



**College of Business and Economics**

**African Center of Excellence in Data Science**

# Prediction of anemia and associated risk factors among under five years children in Rwanda using machine learning techniques

**By**

**Delphine ISHIMWE**

**Registration Number: 221003545**

**A dissertation submitted in partial fulfillment of the requirements for the Degree of Master of Science in Data Science**

**Supervisor: Dr. Pierre Claver RUTAYISIRE**

**May 2023**

## DECLARATION

I here declare that this dissertation, "Prediction of anemia and associated risk factors among under five years children in Rwanda using machine learning techniques," is entirely my original work, and that, to the best of my knowledge, it has not previously been submitted for the award of a degree at any other university, and that all sources of information used in the preparation of this thesis have been properly acknowledged.

**Student Name**: Delphine ISHIMWE

**Signature:** 

## APPROVAL SHEET

This dissertation entitled "**Prediction of anemia and associated risk factors among under five years children in Rwanda using machine learning techniques"** written and submitted by Delphine ISHIMWE in partial fulfillment of the requirements for the degree of Master of Science in Data Science majoring in Data Mining is hereby accepted and approved. The rate of plagiarism tested using Turnitin is 18% which is less than 20% accepted by ACE-DS.

**Dr. Pierre Claver RUTAYISIRE**

Supervisor

**Dr. Ignace KABANO**

Head of Training

## DEDICATION

To the Almighty God

To my mother, brother, and sisters

To my classmate and all lectures

## ACKNOWLEDGEMENTS

Firstly, I give thanks to The Almighty God for his provision during my studies as well as for protecting and favoring me in my life.

I want to express my appreciation to the University of Rwanda (UR), specifically the African Center of Excellence in Data Science (ACE-DS), for helping me develop my knowledge capacity, which was the cornerstone of my exceptional performance during the entire program.

Many thanks to Dr. Pierre Claver RUTAYISIRE, my thesis supervisor, for consenting to serve as my adviser on this project, for providing me with all the tools I required to finish my research project, and for his moral support.

I thank my family and friends in particular for all of their financial, emotional, and moral support.

I would like to thank the administrative team at the UR as well as the entire employees of the ACE-DS for their assistance during my time there. Godspeed to everyone.

# ABSTRACT

The health problem of anemia affects both developed and developing nations, having a significant negative impact on both social and economic development and human health. Rwanda Demographic and Health Survey (RDHS) 2014–2015 reported that 37% of children in Rwanda between the ages of 6 to 59 months are anemic.

The study aimed at identifying the factors affecting anemia in under five years Rwandan children and predicting its status using machine learning techniques.

In this research, secondary data from the RDHS 2019–20 were used. The risk variables linked to anemia in under five years children were determined with the help of bivariate and multivariate analyses. Machine learning models including K-Nearest Neighbor (KNN), Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), and Decision Tree (DT) were used to predict its status. Several evaluation metrics including Area Under the Curve (AUC), Confusion Matrix, Recall, Accuracy, F1 Score, and Precision were used to assess the performance of the created models.

According to the multivariate analysis findings, the factors for anemia in under five years Rwandan children included wealth index, child age, the mother's anemic status, and having recently had diarrhea. In predicting using ML models, the evaluation metrics used show the highest results for the Random Forest model compared to others. The results showed that Random Forest had 70% precision, 70% recall, 70% f1 score, and 70% AUC, indicating that it produced 70% accurate outcomes.

The study concludes that Machine learning is a powerful tool that is being used in health issues for predicting diseases and identifying hidden patterns which couldn't be revealed easily by other traditional methods.



**Keywords**: Anemia, under-five years children, RDHS, Machine Learning

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# LIST OF ACRONYMS AND SYMBOLS

UR: University of Rwanda

ACE-DS: African Center of Excellence in Data Science

WHO: World Health Organization.

HIV/AIDS: Human Immunodeficiency Virus/Acquired Immunodeficiency Syndrome

RDHS: Rwanda Demographic Health Survey

KNN: K-Nearest Neighbor

SVM: Support Vector Machine

KR: Kids Records (KR)

AUC: Area Under the Curve

BDHS: Bangladesh Demographic and Health Survey

LDA: Linear Discriminant Analysis

CART: Classification And Regression Trees (CART)

RF: Random Forest

LR: Logistic Regression

CBC: Complete Blood Count

ANN: Artificial Neural Network

STI: Sexually Transmitted Infections

COVID-19: Coronavirus disease

TP: True Positive

TN: True Negative

FP: False Positive

FN: False Negative

OR: Odds Ratio

CI: Confidence Interval

ROC: Receiver Operating Characteristic

ML: Machine Learning

# CHAPTER ONE. GENERAL INTRODUCTION

## 1.1 Background of Study

In line with the World Health Organization (WHO), anemia is a major problem facing people worldwide (Khan et al., 2019). It is characterized by having a low concentration of hemoglobin in the blood, which affects children's cognitive and motor development and adults' productivity at work(Anand et al., n.d.). Iron deficiency is the main factor contributing to anemia, but other factors include viral disorders like malaria, tuberculosis, HIV, and parasitic infections. The WHO states that when the hemoglobin concentration of a child is below 110 grams per liter (Hgb110g/dl), they are anemic (Kebede et al., 2021). Both rich and developing nations are affected by the global health issue of anemia, which has serious negative effects on the health of individuals. Under five years children are in a period of high demand for iron due to rapid growth, hence those from low-income families have high risks of being affected by anemia due to iron deficiency (Kebede et al., 2021).

Anemia is thought to affect 1.62 billion people worldwide, being high in developing nations (42.8%, 25.7%, and 9.1% respectively, in low, medium, and high-income countries) (Da Silva et al., 2018). In 2011, 273.2 million of under five years children were projected to have anemia, or 42.6% of this population worldwide. Preschoolers in Africa are more likely to be anemic (62.3%) than those in Southeast Asia (53.8%) or the Eastern Mediterranean (48.6%)(Da Silva et al., 2018). 269 million children, or 39.8% of under five years children had anemia in 2019. Anemia was most common in under five years children in the African area (60.2% frequency)(Observatory & Observatory, 2019).

According to RDHS 2014–2015 report, 37% of under five years children were affected by anemia, with 1% of those children having severe anemia, 15% having moderate anemia, and 21% with mild anemia. Referring to WHO guidelines, we define severe anemia as when the concentration of hemoglobin is below 7 g/dl. Anemia is considered moderate in case the level of hemoglobin is between 7.0 and 9.9 g/dl, while mild anemia is for hemoglobin levels between 10.0 and 10.9 g/dl.

Referring to the RDHS 2019–20 report, anemia affects 37% of those under five years of Rwandan children, indicating that the country's anemia prevalence has remained steady during the previous five years. Anemia prevalence increases in children living in rural regions and decreases with age, reaching 70% of children in the ages of 12 to 17 months, and 24% of children in the ages of 48 to 59 months, respectively (NISR, 2020). Anemia prevalence varies by province, with the southern province having the lowest prevalence at 32% and the west and north provinces having the highest incidence at 41%, respectively.

## 1.2 Problem Statement

Childhood anemia in Rwanda remains a public health problem, and if not treated, it affects the development of children's minds, bodies, and social skills (Kebede et al., 2021). Referring to RDHS 2014-15 and RDHS 2019-20 reports, 37% of children are anemic to some extent overall, which shows that this percentage has remained stable over the past five years. Different studies have used traditional statistical methods such as bivariate and multivariate logistic regression analysis (Jembere et al., 2020), multilevel Poisson regression models (Da Silva et al., 2018), and multilevel analysis (Kawo et al., 2018) to identify the factors affecting anemia in under five years children. While these studies help diagnose diseases, they do not handle massive amounts of data that can reveal various links and patterns that could have significant effects. (Khare et al., 2017).

To reduce anemia among under five years children in Rwanda, it is not enough to identify only the associated risk factors, it is also of interest to make predictions to determine whether anemia is likely given the circumstances, and help the medical experts in the early detection of anemia and therefore increase the survival rate of children. The ability to estimate illness status is crucial for establishing policies for community health services, and allocating the resources.

The disease status has been predicted using predictive models like K-Nearest Neighbor, Logistic Regression, Random Forest, Support Vector Machine, and Decision Tree. However, very few studies have applied machine learning in predicting anemia among under five years children in Rwanda.

Compared to traditional statistical methods, machine learning provides opportunities to automatically discover regularities from the data and help to get more accurate prediction results.

Therefore, this study is required to determine the main risks connected to childhood anemia in Rwanda and to develop predictive models using machine learning to make more precise forecasts.

The purpose of this research is to identify risk variables connected to childhood anemia and use machine learning techniques on the RDHS 2019–20 dataset to predict its status.

## 1.3Objective of Study

### 1.3.1 General Objective

Using machine learning approaches based on the RDHS 2019–20 dataset, the overall goal is to identify the risk variables and predict anemia among those under five years of Rwandan children.

### 1.3.2 Specific Objectives

- To determine the key variables influencing childhood anemia in Rwanda.
- To build different predictive models for anemia using machine learning based on RDHS 2019-20.
- To compare the results of the models for predicting anemia among children.

## 1.4Research Questions

- What are the risk variables connected to anemia in under five years children?
- What are the performances of the predictive models for anemia among under five years children?
- Which model is performing better compared to others in predicting childhood anemia?

## 1.5Justification and Significance of Study

The health of pregnant women and children under five is particularly affected by the worldwide health issue of anemia, which has serious effects on their well-being. Anemia affects children under five more severely in Rwanda than it does in pregnant women, according to the RDHS 2019–20 (37% and 25% respectively). The above facts attracted me to conduct this study.

Identifying the main causes and predicting anemia in under five years children in Rwanda will facilitate the Ministry of Health to implement intervention programs to fight against those factors and provide early treatment to children before they get sick, and this will reduce the percentage of children suffering from anemia in Rwanda. Furthermore, this study adds value to the understanding of the problem of anemia in under five years children in Rwanda and will assist the researchers in current evidence about anemia among children in Rwanda.

## 1.6 Scope and Limitation of Study

The purpose of this study is to identify the factors affecting childhood anemia in Rwanda by using the Kids Records (KR) file selected from RDHS 2019-20 dataset. After identifying the main causes, the prediction was done based on the identified factors using machine learning models. Different models were utilized including K-Nearest Neighbor, Logistic Regression, Random Forest, Support Vector Machine, and Decision Tree models. To determine which model performs better than others in predicting under five years old anemia in Rwanda, several evaluation metrics, including Area Under the Curve (AUC), Confusion Matrix, Recall, Accuracy, F1 Score, and Precision were used.

