a high level of accuracy in predicting anaemia risk. The model also performed

well on the recall(sensitivity) metric, achieving a score of 0.86%. This suggests that the model is effective in identifying individuals who are at risk of anaemia, which is crucial for early intervention and treatment. WHO guidelines recommends sensitivity as a key metric for evaluating the performance of screening tests, particularly in the context of public health initiatives aimed at identifying patients at risk of diseases WHO (2023b).

Boosting machine learning algorithms, such as Gradient Boosting, are known for their ability to handle non-linear relationships in data and interactions between features, making them suitable for complex datasets like the one used in this study. The model's performance is consistent with previous research that has demonstrated the effectiveness of boosting machine learning algorithms in predicting anaemia. Dejene (2022b). In this research study, cat boost algorithm outperformed decision tree, random forest, and XGBoost algorithms in predicting anaemia in pregnancy.

Logistic regression had a high score of 0.87% for the recall metric, outperforming all the selected models but did not perform well on the AUC-ROC metric, achieving a score of 0.61%. This indicates that while logistic regression is effective in identifying individuals at risk of anaemia, it may not be as reliable in distinguishing between anaemic and non-anaemic classes. This discrepancy highlights the importance of using multiple evaluation metrics to assess model performance comprehensively.

5.2.2 Risk Factors of Anaemia in pregnancy

1. Maternal Body Mass Index (BMI) and Anemia Risk

One of the strongest predictors of anemia in our study was BMI, a measure of body fat based on weight and height. Our SHAP analysis revealed that individuals with a lower BMI (represented by blue dots) were more likely to be classified as anemic, with SHAP values reaching up to +1. Conversely, those with a higher BMI (represented by red dots) were less likely to be predicted as anemic, with SHAP values as low as -1. This suggests that being underweight is a significant risk factor for anemia, likely due to inadequate nutritional intake and reduced iron stores, while a higher BMI appears to provide some level of protection.

These findings align with prior studies conducted a study in Ethiopia Kara (2020) and found that underweight pregnant women were 1.5 times more likely to develop anemia compared to those with normal BMI (AOR = 1.5, 95% CI: 1.1–2.0) due to inadequate nutritional intake Ahmed (2019). Similarly, a meta-analysis Kara (2020) confirmed that underweight women had a 40% higher likelihood of anemia (pooled OR = 1.4, 95% CI: 1.2–1.7) due to insufficient iron intake.

Interestingly, our study also observed that overweight women had a lower risk of anemia, consistent with Otieno (2024) findings. However, Rahman (2016) cautioned that severe obesity might contribute to anemia due to chronic inflammation impairing iron metabolism (OR = 1.2, 95% CI: 1.0–1.4). While our summary statistics indicated that the highest prevalence of anemia was among individuals with a "Normal" BMI (35.5%), followed by "Overweight" (22.1%) and "Underweight" (2.7%), this apparent discrepancy may be due to sample size limitations or differences in how BMI categories were grouped.

2. Socioeconomic Status (PPI Score) and Anemia Risk

Economic status, measured using the Poverty Probability Index (PPI) score, was another significant predictor of anemia. Our model showed that individuals with a lower PPI score—indicating higher levels of poverty—were at an increased risk of anemia, with SHAP values reaching up to +0.8. Conversely, wealthier individuals (higher PPI scores) were less likely to be anemic, with SHAP values as low as -0.8. This suggests that lower-income individuals face a greater risk of anemia, likely due to poor dietary quality, limited healthcare access, and higher exposure to environmental health risks.

This finding is consistent with previous research. Tripathy (2024) conducted a large-scale review and found that women from lower-income households were nearly twice as likely to develop anemia (AOR = 1.8, 95% CI: 1.5–2.2) due to inadequate nutrition and limited healthcare access Rajula (2020). Similarly, Levy (2023) reported that individuals in the lowest socioeconomic groups were over twice as likely to be anemic (RR = 2.1, 95% CI: 1.8–2.5).

3. Water Source and Anemia Risk

Access to clean water also emerged as a crucial factor influencing anemia risk. Our SHAP analysis indicated that individuals who relied on external water sources—such as shared wells or community taps—were at a higher risk of anemia, with SHAP values increasing up to +0.5. In contrast, those with a personal water source (in-home or yard-based) had a lower risk, with SHAP values dropping to -0.5. This suggests that inadequate water access contributes to anemia, potentially due to increased exposure to waterborne diseases, parasitic infections, and poor hygiene.

This finding aligns with previous studies. Berhe (2022) found that African women without access to clean water had a higher likelihood of anemia (AOR = 1.4, 95% CI: 1.2–1.7), likely due to infections such as hookworm and malaria. Similarly, Stephen (2018) reported that pregnant women in India without piped water access were more likely to be anemic (AOR = 1.5, 95% CI: 1.3–1.8) due to inadequate sanitation and higher disease burden [20].

