UNIVERSITY of RWANDA

COLLEGE OF SCIENCE AND TECHNOLOGY
SCHOOL OF SCIENCES
DEPARTMENT OF MATHEMATICS

PREDICTION OF UNDER-5 CHILDREN MORTALITY IN RWANDA USING MACHINE LEARNING TECHNIQUES

By KANANI Papias

Reg. Number: 221003983

A dissertation Submitted in partial fulfillment of the requirements for the degree of master of science in applied mathematics

Option: Statistical modeling

Supervisor: Innocent NGARUYE, Ph.D.

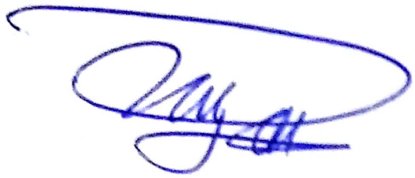
18/7/2024

Declaration

With this certification, I declare that the work included in this dissertation, “PREDICTION OF UNDER-5 CHILDREN MORTALITY IN RWANDA USING MACHINE LEARNING TECHNIQUES”, is entirely my own. To the best of my ability, I do reaffirm that the report is wholly original with no submissions to academic award-granting organizations.

Printed Name: Papias KANANI

Signature :



Date: 18^{th} July, 2024

Approval

This dissertation entitled "PREDICTION OF UNDER-5 CHILDREN MORTALITY IN RWANDA USING MACHINE LEARNING TECHNIQUES" has been written and submitted by KANANI Papias in partial fulfilment of the requirements for the degree of Master of Science in Applied Mathematics (Statistical Modeling). The report has been accepted and approved for submission at University of Rwanda

Supervisor: Innocent NGARUYE, Ph.D.

Signature:

Date: 18^{th} July, 2024

HoD

Signature:

Date: ..../... /2024

Abstract

Controlling under-5 children mortality is crucial as it impacts the future of country's economic growth worldwide. Under support of the World Health Organization (WHO), regular Demographic Health Surveys (DHS) are conducted in several countries to gather socio-demographic data, including mortality rates. In Rwanda, the Demographic and Health Survey is conducted every five years. The mortality rate for under-5 children remains a concern in Rwanda, for instance, the Rwandan Demographic and Health Survey (RDHS-2019/2020) indicated that the infant mortality rate stood at 33 deaths per 1,000 live births, under-5 children mortality at 45 deaths per 1,000 live births, and child mortality at 13 deaths per 1,000 live births. A deep analysis of this report however, portrays that the datasets being extensive, they focused on newborn mortality rather than under-5 children mortality which is a high concern for the present study. Again, the survey used traditional analysis techniques which really manifested incompetent to predict mortality rates accurately. The present study having used machine learning methods, it ensures more conventional and accurate methods in forecasting under-5 children mortality in Rwanda. In this study, thirteen independent variables related to children mortality (Highest Educational level, Births in Last Five Years, Exposure of the mother, Currently Breastfeeding, Number of Living children, Wealth Index combined, Total children Ever Born, Desire for More children, Sex of child, Birth Order number, Number of Antenatal Visits during pregnancy, Had Diarrhea recently, Source of Drinking Water) and the dependent variable "child is Alive", were considered. Leveraging Machine Learning, analysis of this large dataset, containing both numerical and categorical data, was conducted using Python 3.12 and various packages such as Pandas, Matplotlib, and NumPy. Through training and testing the under-5 children dataset, it was discovered that among seven machine learning algorithms used simultaneously, Random Forest was the best predictor, outperforming logistic regression with an average accuracy of 98.2%. Except for Naïve Bayes, all classifiers used scored greater than 95%, indicating their suitability for predicting under-5 children mortality. Features such as the number of living children in the home, source of drinking water, and number of antenatal visits were identified as important predictors of under-5 children mortality. Health care providers should pay attention to these features to forecast the lives of children under-5 years old. Additionally, SHapley Additive exPlanation (SHAP) values revealed that breastfeeding, number of living children in a family, and birth order were significant factors in predicting under-5 children mortality in Rwanda.