## 1.7 Organization of Study

The five chapters that make up this study. The study's background information, significance, and goal are all presented in the first chapter. Chapter two reviews the relevant studies on risk factor identification and the use of ML techniques in the prediction of childhood anemia. The third chapter presents the case study and outlines the procedures used to identify the factors affecting anemia in under five years children in Rwanda. The results from chapter three are explained in chapter four. The fifth chapter concludes with the findings and offers suggestions for the future.

# CHAPTER TWO. LITERATURE REVIEW

## 2.1 Theoretical Review

As stated by WHO, anemia occurs when there is a low concentration of hemoglobin in the body (below 110 grams per liter). Hemoglobin binds oxygen due to the iron it contains. The body's cells need oxygen carried by hemoglobin to make energy. Low hemoglobin levels influence the crucial operation of organs like the heart and lungs because fewer oxygen levels enter to sustain the operations of the body. Additionally, since hemoglobin carries oxygen throughout the body and oxygen helps the body produce energy, fatigue is one of the earliest signs of anemia (Danquah et al., 2013)

Anemia prevalence has been assessed by the WHO to determine the severity of the issue and the need for interventions. The table below lists the importance of anemia for public health based on population prevalence(De Benoist et al., 2008).

| **Presence of anemia (%)** | **Significance grouping for public health** |
|---|---|
| Less than 4.9 | No issue with public health |
| Between 5.0 and 19.9 | Minor public health issue |
| Between 20.0 and 39.9 | The public health issue of moderate severity |
| Greater than or equal to 40.0 | Serious issue with public health |

Table 1: Classification of anemia's importance to public health

Several variables, including sex, age, weight, ethnicity, household head's educational level, and region of residence were identified as factors affecting childhood anemia according to a study by (Keokenchanh et al., 2021). The younger children between the ages of 6 to 11 months were reported to be affected by anemia (72.6%) more than the older ones. Anemia was high in those who had both fever (49.4%) and diarrhea (49.2%) in the two weeks before the survey. Furthermore, there was a presence of anemia in children with illiterate, younger, and female household heads than in children with educated, older, and male household heads.

Additionally, the study on the geographical variation and factors influencing childhood anemia in Ethiopian kids between the ages of 6 to 59 months discovered a strong relationship between

anemia and the income index, religion, child age, the presence of anemia in the mother, the number of under five years children living in the home, employment position of the mother, stunting and a recent two-week fever(Jember et al., 2021). According to the researchers of that study, children older than 12 months are prone to anemia less frequently than those aged 6 to 11 months, and this is because older children are more likely to eat foods like meat, fish, chicken, and cereals that are high in iron. The prevalence of anemia was found to be unaffected by gender (Panchal et al., 2022). However, the researchers who examined the RDHS 2010 discovered that the frequency of anemia was almost 6% higher in boys with more than twice high in the youngest age group than in the oldest age group. All three factors remained significantly correlated with anemia when age group, gender, and malaria status were included in a single logistic regression model (Nkulikiyinka et al., 2015).

## 2.2 Conceptual Framework

Using the conceptual framework that the researchers created (shown in Figure 1) based on the literature, The factors that are connected to anemia in children between the ages of 6 to 59 months are in three categories: background characteristics (place of residence and region), child characteristics (birth weight, age, birth order, sex, and stunting), and mother characteristics (education level, age, employment status, marital status, and wealth status), which have an indirect impact on anemia prevalence by affecting the child's birth weight (Kasonde et al., 2021).

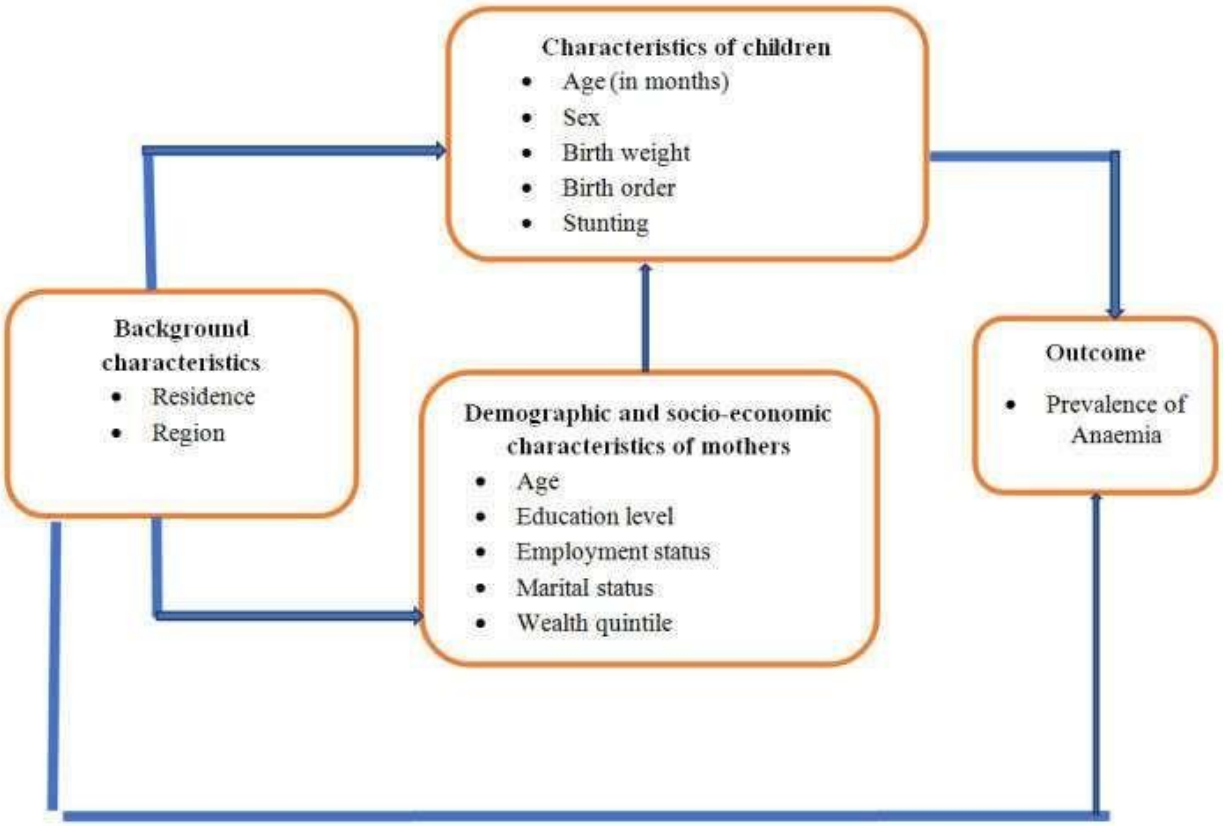



Figure 1: Anemia's conceptual foundation among children

## 2.3The factors that influence anemia in Under five years old children

A study on Anemia risk factors in under five years children in Sub-Saharan African Countries classified the determinants of anemia as Parental/household, neighborhood, or community, and child-related factors; comorbidities (including diarrhea, acute respiratory infection, and fever), maternal education, Age, sex, birth order, mother's anemia status, malnutrition or stunting, maternal age, place of residence, and household affluence were among the factors that researchers discovered (Obasohan et al., n.d.).

A previous study carried out in Guguftu, South Wollo, Northeast Ethiopia used Multivariate logistic regression in identifying the factors for anemia in under five years children and its frequency. The findings revealed that underweight children, children in the ages of 6 to 11, 12 to 23 months, lived in an urban area, had complementary foods introduced to them early (before 6 months), who had mothers with no formal education or only primary education, and who came

from a low-income family were more likely to become anemic (Gebreweld Id et al., 2019). Anemia in under-five children was found to be predicted by low birth weight, dietary variables (poor or not eating foods high in iron such as meat, fruits, and vegetables), and drinking milk in the Arusha district of Tanzania. It was discovered that raising community awareness of the benefits of feeding the baby only breast milk and introducing supplementary meals, teaching mothers and other caregivers about nutrition in general, and exploring the addition of micronutrient powder to supplementary meals to increase their nutritional value could all help to reduce childhood anemia in the Arusha district of Tanzania (Kejo et al., 2018).

Anemia in children between the ages of 6 to 59 months in Ethiopia was studied to determine its geographic pattern and contributing factors. It was discovered that anemia varied geographically depending on the region, which may have been caused by dietary preferences, the prevalence of infectious diseases, the soil's mineral composition, or access to medical facilities; the occurrence of anemia was significantly different due to both community- and individual-level variables, the likelihood of being anemic was higher in those who were younger, had fathers with little education, was from mothers with anemia, lived in the quintile with the least income, were infected, and children with growth delays. About 57 percent of this difference was caused by these community- and individual-level variables (Eshete et al., 2022).

Moreover, different predictors of Anemia in Nigerian children between the ages of 6 to 59 months have been categorized as individual (related to the child) predictors, household, and geographic predictors. The identified significant child-related determinants include age group, gender, length of nursing, deworming status, consumption of iron supplements, Malaria's concomitant conditions, acute respiratory infection, fever, and undernourishment; related to household predictors, the age of the mother, the maternal education level, and household wealth quintile were identified; the regional predictors were reportedly greater in the primarily agricultural zone than in the pastoral region. (Obasohan et al., 2022). The regional results, however, were at odds with (Mohammed et al., 2019), who claimed that children in the agricultural region have high hemoglobin levels than the pastoral region children as a result of the pastoral region's high malaria prevalence.

## 2.4 Empirical Review of Study

Linear Discriminant Analysis (LDA), Classification and Regression Trees (CART), Random Forest, K-nearest neighbors, Support Vector Machines, and logistic regression were used for predicting childhood anemia in Bangladesh Based on Bangladesh Demographic and Health Survey (BDHS) dataset. The Random Forest was identified as the best-performing algorithm than others with 68.53% accuracy, 70.73% sensitivity, 66.41% specificity, and an AUC of 0.6857(Khan et al., 2019). Decision Tree, Naive Bayes, and Random Forest algorithms were used based on complete blood count (CBC) data from pathology centers to predict anemia. The accuracy results demonstrated that the Naive Bayes method outperformed the decision tree and random forest (Jaiswal et al., 2019).