4. Age at Enrollment and Anemia Risk

Age was also a significant predictor of anemia, with older individuals showing a higher risk. Our SHAP analysis indicated that increased age was associated with a higher likelihood of anemia (SHAP values up to +0.4), while younger participants had a lower risk (SHAP values down to -0.4). This may be due to cumulative nutritional deficiencies, increased parity, or age-related health conditions. These results are consistent with global findings. Okube (2022) observed that women aged 35–49 were more likely to have anemia (RR = 1.3, 95% CI: 1.1–1.5) compared to those aged 15–24, likely due to prolonged nutrient depletion Dejene (2022a). Lone et al. Otieno (2024) similarly found that pregnant women aged 30 and above were at increased risk (AOR = 1.6, 95% CI: 1.2–2.0) due to the cumulative effects of multiple pregnancies.

5. Education

Education was another significant predictor, with higher education levels reducing the likelihood of anemia. Our SHAP analysis showed that individuals with a higher education level had lower anemia risk (SHAP values down to -0.2), while those with less education had

an increased risk (SHAP values up to +0.2). This may be due to greater health awareness, improved dietary choices, and better economic opportunities.

This aligns with research [23], which found that women with higher education had a 30% lower risk of anemia (AOR = 0.7, 95% CI: 0.6–0.9) (Harding et al., 2017). Similarly, Dejene (2022a) reported that women with secondary or higher education had lower anemia prevalence (AOR = 0.8, 95% CI: 0.7–0.9).

5.3 Conclusions

The modelling results from this work show that maternal body mass index, wealth level, age of women, source of water for home consumption, and education levels are crucial predictors of anaemia in pregnancy.

The performance of gradient boosting in correctly predicting women with anaemia demonstrates that machine learning and data science approaches can be leveraged in a medical setup to detect diseases, conditions, infections, and health abnormalities early for early initiation of medical interventions.

5.4 Recommendations

Based on the findings from this analysis, the following recommendations can be made to specific stakeholders to help address the risk of women developing anaemia during pregnancy;
i. Financial empowerment programs should be put in place to improve the economic levels of women in the community.

ii. Improving the water sources infrastructure so that the families can access water at their places of residence will reduce the risk of anaemia.

5.5 Limitations of the Study

Although the data collection team was well trained and followed the standard operating procedures to collect the data sometimes self-reported data is subject to recall bias. These biases may have influenced the accuracy of the collected data, potentially affecting the findings of this study. Future research could mitigate this limitation by incorporating objective clinical measurements, such as laboratory-confirmed hemoglobin levels, to validate self-reported responses.

Additionally, this study did not distinguish between different types and stages of anemia. Since anemia presents in various forms—including iron-deficiency anemia, folate-deficiency anemia, and anemia due to chronic disease—each with distinct risk factors and health implications, future research could focus on classifying and analyzing these subtypes separately. A more granular investigation into the different types and progression stages of anemia could provide deeper insights into targeted interventions and treatment strategies.

Furthermore, the study utilized cross-sectional data, which captures information at a single point in time. While this approach provides useful associations between variables, it does not establish causal relationships. A longitudinal study design, where data is collected over multiple time points, would allow for a more comprehensive understanding of how risk factors influence anemia progression over time. Future research could adopt a longitudinal approach and compare its findings with those from cross-sectional studies to enhance the robustness of conclusions.

References

Ahmed, S. (2019). Prevalence and associated factors of anemia among pregnant women receiving antenatal care (anc) at fatima hospital in jashore, bangladesh: A cross-sectional study. *PUBMED*.

Azhar, B. S. (2021). Prevalence of anemia and associated risk factors among pregnant women attending antenatal care in bangladesh: a cross-sectional study,. *Prime Health Care*.

Berhe, K. (2022). Risk factors of anemia among pregnant women attending antenatal care in health facilities of eastern zone of tigray, ethiopia, case-control study. *PUBMED*.

Daru, J. (2018). Risk of maternal mortality in women with severe anaemia during pregnancy and post partum: a multilevel analysis. *Lancet Glob. Health*.

Dejene, B. (2022a). Predicting the level of anemia among ethiopian pregnant women using homogeneous ensemble machine learning algorithm. *BMC*.

Dejene, B. E. (2022b). Predicting the level of anemia among ethiopian pregnant women using homogeneous ensemble machine learning algorithm. *BMC Medical Informatics and Decision Making*.

Kara, W. (2020). Anaemia in pregnancy in southern tanzania: Prevalence and associated risk factors. *Reproductive Health*.

Levy, A. (2023). Maternal anemia during pregnancy is an independent risk factor for low birthweight and preterm delivery,. *Eur. J. Obstet. Gynecol. Reprod*.

Lin, L. (2018). Prevalence, risk factors and associated adverse pregnancy outcomes of anaemia in chinese pregnant women: a multicentre retrospective study. *BMC*.

Mehrnoush, V. (2023). Prediction of postpartum hemorrhage using traditional statistical analysis and a machine learning approach. *AJOG Glob*.

Melku, M. (2014). Prevalence and predictors of maternal anemia during pregnancy in gondar, northwest ethiopia: An institutional based cross-sectional study,. *PUBMED*.