Contents

List of Figures

Acronyms

WHO : World Health Organization

RDHS-2019/2020 : Rwanda Demographic and Health Survey 2019/2020

SDGS : Sustainable Development Goals

ACEDs : African Centre of Excellence in Data Science

ML : Machine Learning

MDGS : Millenium Development Goals

WASH : Water, Sanitation, and Hygiene

SHAP : Shapley Additive exPlanation

LIME : Local Interpretable Model agnostic Explanations

KNN : k^{th} Nearest Neighborhood

SVM : Support Vector Machine

T P : True Positive

F N : False Negative

ROC : Receiver Operating Curves

AUC : Area Under the Curve

LDA : Linear Discriminant Analysis

DT : Decision Tree

LR : Logistic Regression

NB : Naïve Bayes

RF : Random Forest

HE : High Education (S108)

NLC : Number of Living children (V218)

EXP : Exposure

TCEB : Total Number of children Ever Born (V201)

BO : Birth Order (BORD)

SDW : Source of Drinking Water (V113)

DMC : Desire for More hildren (V605)

WI : Worth Index (V190)

BLFY : Birth in the Last Five Years (V208)

CBF : Currently Breast Feeding (V404)

SEX : Sex of a child (B4)

NAV : Number of Antenatal Visits (M14)

List of Tables

Acknowledgment

The preparation of this study is due to the assistance and contributions of a few people. First and foremost, I would like to sincerely thank my supervisor, Dr. Innocent NGARUYE, for his important and ongoing counsel and direction throughout the entire preparation of my dissertation. His zeal and unwavering pursuit of perfection have had a significant influence on my research techniques. I owe a great deal of gratitude to the University of Rwanda's administration and instructors, who gave me all the opportunities I needed to enjoy the Masters′ program and gain the knowledge and skills I need to advance my profession. I am also grateful to my family for their unwavering love and support over the years I have studied, and most of all, I am grateful to God Almighty for giving me the strength, health, and unwavering love that have always encouraged and supported me in my pursuit of this course.

Dedication

I dedicate this report to my wife, my children, and my beloved sisters and brothers,classmates and workmates in appreciation of their unfailing love, support, and care throughout my life and all of my endeavors to be successful. This book is also dedicated to my supervisor, to whom I would like to express my gratitude for his encouragement and assistance throughout my academic pursuits. Furthermore, may the Almighty you reward my close friends and fellow students who toiled diligently to produce the best, to whom I dedicate this effort. Finally, may the Almighty God bless my lecturers, who also made a significant contribution to my accomplishment, I dedicate this book to them.

Chapter 1

Introduction

1.1 Background

Children mortality remains a global concern, with approximately 5 million children under the age of 5 (under-5) dying each year, mainly from preventable causes. Despite a global decline in under-5 children mortality rates, sub-Saharan Africa still faces the highest rates, with 74 deaths per 1000 live births, significantly higher than in Europe and North America [Organization, 2015]. In Rwanda, recent data from the Rwandan Demographic and Health Survey (RDHS-2019/2020) indicates that the infant mortality rate stood at 33 deaths per 1,000 live births, under-5 children mortality at 45 deaths per 1,000 live births, and children mortality at 13 deaths per 1,000 live births [NISR and ICF, 2021].

The leading causes of under-5 children mortality include preterm birth complications, birth asphyxia/trauma, pneumonia, diarrhea, and malaria, all preventable or treatable with access to affordable interventions in health and sanitation [Organization, 2015]. In low-income countries, the under-5 children mortality rate was 69 deaths per 1000 live births in 2017. Given these challenges, and especially the fact that addressing under-5 children mortality aligns with the Sustainable Development Goals (SDGs), particularly goal 3 which aims to protect life and promote the well-being globally, it is very crucial to identify best predicting machine learning models and predict risk factors that influence under-5 children mortality in Rwanda.