Additionally, to classify anemia and malaria using health data determinants, Artificial Neural Networks (ANN), SVM, KNN, RF, and Naive Bayes were used. When their performance was evaluated, it was found that Artificial Neural Networks predicted malaria and anemia with the highest levels of accuracy (94.74% and 84.17%) (Sow et al., 2020).

Cross-sectional research in a hospital was carried out to predict anemia in children between the ages of 6 to 36 months in Jammu. Several ML algorithms, including LDA, CART, KNN, RF, and LR, were developed. By comparing their predictions, Random Forest showed the best prediction accuracy (Gupta et al., 2020).

# CHAPTER THREE. METHODOLOGY

## 3.1Study Design

The RDHS, which is a cross-sectional, nationally representative survey, served as the foundation for the cross-sectional study design used in the current study. To identify the risk variables connected to anemia in under five years children, bivariate and multivariate analyses were used. K-Nearest Neighbor, Logistic Regression, Random Forest, Support Vector Machine, and Decision Tree were five distinct models used to make the prediction. ML methods provide advantages over statistical methods in disease prediction, as it helps to extract hidden patterns from the data without human intervention and get more accurate prediction results. Predicting anemia in under five years children will aid medical professionals in the early detection of anemia and, as a result, raise the survival rate of children.

The Area Under the Curve (AUC), Confusion Matrix, Recall, Accuracy, F1 Score, and Precision were used to determine which model is performing better than others in predicting childhood anemia in Rwanda.

## 3.2Data source

The study's data set came from the cross-sectional, nationally representative Survey (RDHS 2019-20), which was carried out in Rwanda. For the 2019–20 RDHS, data were specifically collected on breastfeeding practices, maternal and child health, adult and pediatric mortality, nutritional status of women and children, fertility, domestic violence, women's empowerment, family planning knowledge and use, knowledge and behavior regarding HIV/AIDS and other STIs, as well as other health-related issues like smoking. Anemia, malaria, HIV, and a few micronutrient markers were also examined for prevalence.

The process of collecting data took place between November 9, 2019, and July 20, 2020. However, it was suspended from March 21 to June 7, 2020, due to the statewide lockdown brought on by the coronavirus pandemic (COVID-19). 13,005 households in total, 2,913 of which are urban and 10,092 of which are rural, made up the sample. 12,951 of the 13,005 households were inhabited, and 12,949 of those households had successful interviews. In this

study, the Kids Records (KR) file from RDHS 2019-20 dataset was used, which contains information for children 6-59 months, with 1216 variables and 8092 records.

## 3.3Study Variables

The RDHS 2019-20 KR file contains 1216 variables with 8092 records. The study did not incorporate all variables. We chose the factors for the analysis based on earlier relevant studies that identified various predictors of childhood anemia. The following is a table showing the variables used in this study.

| Variable name | Counts | Percent |
|---|---|---|
| **Region** | | |
| Kigali | 377 | 11.61 |
| South | 747 | 23.01 |
| West | 821 | 25.29 |
| North | 529 | 16.3 |
| East | 772 | 23.78 |
| **Residence** | | |
| Urban | 699 | 21.53 |
| Rural | 2547 | 78.47 |
| **Highest educational level** | | |
| No primary | 349 | 10.75 |
| Primary | 2140 | 65.93 |
| Secondary | 600 | 18.48 |
| Higher | 157 | 4.84 |
| **Sex of household head** | | |
| Male | 2464 | 75.91 |
| Female | 782 | 24.09 |
| **Wealth index** | | |
| Poorest | 749 | 23.07 |
| Poorer | 664 | 20.46 |
| Middle | 617 | 19.01 |

| | | |
|---|---|---|
| Richer | 611 | 18.82 |
| Richest | 605 | 18.64 |
| **Marital status** | | |
| Never in union | 281 | 8.66 |
| Married | 1708 | 52.62 |
| Living with partner | 1010 | 31.12 |
| Widowed | 48 | 1.48 |
| Divorced | 63 | 1.94 |
| Separated | 136 | 4.19 |
| **Respondent's occupation** | | |
| Not working | 568 | 17.5 |
| Professional/technical/managerial | 108 | 3.33 |
| Clerical | 20 | 0.62 |
| Sales | | |
| Agricultural-self employed | | |
| Agricultural-employee | 1333 | 41.07 |
| Household and Domestic | 28 | 0.86 |
| Services | 343 | 10.57 |
| Skilled manual | 76 | 2.34 |
| Unskilled manual | 770 | 23.72 |
| **Sex of child** | | |
| Male | 1639 | 50.49 |
| Female | 1607 | 49.51 |
| **Anemia status of the mother** | | |
| Not anemic | 2863 | 88.2 |
| Anemic | 383 | 11.8 |
| **Had diarrhea recently** | | |
| No | 2738 | 84.35 |
| Yes, in the last 2 weeks | 508 | 15.65 |
| **Had fever in last 2 weeks** | | |
| No | 2609 | 80.38 |

| | | |
|---|---|---|
| Yes | 637 | 19.62 |
| **Child's age in months** | | |
| 6-11 | 401 | 12.35 |
| 12-23 | 736 | 22.67 |
| 24-35 | 740 | 22.8 |
| 36-47 | 695 | 21.41 |
| 48-59 | 674 | 20.76 |
| **Woman's age in years** | | |
| 15-19 | 42 | 1.29 |
| 20-24 | 422 | 13 |
| 25-29 | 732 | 22.55 |
| 30-34 | 891 | 27.45 |
| 35-39 | 740 | 22.8 |
| 40-44 | 342 | 10.54 |
| 45-49 | 77 | 2.37 |
| **Child's weight in kg** | | |
| 3-8 | 206 | 6.35 |
| 8-13 | 1688 | 52 |
| 13-18 | 1258 | 38.76 |
| 18-23 | 94 | 2.9 |
| **Birth weight in kg** | | |
| 0 - 2.5 | 201 | 6.19 |
| 2.5 - 4 | 2601 | 80.13 |
| >=4 | 444 | 13.68 |
| **Anemia status** | | |
| Not anemic | 2059 | 63.43 |
| Anemic | 1187 | 36.57 |

Table 2: Variables used in the study

## 3.4Study sample size

The KR file consists of 8092 records for children 6 to 59 months old, recorded from mothers age d 15 to 49, based on the KR file of RDHS 2019-20. After removing the missing values and outlie rs, 3246 records were used as the sample size in this study.

## 3.5Data Pre-processing

To reduce prediction errors and improve efficiency, careful cleaning of the dataset is needed. Data pre-processing is a task that includes the preparation and transformation of raw data into a suitable format for transmission to an ML model. It involves four main stages namely, data cleaning, data reduction or dimension reduction, data transformation, and data integration (Alasadi & Bhaya, 2017). Data cleaning is the first step in data preprocessing which is done to fill in the incomplete values, smooth the noisy data, eliminate the outliers, and fix the discrepancy. The KR file used in the study was not completely accurate and error-free. We performed data cleaning on the dataset for the treatment of the null values and outliers.

To improve the performance of the model, it is essential to take the most significant and vital variables and limit the number of features before developing the predictive models. By using bivariate and multivariate analysis approaches we identified the main variables contributing to anemia in under five years children. Selected attributes were then used to build ML predictive models.

## 3.6Building Predictive Models

Predicting anemia status in under five years Rwandan Children; LR, DT, RF, SVM, and KNN models have been used based on the KR file from RDHS 2019-20 dataset. To train the models and evaluate their performance, the train and test data were used. The train data count for 80% of the dataset, while the remaining 20% was the test data used for model evaluation. The above-mentioned models were chosen because they had been successfully applied in other studies of a similar nature and had produced reliable prediction outcomes. Below, we give a quick overview of the models that were applied to this investigation.

### 3.6.1 Logistic Regression

Logistic regression falls under the genre of Supervised Learning and is one of the Machine Learning algorithms that are most frequently used. It is used to predict the outcome of a categorical dependent variable using independent variables. The outcome must therefore be a discrete or categorical value such as True or False, 0 or 1, or Yes or No(Hamadi et al., 2020).

After identifying the significant variables contributing to anemia among children, Logistic Regression has been used to provide classification results for anemic and not anemic children given the factors.

### 3.6.2 Decision Tree

A Decision Tree is a supervised ML method used to solve classification and regression problems, but mostly it is used for classification-related problems. It is a tree-structured classifier, with internal nodes representing the properties of the dataset, branches representing the decision-making process, and each leaf node representing the classification outcome.

It has been used in this research to extract patterns for the prediction of pediatric anemia. It has been suggested for use in a variety of domains, including the analysis of medical diseases. We employed the Decision Tree classification technique to predict anemia status among under-five-year-old children based on several decision rules derived from the independent variables used since the target variable of the dataset used is categorial.

### 3.6.3 Random Forest

The Random Forest is a supervised ML method used for applications in classification and regression. It is an ensemble model meaning that it builds multiple Decision Trees on different subsets of the dataset and averages their predictions to make the final and accurate predictions. Since the Random Forest technique has been demonstrated to perform better than other algorithms used in previous research to predict childhood anemia, it has been used in this work to create predictions of anemia in under five years children in Rwanda using the Kids Records file of the RDHS 2019–20.

### 3.6.4 Support Vector Machine

The Support Vector Machine is an ML technique designed for classification and regression issues (Bansal et al., 2022). The SVM method seeks to provide the best hyperplane, a decision boundary that divides the data points into classes and allows us to classify new data points correctly. SVM algorithms come in two different varieties. The non-linear type SVM is used when the dataset cannot be divided into two classes by a straight line, and the linear type SVM is used when the dataset can be divided into two classes by a straight line.

In this study, the SVM algorithm was used to build classification predictions for anemia among under five years children in Rwanda.

### 3.6.5 K-Nearest Neighbor

Usually abbreviated as KNN, the K-Nearest Neighbor model is a supervised machine-learning approach used to address regression and classification problems. The method is referred to as K-Nearest Neighbor (KNN) Classification, where k nearest neighbors are utilized to determine the class because it is frequently advantageous to take into account more than one neighbor. Examples are categorized in Nearest Neighbor Classification according to the class of their closest neighbors (Cunningham & Delany, 2021).