Mostafa, E. (2022). Prevalence and risk factors of iron deficiency anaemia with pregnancy at minia university hospital,. *Minia Journal of Medical Research*.

Odhiambo, J. N. (2020). Mapping of anaemia prevalence among pregnant women in kenya (2016–2019). *BMC Pregnancy and Childbirth*.

Okube, O. (2022). Prevalence and factors associated with anaemia among pregnant women attending antenatal clinic in the second and third trimesters at pumwani maternity hospital, kenya. *Scientific Research*.

Ondimu, K. N. (2000). Severe anaemia during pregnancy in kisumu district of kenya: prevalence and risk factors. *PUBMED*.

Otieno, B. (2024). Prevalence and predictors of anaemia in pregnant women receiving antenatal care at kilifi county referral hospital, kenya. *Africa Health Science*.

Rahman, M. (2016). Maternal anemia and risk of adverse birth and health outcomes in low- and middle-income countries: systematic review and meta-analysis12. *Am. J. Clin. Nutr.*

Rajula, H. S. R. (2020). Comparison of conventional statistical methods with machine learning in medicine: Diagnosis, drug development, and treatment,. *PUBMED*.

Stephen, G. (2018). Anaemia in pregnancy: Prevalence, risk factors, and adverse perinatal outcomes in northern tanzania,. *PUBMED*.

Tripathy, N. (2024). Comparative analysis of diabetes prediction using machine learning and deep learning algorithms in healthcare. *IEEE Computer Society*.

UN (2015). The 17 goals | sustainable development. Accessed: 2024-09-11.

von Dadelszen, P. (2020). The precise (pregnancy care integrating translational science, everywhere) network's first protocol: deep phenotyping in three sub-saharan african countries. *Reproductive Health*.

Wemakor, A. (2018). Prevalence and determinants of anaemia in pregnant women receiving antenatal care at a tertiary referral hospital in northern ghana,. *PUBMED*.

Weng, S. F. (2020). Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PUBMED*.

WHO (2011). Who. haemoglobin concentrations for the diagnosis of anaemia and assessment of severity. vitamin and mineral nutrition information system. geneva, world health organization, 2011 (who/nmh/nhd/mnm/11.1). Accessed: 2024-09-10.

WHO (2023a). Anaemia. Accessed: 2024-09-10.

WHO (2023b). Developing guideline recommendations for tests or diagnostic tools. Accessed: 2025-05-23.

Y. Balarajan, U. R. (2011). Anaemia in low-income and middle-income countries. *The Lancet*.

Appendix A

Similarity Report

Moses Mukhanya

Moses_Mukhanya_Masters_Thesis_Report.pdf

Strathmore University (Main Account)

Document Details

Submission ID
trn:oid:::2945:275213689

Submission Date
Mar 28, 2025, 8:55 PM GMT+3

Download Date
Apr 4, 2025, 3:58 PM GMT+3

File Name
Moses_Mukhanya_Masters_Thesis_Report.pdf

File Size
1.5 MB

69 Pages

14,388 Words

80,018 Characters

turnitin Page 2 of 85 - Integrity Overview Submission ID trn:oid:::2945:275213689

24% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text

Match Groups

250 Not Cited or Quoted 22%
Matches with neither in-text citation nor quotation marks

32 Missing Quotations 2%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

16% Internet sources
13% Publications
20% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Appendix B

Ethical approval

Strathmore UNIVERSITY

19th December 2024

Mr Mukhanya Moses,
moses.mukhanya@strathmore.edu

Dear Mr Mukhanya,

RE: Use Machine Learning Approaches to Identify Risk Factors of Anaemia in Pregnancy

This is to inform you that SU-ISERC has reviewed and **approved** your above **SU-masters** proposal. Your application reference number is **SU-ISERC2477/24.** The approval period is from **19th December 2024 to 18th December 2025.**

This approval is subject to compliance with the following requirements:

i. Only approved documents including (informed consents, study instruments, MTA) will be used.
ii. All changes including (amendments, deviations, and violations) are submitted for review and approval by SU-ISERC.
iii. Death and life-threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to SU-ISERC within 72 hours of notification.
iv. Any changes anticipated or otherwise that may increase the risks or affected safety or welfare of study participants and others or affect the integrity of the research must be reported to SU-ISERC within 72 hours.
v. Clearance for the export of biological specimens must be obtained from relevant institutions.
vi. Submission of a request for renewal of approval at least 60 days prior to the expiry of the approval period. Attach a comprehensive progress report to support the renewal.
vii. Submission of an executive summary report within 90 days of completion of the study to SU-ISERC.

Before commencing your study, you will be expected to obtain a research license from National Commission for Science, Technology, and Innovation (NACOSTI) https://research-portal.nacosti.go.ke/ and obtain other clearances needed.

Yours sincerely,

Mr Ambrose Rachier,
Chairperson; SU-ISERC

Ole Sangale Rd, Madaraka Estate. PO Box 59857-00200, Nairobi, Kenya. Tel +254 (0)703 034000
Email admissions@strathmore.edu www.strathmore.edu