1.2 Problem statement

Under-5 children mortality remains a concern globally. With a particular focus in Rwanda, where previous studies have explored various risk factors associated with infant mortality in Rwanda, highlighting the importance of addressing high fertility rates, short inter-pregnancy intervals, and other socio-economic and demographic factors. However, predicting under-5 children mortality accurately remains a challenge. The advent of Artificial Intelligence (AI) and Machine Learning (ML) presents an opportunity to enhance predictive

capabilities in health sciences by leveraging large datasets. Yet, there is a gap in research specifically focusing on under-5 children mortality prediction using ML techniques in Rwanda.

1.3 Objective of the Study

1.3.1 General Objective

This study aims to train, evaluate, and select the best machine learning classification model for predicting under-5 children mortality and predicting the most influential risk factors in Rwanda.

1.3.2 Specific Objectives

Here our specific objectives are stated.

1. Identify from the RDHS-2019/2020 data significant risk factors associated with under-5 children mortality in Rwanda using determinants from literature review.
2. Evaluate the performance of various machine learning algorithms, compare their outcomes, and select the most accurate predictor.
3. Rank the risk factors by importance in predicting under-5 children mortality using the selected best machine learning algorithm.

1.4 Significance of the Study

This study contributes to under-5 child healthcare by applying machine learning to extract hidden patterns from data, thus improving decision-making and policy formulation. The findings will aid health administrators in enhancing service quality and inform national healthcare policies, especially policies by the Ministry of Health (MoH). Additionally, the study serves as a benchmark for future research on the subject of under-5 children mortality, guiding interventions to minimize identified risk factors and improve child healthcare services.

1.5 Scope and Limitation of the Study

The study utilizes statistical analysis and machine learning techniques on secondary data, namely the RDHS-2019/2020 data obtained from the National Institute for Statistics of Rwanda (NISR) to predict under-5 children mortality in Rwanda. It focuses on socio-economic and demographic factors to develop predictive models. Limitations include challenges in investigating associations between variables and the need for techniques to explore these relationships effectively.

1.6 Research Questions

1. What major risk factors, as determined by the RDHS-2019/2020 data, are connected to under-5 child mortality in Rwanda, and how do these factors differ from those documented in the body of current literature?

2. With RDHS-2019/2020 data, how well do several machine learning algorithms predict the death of children under five? Which method shows the best accuracy and dependability in this prediction?

3. What is the best way to rank the predictive importance of the risk factors linked to child mortality among children under five years old using the most accurate machine learning algorithm?

1.7 Organization of the study

The study follows a structured framework, beginning with data collection and exploratory analysis, followed by data preparation and pre-processing to enhance modeling. Machine learning models are then fitted, and their performance evaluated using specific metrics. The study is organized into five chapters, evolving around the introduction, literature review, methodology, data analysis, and conclusion and recommendations respectively. The chart in Figure 1.7-1 represent the deployment of the ML models.