## 3.7Model Evaluation

Using various evaluation metrics, including Area Under the Curve (AUC), Confusion Matrix, Recall, Accuracy, F1 Score, and Precision, the prediction outcomes of the aforementioned models were evaluated. The effectiveness of the trained models was assessed using these evaluation criteria, and the top model for Rwandan children's anemia prediction was chosen after that.

### 3.7.1 Confusion matrix

It is a cross-table that displays the number of actual classification values and predicted classification values (Grandini et al., 2020). The columns show the model predictions while the rows display the true classifications. The correctly classified values are located on the main diagonal of the matrix. A confusion matrix aids in determining which categories are consistent with the correct model and which are not. It has the following four fundamental properties, according to (Actual, n.d.).

**True Positive (TP):** True Positive denotes both a positive actual value and a positive model forecast.

**True Negative (TN):** The actual value and the predicted value are both negative.

**False Positive (FP):** The positive outcomes that the model incorrectly predicts as positive to the contrary they are negative

**False Negative (FN):** False negatives are the negative outcomes that the model incorrectly predicts as negative to the contrary, they are positive.

### 3.7.2 Accuracy

Accuracy is a metric for evaluating classification models (Kulkarni et al., 2020). The accuracy of the model is the proportion of correctly classified values to all values. It can be computed as follows:

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

The percentage of all samples that the model successfully classified provides the overall accuracy of the model. The issue with the accuracy metric is that it can be misleading when dealing with imbalanced datasets because the majority class will have more examples than the minority class, allowing even an incompetent model to achieve 95% accuracy. The alternative is to use recall as well as precision metrics in the case of imbalanced dataset classification. (Brownlee, n.d.). As a result, alternative measures for classifier evaluation were utilized in this work to evaluate the classifiers' performance.

### 3.7.3 Precision and Recall

Precision denotes the proportion of identifications that was correct. It enables us to see how accurately the model categorizes the positive values. The percentage of positive samples that the classifier correctly recognized as positive represents the recall also known as the Sensitivity or True Positive Rate (Martin Ward Powers, 2011).

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

### 3.7.4 F1 Score

The F1 Score is a metric used to assess the model's efficacy. It can be calculated from the confusion matrix by combining precision and recall using the Harmonic mean approach(Grandini et al., 2020).

$$F1\ Score = \frac{2(Precision * Recall)}{Precision + Recall}$$

To evaluate the models used in this study, F1 Score was among the metrics used.

# CHAPTER FOUR. RESULTS AND DISCUSSIONS

## 4.1. Introduction

The results of the bivariate and multivariate logistic regression analysis performed to identify the factors affecting anemia in Rwandan children under the age of five are presented in this chapter. The findings of K-Nearest Neighbor, Logistic Regression, Random Forest, Support Vector Machine, and Decision Tree are also covered in this chapter. Using the Area Under the Curve (AUC), Confusion Matrix, Recall, Accuracy, F1 Score, and Precision, we assessed the models' performance. The outcomes are also discussed.

## 4.2 Characteristics of the study population

This study included 3246 total children, with 2059(63.4%) of them being not anemic, and 1187(36.6%) being anemic. Most of the anemic children come from Western regions where 42% of children are anemic. By residence, children from rural regions are more anemic (38%) than those in urban areas (33%). Children from the poorest families are more anemic (42%) than the ones from the richest families (30%). Children whose mothers have some levels of anemia are more anemic (47%) than those from non-anemic mothers. Children between the ages of 6 to 11 months are more anemic than older children, and children whose mothers are aged 15-19 years are more anemic than those from older mothers. Table 3 shows further details

## 4.3. Bivariate analysis of risk factors for anemia in under five years children

The results from the bivariate analysis showed that Region ($x^2 = 28.86$, $p < 005$), residence ($x^2 = 5.36$, $p < 005$), wealth index ($x^2 = 21.45$, $p < 005$), marital status ($x^2 = 12.97$, $p < 005$), respondent's occupation ($x^2 = 16.28$, $p < 005$), woman's age ($x^2 = 24.39$, $p < 005$), child weight ($x^2 = 136.03$, $p < 005$), child age ($x^2 = 263.11$, $p < 005$), anemia status of the mother ($x^2 = 20.87$, $p < 005$), had diarrhea recently ($x^2 = 29.05$, $p < 005$), and had fever in last two weeks ($x^2 = 20.68$, $p < 005$) were strongly correlated with childhood anemia in Rwanda. Table 3 presents the results.

| Variables | Anemia Status | | P-value |
|---|---|---|---|
| | Anemic N (%) | Not anemic N (%) | |
| **Region** | | | |
| Kigali | 131(35.0%) | 246(65.0%) | 2.11e-05 |
| South | 239(32.0%) | 508(68.0%) | |
| West | 348(42.0%) | 473(58.0%) | |
| North | 214(40.0%) | 315(60.0%) | |
| East | 255(33.0%) | 517(67.0%) | |
| **Residence** | | | |
| Urban | 229(33.0%) | 470(67.0%) | 0.02 |
| Rural | 958(38.0%) | 1589(62.0%) | |
| **Highest educational level** | | | |
| No primary | 141(40.0%) | 208(60.0%) | 0.168 |
| Primary | 773(36.0% | 1367(36.0%) | |
| Secondary | 225(38.0%) | 375(62.0%) | |
| Higher | 48(31.0%) | 109(69.0%) | |
| **Sex of household head** | | | |
| Male | 915(37.0%) | 1549(63.0%) | 0.25 |
| Female | 272(35.0%) | 510(65.0%) | |
| **Wealth index** | | | |
| Poorest | 315(42.0%) | 434(58.0%) | 0.0002 |
| Poorer | 248(37.0%) | 416(63.0%) | |
| Middle | 231(37.0%) | 386(63.0%) | |
| Richer | 209(34.0%) | 402(66.0%) | |
| Richest | 184(30.0%) | 421(70.0%) | |
| **Marital status** | | | |
| Never in union | 107(38.0%) | 174(62.0%) | 0.023 |

| | | | |
|---|---|---|---|
| Married | 605(35.0%) | 1103(65.0%) | |
| Living with partner | 402(40.0%) | 608(60.0%) | |
| Widowed | 16(33.0%) | 32(67.0%) | |
| Divorced | 22(35.0%) | 41(65.0%) | |
| Separated | 35(26.0%) | 101(74.0%) | |
| **Respondent's occupation** | | | |
| Not working | 234(41.0%) | 334(59.0%) | 0.022 |
| Professional/technical/managerial | 26(24.0%) | 82(76.0%) | |
| Clerical | 8(40.0%) | 12(60.0%) | |
| Sales | | | |
| Agricultural-self employed | | | |
| Agricultural-employee | 483(36.0%) | 850(64.0%) | |
| Household and Domestic | 7(25.0%) | 21(75.0%) | |
| Services | 119(35.0%) | 224(65.0%) | |
| Skilled manual | 23(30.0%) | 53(70.0%) | |
| Unskilled manual | 287(37.0%) | 483(63.0%) | |
| **Sex of child** | | | |
| Male | 620(38.0%) | 1019(62.0%) | 0.14 |
| Female | 567(35.0%) | 1040(65.0%) | |
| **Anemia status of the mother** | | | |
| Not anemic | 1006(35.0%) | 1857(65.0%) | 4.9e-06 |
| Anemic | 181(47.0%) | 202(53.0%) | |
| **Had diarrhea recently** | | | |
| No | 947(35.0%) | 1791(65.0%) | 7.05e-8 |
| Yes, in the last 2 weeks | 240(47.0%) | 268(53.0%) | |
| **Had fever in last 2 weeks** | | | |
| No | 904(35.0%) | 1705(65.0%) | 5.42e-6 |
| Yes | 283(44.0%) | 354(56.0%) | |
| **Child's age in months** | | | |

| | | | |
|---|---|---|---|
| 6-11 | 270(67.0%) | 131(33.0%) | 9.7e-56 |
| 12-23 | 331(45.0%) | 405(55.0%) | |
| 24-35 | 229(31.0%) | 511(69.0%) | |
| 36-47 | 194(28.0%) | 501(72.0%) | |
| 48-59 | 163(24.0%) | 511(76.0%) | |
| **Woman's age in years** | | | |
| 15-19 | 25(60.0%) | 17(40.0%) | 0.0004 |
| 20-24 | 182(43.0%) | 240(57.0%) | |
| 25-29 | 275(38.0%) | 457(62.0%) | |
| 30-34 | 313(35.0%) | 578(65.0%) | |
| 35-39 | 261(35.0%) | 479(65.0%) | |
| 40-44 | 109(32.0%) | 233(68.0%) | |
| 45-49 | 22(29.0%) | 55(71.0%) | |
| **Child's weight in kg** | | | |
| 3-8 | 128(62.0%) | 78(38.0%) | 2.7e-29 |
| 8-13 | 701(42.0%) | 987(58.0%) | |
| 13-18 | 337(27.0%) | 921(73.0%) | |
| 18-23 | 21(22.0%) | 73(78.0%) | |
| **Birth weight in kg** | | | |
| 0 - 2.5 | 84(42.0%) | 117(58.0%) | 0.26 |
| 2.5 - 4 | 946(36.0%) | 1655(64.0%) | |
| >=4 | 157(35.0%) | 287(65.0%) | |

Table 3: Bivariate analysis of risk factors of anemia among under five years children

## 4.4 Multivariate analysis of risk factors for anemia in under five years children

Using a multivariate analysis, the risk factors for anemia in children under the age of five years were identified. The results showed that wealth index [OR=0.89, 95% CI (0.84-0.94), p<0.05], c hild age [OR=0.65, 95% CI (0.61-0.69), p<0.05], anemia status of the mother [OR=1.77, 95% CI (1.41-2.21), p<0.05], and had diarrhea recently before the survey [OR=1.14, 95% CI (1.03-1.26), p<0.05] were the identified factors. Children from the richest family [OR=0.59, 95% CI (0.43-0. 8), p<0.05] had fewer odds to be anemic than those from the poorest family. A child's likelihood of being anemic decreases with age. The older children in the ages of 48 to 59 months [OR=0.13, 95% CI (0.08-0.19), p<0.05] are less likely to be anemic than younger children in the ages of 6 to 11 months. Children whose mothers are anemic [OR=1.82, 95% CI (1.45-2.3), p<0.05] had 82% more risks of being anemic than those whose mothers are not anemic. Children who had diarrhea within the two weeks before the survey [OR=1.23, 95% CI (0.99-1.5), p<0.05] had more risks of being anemic than those who did not have diarrhea. Table 4 illustrates the results

| Variables | Odds ratio | 95% CI | | P-Value |
|---|---|---|---|---|
| | | Lower | Upper | |
| **Region** | | | | |
| Kigali | 1 | | | |
| South | 0.72 | 0.52 | 0.98 | 0.039 |
| West | 1.27 | 0.94 | 1.7 | 0.119 |
| North | 1.21 | 0.87 | 1.67 | 0.248 |
| East | 0.86 | 0.63 | 1.17 | 0.341 |
| **Residence** | | | | |
| Urban | 1 | | | |
| Rural | 1.1 | 0.88 | 1.4 | 0.378 |
| **Wealth index** | | | | |
| Poorest | 1 | | | |

| | | | | |
|---|---|---|---|---|
| Poorer | 0.73 | 0.58 | 0.92 | 0.010 |
| Middle | 0.75 | 0.59 | 0.95 | 0.021 |
| Richer | 0.65 | 0.50 | 0.84 | 0.001 |
| Richest | 0.59 | 0.43 | 0.8 | 0.001 |
| **Marital status** | | | | |
| Never in union | 1 | | | |
| Married | 1.12 | 0.82 | 1.53 | 0.482 |
| Living with partner | 1.21 | 0.89 | 1.64 | 0.208 |
| Widowed | 0.9 | 0.44 | 1.86 | 0.792 |
| Divorced | 1.13 | 0.61 | 2.07 | 0.690 |
| Separated | 0.66 | 0.4 | 1.07 | 0.095 |
| **Respondent's occupation** | | | | |
| Not working | 1 | | | |
| Professional/technical/managerial | 0.56 | 0.34 | 0.95 | 0.031 |
| Clerical | 1.11 | 0.41 | 3.02 | 0.832 |
| Sales | | | | |
| Agricultural-self employed | | | | |
| Agricultural-employee | 0.83 | 0.66 | 1.04 | 0.106 |
| Household and Domestic | 0.57 | 0.23 | 1.45 | 0.243 |
| Services | 0.95 | 0.7 | 1.29 | 0.770 |
| Skilled manual | 0.64 | 0.37 | 1.12 | 0.124 |
| Unskilled manual | 0.78 | 0.61 | 1.01 | 0.065 |
| **Anemia status of the mother** | | | | |
| Not anemic | 1 | | | |
| Anemic | 1.82 | 1.45 | 2.3 | 0.000 |
| **Had diarrhea recently** | | | | |
| No | 1 | | | |
| Yes, in the last 2 weeks | 1.23 | 0.99 | 1.5 | 0.05 |
| **Had fever in last 2 weeks** | | | | |

| | | | | |
|---|---|---|---|---|
| No | 1 | | | |
| Yes | 1.18 | 0.97 | 1.44 | 0.089 |
| **Child's age in months** | | | | |
| 6-11 | 1 | | | |
| 12-23 | 0.35 | 0.26 | 0.47 | 0.000 |
| 24-35 | 0.18 | 0.13 | 0.25 | 0.000 |
| 36-47 | 0.16 | 0.11 | 0.23 | 0.000 |
| 48-59 | 0.13 | 0.087 | 0.19 | 0.000 |
| **Woman's age in years** | | | | |
| 15-19 | 1 | | | |
| 20-24 | 0.56 | 0.28 | 1.14 | 0.114 |
| 25-29 | 0.52 | 0.26 | 1.05 | 0.069 |
| 30-34 | 0.51 | 0.25 | 1.02 | 0.059 |
| 35-39 | 0.53 | 0.26 | 1.08 | 0.084 |
| 40-44 | 0.47 | 0.23 | 0.99 | 0.047 |
| 45-49 | 0.39 | 0.16 | 0.93 | 0.035 |
| **Child's weight in kg** | | | | |
| 3-8 | 1 | | | |
| 8-13 | 1.27 | 0.87 | 1.85 | 0.201 |
| 13-18 | 1.37 | 0.88 | 2.15 | 0.162 |
| 18-23 | 1.39 | 0.7 | 2.76 | 0.342 |

Table 4: Multivariate analysis of risk factors of anemia among under five years children

## 4.5 Predicting Anemia in under five years Children in Rwanda

Before building predictive models, the feature selection technique was used to identify the most relevant variables linked to anemia in children under the age of five years children. The identified features were used in predicting anemia among under five years children in Rwanda. We have used five machine learning models, and six evaluation metrics to compare their performance. Below are presented the results obtained from the classifiers.

### 4.5.1 Logistic Regression

The results show that Logistic regression has an accuracy of 63% which means that 63% of the total predictions are accurate. Precision and recall are 64% and 63% respectively. It implies that 64% of children who were predicted anemic were anemic. Below are further results presented in a table.

| Model | Accuracy | Precision | Recall | F1 score | AUC |
|---|---|---|---|---|---|
| **Logistic Regression** | 0.63 | 0.64 | 0.63 | 0.63 | 0.63 |

The ROC curve has also been used to help us visualize how well our machine-learning classifiers are performing. For logistic regression, The AUC-ROC curve shows that Logistic regression can distinguish anemic and not anemic classes at 63 %. The graph is presented below.

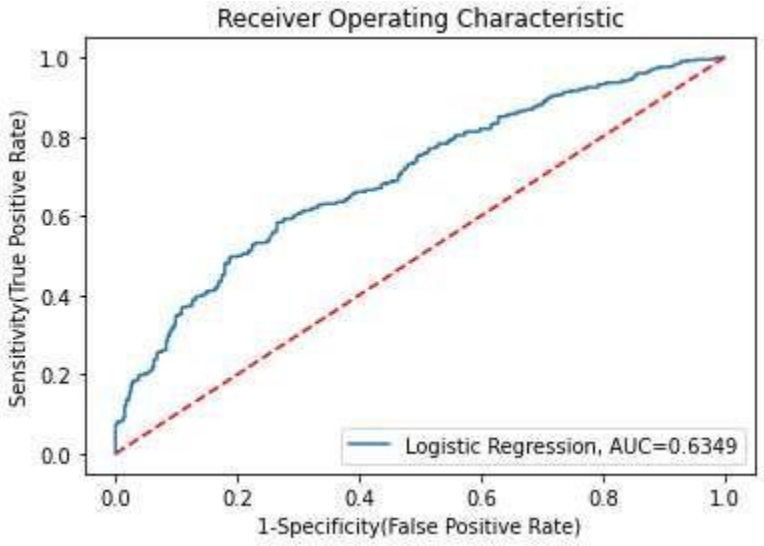



Figure 2: Logistic Regression ROC Curve

### 4.5.2. Decision Tree

The Decision tree is 64% accurate in its prediction results, 64% of predicted anemic children are anemic, and the model is wrongly predicting 133 children as not anemic while they have some levels of anemia. The results are shown below

| Model | Accuracy | Precision | Recall | F1 score | AUC |
|---|---|---|---|---|---|
| **Decision Tree** | 0.64 | 0.64 | 0.64 | 0.64 | 0.64 |

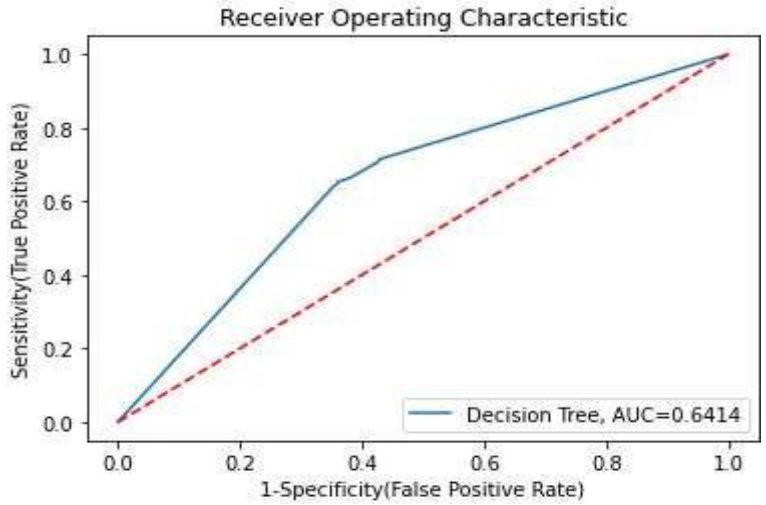



Figure 3: Decision Tree ROC Curve

## 4.5.3 Random Forest

The prediction results of the random forest model are 70% accurate. Precision and recall are 70% and 70% respectively. meaning that 70% of predicted anemic children were anemic.

| Model | Accuracy | Precision | Recall | F1 score | AUC |
|---|---|---|---|---|---|
| **Random Forest** | 0.70 | 0.70 | 0.70 | 0.70 | 0.70 |

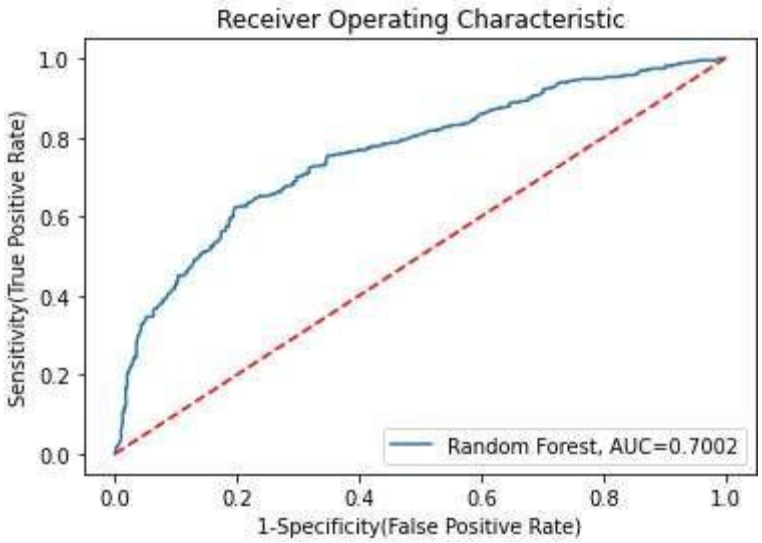



Figure 4: Random Forest ROC Curve

### 4.5.4 K Nearest Neighbor

The KNN model is 69% accurate in the prediction results as it is shown by its accuracy score. The precision score is 79% signifies that 70% of predicted anemic children were anemic.