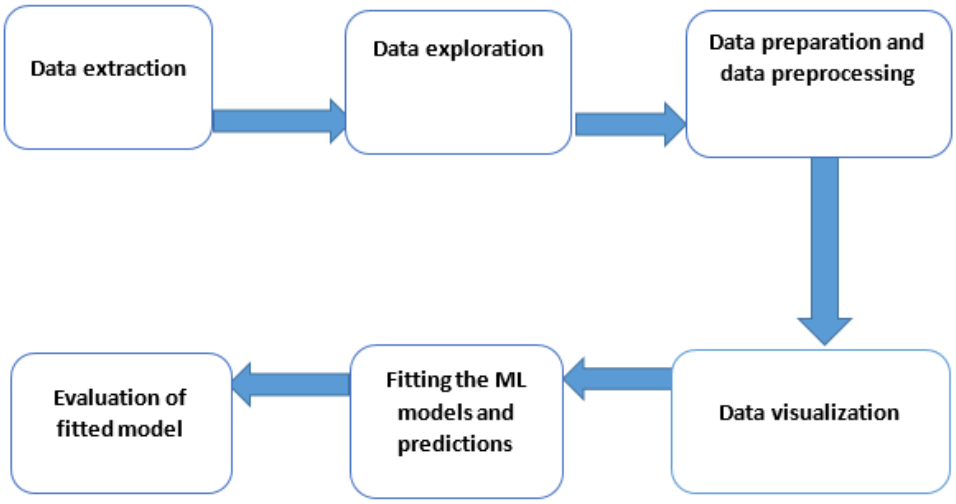


Figure 1.7-1: Deployment of ML algorithms flow chart

Chapter 2

Literature review

2.1 Overview

The literature reviewed focuses on under-5 children mortality and machine learning techniques applied to mortality prediction, drawing evidence from various sources. The review provides insights into under-5 children mortality globally and specifically in Rwanda, as well as descriptions of selected machine learning approaches and techniques used in prediction. It generally underlines the complex interplay of factors predicting under-5 children mortality, the significance of machine learning techniques in prediction, and the need for targeted interventions to address risk factors. By leveraging data-driven approaches and understanding the multifaceted nature of under-5 children mortality in Rwanda, stakeholders can formulate effective strategies to reduce child deaths and enhance overall child well-being, both in Rwanda and globally.

2.2 Under-5 children mortality

Under-5 children mortality, as defined by RDHS-2019/2020, encompasses the likelihood of death between birth and the fifth birthday, including infant mortality (death before one year) and children mortality. According to World Health Organization (WHO), in 2014, up to 5.9 million children died before their fifth birthday, and 2.7 million, that is 4.5% of the babies die every year in their first month of life (WHO, 2015).

The death of a child can leave bereaved parents and relatives in a difficult situation. Since many child fatalities went unreported, it was hard to trace every case, which further delayed the process of making swift adjustments. Promising is that global efforts have led to substantial progress in reducing child deaths since 1990; and as a matter of fact, reduction of children mortality has ranked a priority in the United Nations' Millennium Development Goals (MDGs), and again succeeded by Sustainable Development Goal (SDG-3) number 3, which seeks to ensure healthy lives and promote the well-being of all at all ages by 2030 [Mfateneza et al., 2022].

2.3 Risk factors influencing Under-5 children mortality in Rwanda

Several scholars, including JA [2020], Gupta et al. [2018], and others, have looked into the risk factors impacting under-5 children mortality in Rwanda and in most other African countries. These last, found that the risk factors contributing to under-5 children mortality include infectious diseases, malnutrition, highest educational level of parents, exposure of the mother, currently breastfeeding, wealth index of the family combined, desire for more children, sex of child, Number of antenatal visits during pregnancy, had diarrhea recently, underrated maternal education, gender inequality, immunization coverage, poor water, sanitation, hygiene provisions, total children ever born, birth order number, maternal and newborn health issues, limited healthcare access, natural disasters, source of drinking water, births in last five years, number of living children, and poverty in general [JA, 2020]. Addressing these interconnected risk factors requires comprehensive and multi-sectoral approaches. While Rwanda has made significant strides in reducing under-5 children mortality through healthcare improvements, nutrition programs, immunization campaigns, and poverty reduction initiatives, sustained efforts are essential for further progress [Saroj et al., 2022].

2.4 Machine learning techniques applied to under-5 children mortality prediction

Machine learning techniques have become more and more popular in the last several years as a way to analyze big datasets and identify patterns that traditional techniques would overlook. Using advanced machine learning algorithms, researchers such as Madakkatel et al. [2021], and others discovered that under-5 children mortality could be predicted using techniques like; Logistic Regression, KNN classifier, Decision Trees classifier, Support Vector Machines, Random Forest, Linear Discriminant Analysis, and Naïve Bayes.