| Model | Accuracy | Precision | Recall | F1 score | AUC |
|---|---|---|---|---|---|
| **KNN** | 0.69 | 0.70 | 0.70 | 0.70 | 0.69 |

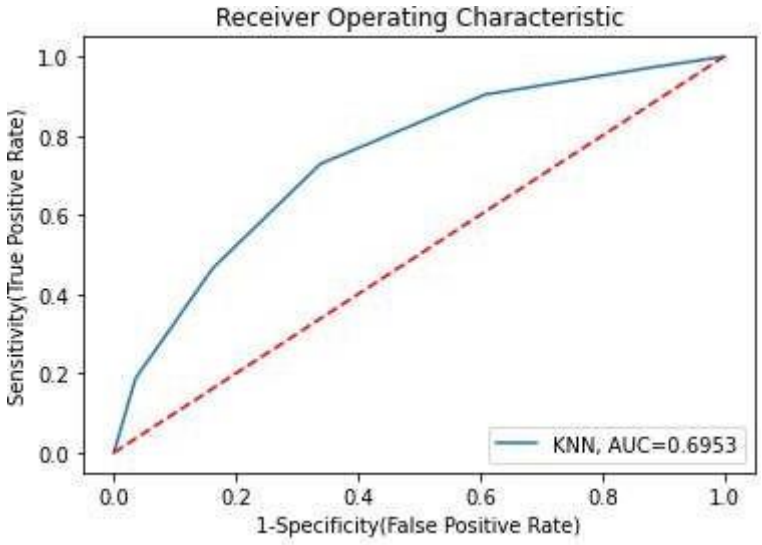



Figure 5: KNN ROC Curve

### 4.5.5 Support Vector Machine

The results from the analysis show that the SVM model's accuracy is 64%, and the precision and recall scores are all 65% respectively. In general, it suggests that the SVM model's prediction results are 65% accurate.

| Model | Accuracy | Precision | Recall | F1 score | AUC |
|---|---|---|---|---|---|
| **SVM** | 0.64 | 0.65 | 0.65 | 0.65 | 0.647 |

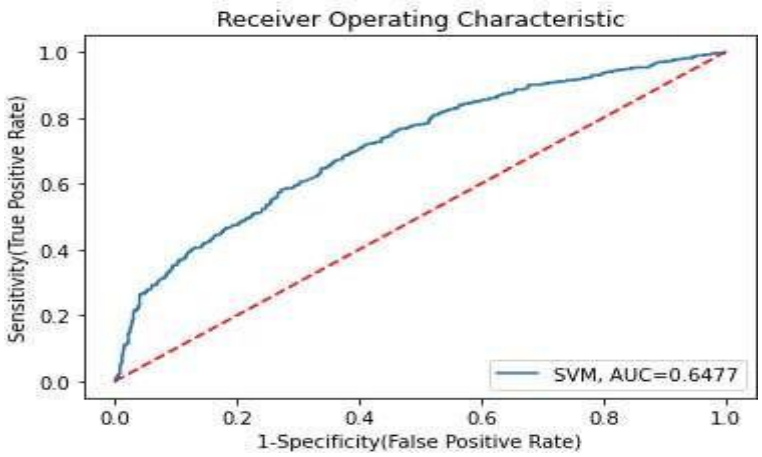
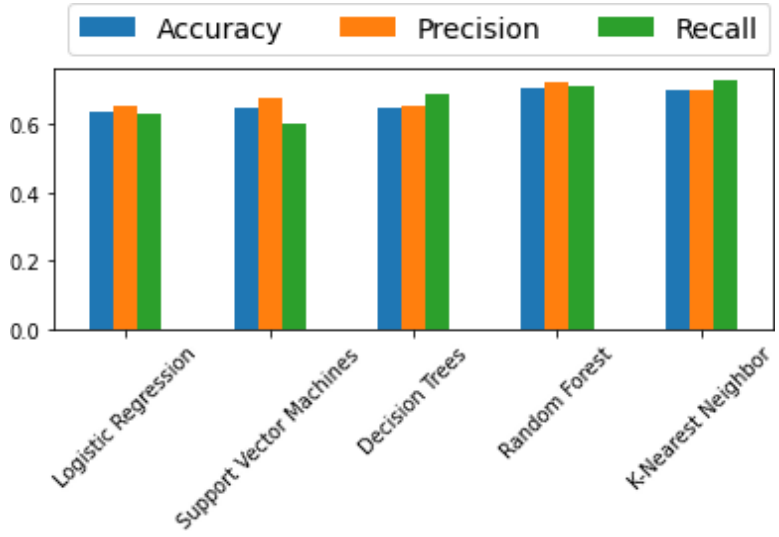



Figure 6: SVM ROC Curve

Here is a quick bar chart and the ROC Curves for all the classifiers showing how good the classifiers are in their predictions.



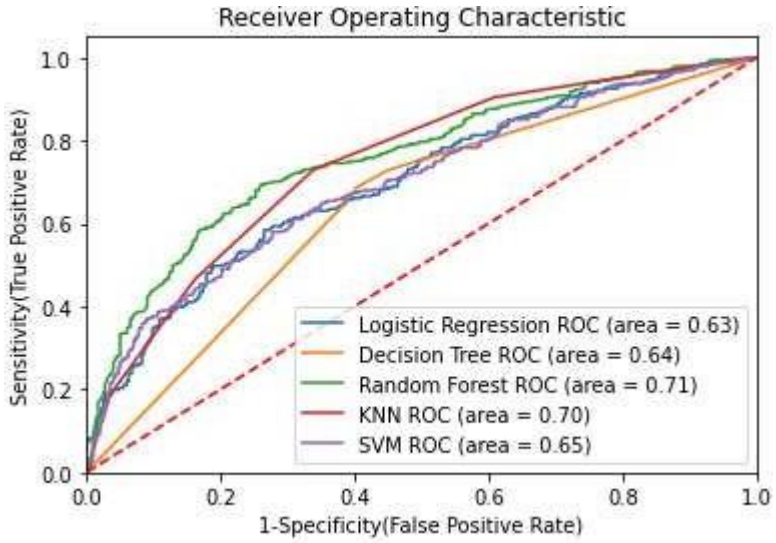



Figure 7: ROC Curve for the ML models used

The predictive models above were trained on 80% of the dataset and tested on the remaining 20% of the dataset. For the test data used, the confusion matrix results show that logistic regression correctly classifies 523 children while 301 are the wrong predictions.

With the Decision Tree classifier, the correct predictions are 531 children, and the wrongly classified children are 293. The Random Forest predicted correctly 577 children; however, it wrongly predicted 119 children as anemic while they are not anemic, and 128 children predicted as not having anemia while they are anemic. The KNN model correctly predicted negative 262 children, 312 children correctly predicted positive, while 134 and 116 children wrongly predicted positive and negative concerning the confusion matrix results. The Support Vector Machine correctly predicted 532 children, but 292 children are wrongly predicted. Below is a table showing the evaluation metrics results for all the ML models used.

| **Evaluation metrics** | **ML models** | | | | |
|---|---|---|---|---|---|
| | **LR** | **DT** | **RF** | **KNN** | **SVM** |
| Accuracy | 0.63 | 0.64 | 0.70 | 0.69 | 0.64 |
| Precision | 0.64 | 0.64 | 0.70 | 0.70 | 0.65 |
| Recall | 0.63 | 0.64 | 0.70 | 0.70 | 0.65 |
| F1 score | 0.63 | 0.64 | 0.70 | 0.70 | 0.65 |
| AUC | 0.63 | 0.64 | 0.70 | 0.69 | 0.647 |

Table 5: Evaluation metrics results for the ML models used

## 4.6 Discussion of the Findings

The purpose of this study is to identify the factors affecting anemia in Rwandan children below the age of five. After identifying such factors, the study also predicts anemia status using ML models. Referring to the results of the multivariate analysis wealth index, child age, the mother's anemia status, and having recently experienced diarrhea were the factors affecting anemia in Rwandan children under the age of five. These findings were in line with those of the earlier studies(Eshete et al., 2022; Obasohan et al., 2022).

The evaluation metrics used reveal that the Random Forest model outperforms the other ML models for prediction. For the 20% test data used, the Random Forest model correctly predicted negative 277 children, and 300 children correctly predicted positive, however, 119 and 128 children wrongly predicted positive and negative. It was found that Random Forest provided 70% accurate results with a precision of 70%, recall of 70%, f1 score of 70%, and AUC of 70%. These results collaborated with the previous studies(Gupta et al., 2020; Khan et al., 2019). The results from predictive models show that Logistic Regression achieved an accuracy of 63% with 64% precision, 63% recall, and 63% f1 score and AUC. It provided 253 and 270 children as the correct predictions for not anemic and anemic children. It has wrongly predicted 143 children as anemic while they are not, and 158 children as not anemic while they are anemic. Decision Tree provided the accuracy of 64%, 64% precision, and 64% of recall, f1 score, and AUC. Those results lead to the correct predictions of 236, not anemic children, and 295 were predicted correctly as anemic children. The results also show that it has wrongly predicted 133 anemic children as not having anemia, and 160 children were wrongly predicted as anemic while they don't have anemia. The Random Forest predicted correctly 577 children; however, it wrongly predicted 119 children as anemic while they are not anemic, and 128 children predicted as not having anemia while they are anemic. The KNN model correctly predicted negative 262 children, 312 children correctly predicted positive, while 134 and 116 children wrongly predicted positive and negative concerning the confusion matrix results. The Support Vector Machine correctly predicted 532 children, but 292 children are wrongly predicted. By analyzing the effectiveness of the predictive ML models using the assessment metrics, it was discovered that the Random Forest offered the greatest scores when compared to others used in this study.

# CHAPTER FIVE. CONCLUSION AND RECOMMENDATIONS

## 5.1 Conclusion

The purpose of this study is to identify the factors affecting anemia in Rwandan children below the age of five and predicts its status using ML techniques. Such factors were identified using bivariate and multivariate logistic regression. The findings revealed that the anemia status of the mother, wealth index, child age, and having recently experienced diarrhea before the survey were the risk variables linked to anemia in under five years children in Rwanda.