Machine learning involves a series of steps to develop and deploy models that can make predictions or provide insights from data. Its process generally involves defining the problem, gathering and preprocessing data, selecting a machine learning algorithm, splitting the data, training the model, validating and tuning the model, evaluating the model, deploying the model, monitoring and maintaining the model, and finally iterating and improving the model [Madakkatel et al., 2021].

In order to predict Ethiopia's under-5 children mortality determinants, for example, the 2016 Ethiopian Demographic and Health Survey used the majority of those machine learning models. The results showed that the Random Forest model performed the best, with a prediction ability of 67.2%. It emphasized certain elements as significant determinants, like household size Bitew et al. [2020]. That same year, the most effective machine learning model, with an accuracy of 95.29% to 95.96%, was found by comparing data from the National Family Health Survey of Uttar Pradesh in India. In particular, the number of living children identified

to predict under-5 children mortality was found to be associated with the neural network model.
A recent study conducted in Rwanda and published in 2022 applied machine learning techniques to forecast infant mortality using the dataset from the Rwanda Demographic and Health Survey 2014/2015. As demonstrated by the accuracy of 84.3%, recall of 91.3%, precision of 80.3%, F1 score of 85.5%, and AUROC of 84.2%, the Random Forest model proved to be the most successful method. The study came to the conclusion that machine learning techniques could be a potent tool for predicting infant mortality because they could identify some hidden information that standard statistical methods would miss [Mfateneza et al., 2022].
These papers show how machine learning models can be used to solve important public health challenges such as death rates among children under-5.

It would be preferable in Rwanda to look beyond infant mortality and investigate the predictability of mortality for children under-5 using machine learning techniques in the context of the RDHS-2019/2020 dataset. Contemporary, up to seven learning machine technics namely Logistic regression, KNN classifier, Decision trees classifier, Support Vector machines, Random Forest, Linear discriminant analysis, and Naïve Bayes (NB), would be widely used in the estimation of mortality [Kananura, 2022].

2.4.1 Logistic regression

The literature review on the use of logistic regression algorithms as a machine learning technique in predicting the death of children under-5 provides a comprehensive analysis of the use and effectiveness of this algorithm in healthcare research.

According to Rymarczyk et al. [2019], the logistic model, while widely employed to address classification problems, can also be applied to predict under-5 children mortality. The result falls between 0 and 1 and is derived from the logistic function, which is used to forecast an event's likelihood.
In order to model the probability of a binary outcome based on a linear combination of predictors, one can use the log-odds form of logistic regression represented by Equation 2.1:

$$ln(\frac{P(Y_i = 1|X_i)}{P(Y_i = 0|X_i)}) = ln(\frac{P(Y_i = 1|X_i)}{1 - P(Y_i = 1|X_i)}) = a_0 + a_1X_{i1} + a_2X_{i2} +, ..., +a_kX_{ik} \tag{2.1}$$

where $i = 1, 2, 3, ..., n$ and $a_0, a_1, a_2, ..., a_k$ are the regression coefficients associated to predictors X_i, and Y_i denotes the binary outcome of variable which is either 0 or 1, in the above Equation 2.1, $P(Y_i = 1|X_i)$ stands for the probability that the dependent variable Y equals 1 given the independent variables X for the i^{th} observation. This ML algorithm has been used in predicting viral load non suppression among people living with HIV in Uganda. This last were then deployed, assessed for accuracy, F1 score, and area under the curve (AUC), which varied between 0.95 and 0.98, suggesting that the classifier had the ability to detect suppression of viral load [Rymarczyk et al., 2019].

2.4.2 Linear Discriminant Analysis

In healthcare research, some studies investigated the efficacy and methodology of linear discriminant analysis (LDA) to predict the death of children under-5 children mortality. The application of LDA, a statistical method for classification problems, is increasingly being employed to identify factors influencing child survival outcomes. To assess huge, intricate datasets that include demographic information, they employ LDA. As supervised learning technique applied to classification tasks, its purpose is to determine which linear combination of features best distinguishes between two or more classes of objects or occurrences.