After identifying the risk factors, K-Nearest Neighbor, Logistic Regression, Random Forest, Support Vector Machine, and Decision Tree were used for prediction. We have trained the predictive models on 80% of the dataset and tested them on 20% of the dataset. Different evaluation metrics including Area Under the Curve (AUC), Confusion Matrix, Recall, Accuracy, F1 Score, and Precision have been used to evaluate the performance of the model and choose the best in predicting anemia in under five years children in Rwanda. it was discovered that the Random Forest offered the greatest scores when compared to other models that have been used in this research in predicting anemia in children below the age of five in Rwanda using the RDHS 2019/2020 KR dataset.

## 5.2 Recommendation and further work

The findings of this research will help academics in understanding the problem of anemia in young children in Rwanda by providing them with up-to-date data. To combat the identified risk factors, the Ministry of Health can use the study's findings to establish intervention programs. Random Forest identified as the best classifier will be used to predict anemia among under five years children in Rwanda which will facilitate providing early treatment to children before they get sick, and this will reduce the percentage of children suffering from anemia in Rwanda.

The dataset used in this study is RDHS 2019-20 KR file which contains information for under five years children. I would suggest other researchers use different RDHS data taken at different times to see how anemia is affecting under five years children in Rwanda over time.

As this study used only K-Nearest Neighbor, Logistic Regression, Random Forest, Support Vector Machine, and Decision Tree, I am suggesting other researchers use more ensemble ML models like XGBoost and GradientBoosting because it was found that ensemble models provide good results than a single model. It will be also important to focus on time series machine learning models to analyze anemia among children in Rwanda over time.

# REFERENCES


Actual. (n.d.). *Confusion Matrix for Multi-Class Classification*. https://www.analyticsvidhya.com/blog/2021/06/confusion-matrix-for-multi-class-classification/2/9

Alasadi, S. A., & Bhaya, W. S. (2017). Review of data preprocessing techniques in data mining. *Journal of Engineering and Applied Sciences*, *12*(16), 4102–4107. https://doi.org/10.3923/jeasci.2017.4102.4107

Anand, P., Gupta, R., & Sharma, A. (n.d.). *Prediction of Anaemia among children using Machine Learning Algorithms*. 469–480.

Bansal, M., Goyal, A., & Choudhary, A. (2022). A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning. *Decision Analytics Journal*, *3*, 100071. https://doi.org/10.1016/J.DAJOUR.2022.100071

Brownlee, J. (n.d.). *How to Calculate Precision, Recall, and F-Measure for Imbalanced Classification https://machinelearningmastery.com/precision-recall-and-f-measure-for-imbalanced-classification/#:~:text=Precision quantifies the number of and rec… 1/33 ⊃⊃ Navigation Navigation How to Calculate Precision, Recall, and F-Measure for Imbalanced Classification Start Machine Learning*. https://machinelearningmastery.com/precision-recall-and-f-measure-for-imbalanced-classification/#:~:text=Precision

Cunningham, P., & Delany, S. J. (2021). K-Nearest Neighbour Classifiers-A Tutorial. In *ACM Computing Surveys* (Vol. 54, Issue 6). Association for Computing Machinery. https://doi.org/10.1145/3459665

Da Silva, L. L. S., Fawzi, W. W., & Cardoso, M. A. (2018). Factors associated with anemia in young children in Brazil. *PLoS ONE*, *13*(9), 1–12. https://doi.org/10.1371/journal.pone.0204504

Danquah, I., Gahutu, J.-B., Zeile, I., Musemakweri, A., & Mockenhaupt, F. P. (2013). *Anaemia, iron deficiency, and a common polymorphism of iron-regulation, TMPRSS6 rs855791, in Rwandan children*. https://doi.org/10.1111/tmi.12216

De Benoist, Bruno., World Health Organization., & Centers for Disease Control and Prevention (U.S.). (2008). *Worldwide prevalence of anemia 1993-2005 of WHO Global Database of anemia*. World Health Organization.

Eshete, T., Anteneh, K., Getahun, F., Marie, T., Sahile, E., Kidane, R., & Bazezaw, Y. (2022). Geographical pattern and associated factors of anemia among children aged 6–59 months in Ethiopia: Further analysis of Ethiopian demographic and health survey 2016. *International Journal of Africa Nursing Sciences*, *16*. https://doi.org/10.1016/j.ijans.2022.100420

Grandini, M., Bagli, E., & Visani, G. (2020). *Metrics for Multi-Class Classification: an Overview*. http://arxiv.org/abs/2008.05756

Gupta, R., Anand, P., & Sharma, A. (2020). *Prediction of Anaemia among children using Machine Learning Algorithms*. *June*.

Hamadi, A. H., Mohammed, Y. A., & Saleh, G. (2020). Investigating the Applicability of Logistic Regression and Artificial Neural Networks in Predicting Breast Cancer. *Journal of Al-Qadisiyah for Computer Science and Mathematics*, *12*(2), 63. https://doi.org/10.29304/jqcm.2020.12.2.697

Jaiswal, M., Srivastava, A., & Siddiqui, T. J. (2019). Machine learning algorithms for anemia disease prediction. *Lecture Notes in Electrical Engineering*, *524*, 463–469. https://doi.org/10.1007/978-981-13-2685-1_44

Jember, T. A., Teshome, D. F., Gezie, L. D., & Agegnehu, C. D. (2021). Spatial variation and determinants of childhood anemia among children aged 6 to 59 months in Ethiopia: further analysis of Ethiopian demographic and health survey 2016. *BMC Pediatrics*, *21*(1). https://doi.org/10.1186/s12887-021-02901-y

Jembere, M., Kabthymer, R. H., & Deribew, A. (2020). Determinants of Anemia Among Children Aged 6 to 59 Months in Dilla Town, Southern Ethiopia: A Facility Based Case-Control Study. *Global Pediatric Health*, *7*, 1–22. https://doi.org/10.1177/2333794X20974232

Kasonde, M. E., Banda, C., & Phiri, M. (2021). DETERMINANTS OF PREVALENCE OF ANAEMIA: MULTIVARIATE ANALYSIS OF 2018 DHS Monitoring and Evaluation Capacity and System Strengthening in Zambia View project Population Studies View project DETERMINANTS OF PREVALENCE OF ANAEMIA MULTIVARIATE ANALYSIS OF 2018 DHS A R T I C L E I N F O. *Article in International Journal of Current Advanced Research*. https://doi.org/10.24327/ijcar.2021.23752.4708

Kawo, K. N., Asfaw, Z. G., & Yohannes, N. (2018). Multilevel Analysis of Determinants of Anemia Prevalence among Children Aged 6-59 Months in Ethiopia: Classical and Bayesian Approaches. *Anemia*, *2018*. https://doi.org/10.1155/2018/3087354

Kebede, D., Getaneh, F., Endalamaw, K., Belay, T., & Fenta, A. (2021). Prevalence of anemia and its associated factors among under-five age children in Shanan gibe hospital, Southwest Ethiopia. *BMC Pediatrics*, 1–9. https://doi.org/10.1186/s12887-021-03011-5

Kejo, D., Petrucka, P. M., Martin, H., Kimanya, M. E., & Mosha, T. C. (2018). Prevalence and predictors of anemia among children under 5 years of age in Arusha District, Tanzania. *Pediatric Health, Medicine and Therapeutics*, *9*. https://doi.org/10.2147/PHMT.S148515

Keokenchanh, S., Kounnavong, S., Midorikawa, K., Ikeda, W., Morita, A., Kitajima, T., & Sokejima, S. (2021). Prevalence of anemia and its associated factors among children aged 6-59 months in the Lao People's Democratic Republic: A multilevel analysis. *PLoS ONE*, *16*(3 March), 6–13. https://doi.org/10.1371/journal.pone.0248969

Khan, J. R., Chowdhury, S., Islam, H., Raheem, E., & Management, H. (2019). *MACHINE LEARNING ALGORITHMS TO PREDICT THE CHILDHOOD*. *17*(1), 195–218. https://doi.org/10.6339/JDS.201901

Khare, S., Kavyashree, S., Gupta, D., & Jyotishi, A. (2017). Investigation of Nutritional Status of Children Based on Machine Learning Techniques Using Indian Demographic and Health Survey Data. *Procedia Computer Science*, *115*, 338–349. https://doi.org/10.1016/J.PROCS.2017.09.087

Kulkarni, A., Chong, D., & Batarseh, F. A. (2020). Foundations of data imbalance and solutions for a data democracy. *Data Democracy: At the Nexus of Artificial Intelligence, Software Development, and Knowledge Engineering*, 83–106. https://doi.org/10.1016/B978-0-12-818366-3.00005-8

Martin Ward Powers, D. (2011). *EVALUATION: FROM PRECISION, RECALL, AND F-MEASURE TO ROC, INFORMEDNESS, MARKEDNESS & CORRELATION*. *2*(1), 37–63. https://doi.org/10.9735/2229-3981

Mohammed, S. H., Habtewold, T. D., & Esmaillzadeh, A. (2019). Household, maternal, and child related determinants of hemoglobin levels of Ethiopian children: Hierarchical regression analysis. *BMC Pediatrics*, *19*(1). https://doi.org/10.1186/s12887-019-1476-9

NISR. (2020). *[National Institute of Statistics of Rwanda] Rwanda Demographic and Health Survey 2019-20 Key Indicators Report.* https://www.statistics.gov.rw/publication/demographic-and-health-survey-20192020-key-indicators

Nkulikiyinka, R., Binagwaho, A., & Palmer, K. (2015). The changing importance of key factors associated with anemia in 6- to 59-month-old children in a sub-Saharan African setting where malaria is on the decline: Analysis of the Rwanda Demographic and Health Survey 2010. *Tropical Medicine and International Health*, *20*(12), 1722–1732. https://doi.org/10.1111/tmi.12604

Obasohan, P. E., Walters, S. J., Jacques, R., & Khatab, K. (n.d.). *A Scoping Review of the Risk Factors Associated with Anaemia among Children Under Five Years in Sub-Saharan African Countries*. https://doi.org/10.3390/ijerph17238829

Obasohan, P. E., Walters, S. J., Jacques, R., & Khatab, K. (2022). Individual, household, and area predictors of anemia among children aged 6–59 months in Nigeria. *Public Health in Practice*, *3*, 100229. https://doi.org/10.1016/J.PUHIP.2022.100229

Observatory, G. H., & Observatory, G. H. (2019). *Anemia in women and children*. 1–7.