According to Tharwat et al. [2017], LDA seeks to maintain as much class discriminating information as possible while reducing the dimensionality of the feature space. In LDA, in order to achieve this, onto a subspace of lower dimensions while maximizing the distance between classes. The LDA mathematical model involves finding a linear combination of features that best separate different classes into data.

With this study, given the dataset (x_i, y_i), $i = 1, 2, 3, ..., N$ also x_i represents the i^{th} feature vector and y_i represents the corresponding class label. Thus, the mathematical model of LDA is the projection:

$$y = W^T x, \tag{2.2}$$

where W represents the transformation matrix of LDA, $y = (y_1, y_2, y_3, ...y_N)^T$, and $x = (x_1, x_2, x_3, ..., x_N)$. Although this strategy was producing decent results, it was not ranking highly. It would be best to investigate it further to determine its true prevalence.

2.4.3 Random Forest

The Random Forest method strategy to investigate.It has been demonstrated by researchers to be a good predictor of death for Rwandan children under the age of five. This is due to its capacity to handle intricate datasets and capture complex interactions among different predictors. So, according to Kananura [2022], when integrating numerous decision trees to improve forecast accuracy, Random Forest presents a potential strategy in a nation where lowering children mortality remains a major developmental priority. Random Forest model randomly drafts N training subsets $M = \{M_1, ..., M_N\}$ also known as bootstrapping ensemble that fits multiple models on different subsets of a training dataset used to create training samples from the original training data with replacement called bootstraps.

Depending on features, Random Forest was used in so many studies to create required training samples using the following formula:

$$P = \left(1 - \frac{1}{N}\right)^N \tag{2.3}$$

which is the probability at N^{th} training samples. With the above formula, each decision tree in the ensemble made a prediction of the target variable, and the final output was based on the majority with N decision trees

voting from all the trees.

Also, it examined Gini index which was used for Classification And Regression Trees (CARTS) to select tree nodes useful also for decision tree as follows:

$$G(M) = \sum_{k=1}^{N} P_k(1 - P_k) = 1 - \sum_{k=1}^{k} P_k^2 \tag{2.4}$$

The Gini impurity at node M is represented by $G(M)$, and the proportion of class N observations at note M is represented by P_k. Finding the decision rule that causes splits to increase impurity is a recursive procedure that is performed until all nodes are pure or the cut-off is reached.

Random Forest was in so many studies classified as the best predictor of the features which are presenting huge datasets. In Bangladesh, for instance, SA and MI [2021] argued that the under-5 malnutrition rate is predicted using a Random Forest model and classification tree. The Random Forest predictive model has an accuracy of 70.1% and 72.4%, and an area under the receiver operating characteristic curve of 69.8% and 70% for underweight and stunting, respectively. Also for infant mortality in Rwanda, Mfateneza et al. [2022] found that the RF model was the best predictive model of infant mortality with accuracy 84.3%, recall 91.3%, precision 80.3%, F1 score 85.5%, and AUROC 84.2% which showed that Random Forest could be taken into consideration for this study.

2.4.4 Decision tree classifiers

Beside random foerest some researcher like Charbuty and Abdulazeez [2021] found that among ML aligorithms, decision trees was also important algorithm to predict under-5 children mortality. Decision trees iteratively partition the data into subsets that best distinguish between high and low mortality risk groups, allowing them to identify key predictors and their interactions. These models can provide useful information to healthcare providers and policymakers in Rwanda by highlighting the primary causes of death for children under-5 and suggesting targeted interventions to improve child health outcomes for different population segments and geographic regions. with this machine learning classification model, collection of splitting rules used to divide the dataset into segments based on the target variable is represented by a tree, as a tree represents the set of splitting rules used to segment the dataset based on the target variable [Charbuty and Abdulazeez, 2021].