Panchal, S. S., Mishra, U., Kothari, C., Kothari, V., Dalai, S., Mecwan, M., Chaudhary, S., Sharma, J., & Shah, P. (2022). Prevalence of anemia in pre-school tribal children concerning parasitic infections and nutritional impact. *Journal of Taibah University Medical Sciences*. https://doi.org/10.1016/j.jtumed.2022.05.002

Sow, B., Mukhtar, H., Ahmad, H. F., & Suguri, H. (2020). Assessing the relative importance of social determinants of health in malaria and anemia classification based on machine learning techniques. *Informatics for Health and Social Care*, *45*(3), 229–241. https://doi.org/10.1080/17538157.2019.1582056

# Dissertation final

ORIGINALITY REPORT

| 18% | 16% | 11% | % |
|---|---|---|---|
| SIMILARITY INDEX | INTERNET SOURCES | PUBLICATIONS | STUDENT PAPERS |

PRIMARY SOURCES

| | | |
|---|---|---|
| 1 | dr.ur.ac.rw<br>Internet Source | 4% |
| 2 | www.researchgate.net<br>Internet Source | 1% |
| 3 | www.frontiersin.org<br>Internet Source | 1% |
| 4 | "Applications of Artificial Intelligence and Machine Learning", Springer Science and Business Media LLC, 2022<br>Publication | 1% |
| 5 | www.studymode.com<br>Internet Source | 1% |
| 6 | repository.smuc.edu.et<br>Internet Source | 1% |
| 7 | scholarworks.waldenu.edu<br>Internet Source | 1% |
| 8 | link.springer.com<br>Internet Source | <1% |
| | impactfactor.org | |

| | | |
|---|---|---|
| 9 | Internet Source | <1% |
| 10 | www.coursehero.com<br>Internet Source | <1% |
| 11 | hdl.handle.net<br>Internet Source | <1% |
| 12 | assets.researchsquare.com<br>Internet Source | <1% |
| 13 | eprints.whiterose.ac.uk<br>Internet Source | <1% |
| 14 | pubmed.ncbi.nlm.nih.gov<br>Internet Source | <1% |
| 15 | bmcpublichealth.biomedcentral.com<br>Internet Source | <1% |
| 16 | www.mdpi.com<br>Internet Source | <1% |
| 17 | Akhil Kadiyala. "Application of MATLAB to select an optimum performing genetic algorithm for predicting in-vehicle pollutant concentrations", Environmental Progress & Sustainable Energy, 12/2010<br>Publication | <1% |
| 18 | Dagnew Getnet Adugna, Anteneh Ayelign Kibret, Hailu Aragie, Engidaw Fentahun Enyew et al. "Prevalence and determinants of | <1% |

anemia among children aged from 6 to 59 months in Liberia: a multilevel analysis of the 2019/20 Liberia demographic and health survey data", Frontiers in Pediatrics, 2023
Publication

| | | |
|---|---|---|
| 19 | cris.maastrichtuniversity.nl<br>Internet Source | <1% |
| 20 | topsecretapiaccess.dovepress.com<br>Internet Source | <1% |
| 21 | www.minervamedica.it<br>Internet Source | <1% |
| 22 | erepository.mkuit.ac.rw<br>Internet Source | <1% |
| 23 | dspace.nm-aist.ac.tz<br>Internet Source | <1% |
| 24 | repository.ju.edu.et<br>Internet Source | <1% |
| 25 | doctorpenguin.com<br>Internet Source | <1% |
| 26 | www.dovepress.com<br>Internet Source | <1% |
| 27 | bmcwomenshealth.biomedcentral.com<br>Internet Source | <1% |
| 28 | etd.aau.edu.et<br>Internet Source | <1% |

| | | |
|---|---|---|
| 29 | bmcinfectdis.biomedcentral.com<br>Internet Source | <1% |
| 30 | academic.oup.com<br>Internet Source | <1% |
| 31 | www.lse.ac.uk<br>Internet Source | <1% |
| 32 | "Recent Trends in Communication, Computing, and Electronics", Springer Nature, 2019<br>Publication | <1% |
| 33 | erl.ucc.edu.gh:8080<br>Internet Source | <1% |
| 34 | Md. Mohsan Khudri, Kang Keun Rhee, Mohammad Shabbir Hasan, Karar Zunaid Ahsan. "Predicting nutritional status for women of childbearing age from their economic, health, and demographic features: A supervised machine learning approach", PLOS ONE, 2023<br>Publication | <1% |
| 35 | jscholaronline.org<br>Internet Source | <1% |
| 36 | udsspace.uds.edu.gh<br>Internet Source | <1% |
| 37 | Fekadu Yihunie, Eman Abdelfattah, Amish Regmi. "Applying Machine Learning to | <1% |

Anomaly-Based Intrusion Detection Systems", 2019 IEEE Long Island Systems, Applications and Technology Conference (LISAT), 2019
Publication

---

38 Misganaw Gebrie Worku, Tesfa Sewunet Alamneh, Achamyeleh Birhanu Teshale, Yigizie Yeshaw et al. "Multilevel analysis of determinants of anemia among young women (15-24) in sub-Sahara Africa", PLOS ONE, 2022 <1%
Publication

---

39 Mohammad Shahin, F. Frank Chen, Ali Hosseinzadeh, Neda Zand. "Using Machine Learning and Deep Learning Algorithms for Downtime Minimization in Manufacturing Systems: An Early Failure Detection Diagnostic Service", Research Square Platform LLC, 2023 <1%
Publication

---

40 Richard Nkulikiyinka, Agnes Binagwaho, Katie Palmer. "The changing importance of key factors associated with anaemia in 6- to 59-month-old children in a sub-Saharan African setting where malaria is on the decline: analysis of the Rwanda Demographic and Health Survey 2010", Tropical Medicine & International Health, 2015 <1%
Publication

---

| | | |
|---|---|---|
| 41 | dokumen.pub<br>Internet Source | <1% |
| 42 | open.uct.ac.za<br>Internet Source | <1% |
| 43 | umu.diva-portal.org<br>Internet Source | <1% |
| 44 | www.researchsquare.com<br>Internet Source | <1% |
| 45 | Malti Bansal, Apoorva Goyal, Apoorva Choudhary. "A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning", Decision Analytics Journal, 2022<br>Publication | <1% |
| 46 | Peter Brooker. "Air Traffic Management accident risk. Part 2: Repairing the deficiencies of ESARR4", Safety Science, 2006<br>Publication | <1% |
| 47 | fau.digital.flvc.org<br>Internet Source | <1% |
| 48 | library.kisiiuniversity.ac.ke:8080<br>Internet Source | <1% |
| 49 | ouci.dntb.gov.ua<br>Internet Source | <1% |

| | | |
|---|---|---|
| 50 | repository.out.ac.tz<br>Internet Source | <1% |
| 51 | tigerprints.clemson.edu<br>Internet Source | <1% |
| 52 | www.scirp.org<br>Internet Source | <1% |
| 53 | "Soft Computing: Theories and Applications", Springer Science and Business Media LLC, 2022<br>Publication | <1% |
| 54 | Chia-Lun Yeh, Bo-Rong Chen, Ling-Yi Tseng, Ping Jao, Tung-Hung Su, Pai-Chi Li. "Shear-wave elasticity imaging of a liver fibrosis mouse model using high-frequency ultrasound", IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control, 2015<br>Publication | <1% |
| 55 | Francis Kipkogei, Ignace H. Kabano, Belle Fille Murorunkwere, Nzabanita Joseph. "Business success prediction in Rwanda: a comparison of tree-based models and logistic regression classifiers", SN Business & Economics, 2021<br>Publication | <1% |
| 56 | Nebyu Daniel Amaha. "What factors are associated with the increase of anemia in Ethiopian children aged 6 to 59 months | <1% |

between 2011 and 2016?", BMC Nutrition, 2020
Publication

| | | |
|---|---|---|
| 57 | Stefano Lise. "Prediction of hot spot residues at protein-protein interfaces by combining machine learning and energy-based methods", BMC Bioinformatics, 2009<br>Publication | <1% |
| 58 | Tesfa Sewunet Alamneh, Alemakef Wagnew Melesse, Kassahun Alemu Gelaye. "Determinants of anemia severity levels among children aged 6-59months in Ethiopia: Multilevel Bayesian statistical approach", Scientific Reports, 2023<br>Publication | <1% |
| 59 | acikerisim.karabuk.edu.tr:8080<br>Internet Source | <1% |
| 60 | ageconsearch.umn.edu<br>Internet Source | <1% |
| 61 | arno.uvt.nl<br>Internet Source | <1% |
| 62 | dspace.unza.zm<br>Internet Source | <1% |
| 63 | journals.plos.org<br>Internet Source | <1% |
| 64 | repository.lib.tottori-u.ac.jp<br>Internet Source | |

| | | |
|---|---|---|
| | | <1% |
| 65 | repository.nusamandiri.ac.id<br>Internet Source | <1% |
| 66 | shaper-api.kadist.org<br>Internet Source | <1% |
| 67 | www.ijikm.org<br>Internet Source | <1% |
| 68 | Abebew Aklog Asmare, Yitateku Adugna Agmas. "Bivariate binary logistic regression analysis on undernutrition and anemia among under-five children in Rwanda; evidence from 2019/20 Rwanda demographic health survey", Research Square Platform LLC, 2022<br>Publication | <1% |
| 69 | Destaw Kebede, Fantahun Getaneh, Kirubel Endalamaw, Tariku Belay, Abebe Fenta. "Prevalence of anemia and its associated factors among under-five age children in Shanan gibe hospital, Southwest Ethiopia", BMC Pediatrics, 2021<br>Publication | <1% |
| 70 | Kindie Fentahun Muchie. "Determinants of severity levels of anemia among children aged 6-59 months in Ethiopia: further | <1% |

analysis of the 2011 Ethiopian demographic and health survey", BMC Nutrition, 2016

Publication

| | | | |
|---|---|---|---|
| Exclude quotes | Off | Exclude matches | Off |
| Exclude bibliography | On | | |