Since most of studies required this kind of the splitting of rules, the Decision Trees Classification Algorithm was used with the following formula that describes the relationship between the outcome y which is either 0 or 1 and feature instance x:

$$\hat{y} = \hat{f}(x) = \sum_{m=1}^{M} C_m I_{\{x \in R_m\}}, \tag{2.5}$$

where C_m stands for node outcome at each instance which falls into exactly one leaf node equal subset R_m and $I_{\{x \in R_m\}}$ is the identity function that returns 1 if x is the subset of R_m and 0 otherwise. If an instance falls into a leaf node R_l the predicted outcome is $\hat{y} = c_l$ where c_l is the average of all training instances in leaf node R_l so the prediction of an individual instance is the target outcome plus the sum of all contributions of the splits that occur between the root node and the terminal node where the instance ends up.
This method was used by Mfateneza et al. [2022], and found that decision tree model successfully predicted infant mortality in Rwanda with accuracy 83%, recall 91%, precision 79%, F1 score 84.67% and AUROC 82.9%.

2.4.5 K-Nearest Neighbors classifier (KNN)

According to Zhang et al. [2017], the KNN classifier uses the majority class among its closest neighbors to classify data points, according to the similarity principle. In order to estimate the chance of death in the particular scenario of under-5 children mortality prediction, KNN can examine a number of variables, including socioeconomic indicators, healthcare accessibility, and demographic traits. The KNN classifier uses "feature similarity" to produce an outcome of a new data point, and that new data point is given an outcome based on how closely it resembles the data in the training set. Since this study required finding the Euclidean distance K between points A and B in a feature space, the algorithm was used to find the nearest K distance. Assuming that feature vectors represent A and B, the model used for this machine learning algorithm is:

$$dist(A,B) = \sqrt{\frac{\sum_{i=1}^{m}(x_i - y_i)^2}{m}}, \tag{2.6}$$

where $A = (x_1, x_2, ..., x_m)$ and $B = (y_1, y_2, ..., y_m)$, such that x_i is the feature component of A, y_i is the feature component of B and m is the feature space's dimensionality. The findings of Shichao Zhang's research demonstrated that the prior KNN approaches were unfeasible in actual applications since they set every test data point to the identical k value [Zhang et al., 2017].

2.4.6 Naïve Bayes

Researchers in this field also stated that the Naïve Bayes classifier can probabilistically predict the likelihood of under-5 children mortality in Rwanda based on the presence or absence of particular traits by leveraging historical data on various socio-economic, demographic, and health-related factors. So, according to Greiner et al. [Technical report TR03-09, Department of Computing Science, University of Alberta, Canada, 2003], the Naïve Bayes classifier's mathematical model applies the Bayes theorem, which is quantitatively stated and suggests that C represents the class of observations X. When utilizing posterior probability as the basis for the

Bayes rule, it is possible to forecast the class X defined by the assumption $\{X_1, X_2, ..., X_n\}$.

$$P(C|X) = \frac{P(C)P(X|C)}{P(X)}. \tag{2.7}$$

$P(C|X)$ represents the posterior probability of the class $P(C)$ indicates the prior probability of the class $P(X|C)$ represents the likelihood which is the probability of predictor $P(X)$ presents the prior probability of the predictor.

Those evidences are given by class labels, the assumption of Naïve Bayes is that the features which are conditionally independent and are stated, in terms of mathematical formula, as:

$$P(X_1, X_2, ..., X_N|C) = P(X_1|C)P(X_2|C)P(X_3|C)...P(X_N|C). \tag{2.8}$$

In Equation 2.8, the values $X_1, X_2, ..., X_n$ are the features, and C is the class label.
As stated by Dukuzumuremyi [2020] in his findings, Naïve performed well enough to rank among the best predictors of under-5 children mortality in Rwanda, even if it was not the top classifier in the majority of the studies.

2.4.7 Support vector machines

Following the above assertion, it has been demonstrated by numerous researchers that also support vector machines (SVMs) are very good at handling high-dimensional data and identifying intricate relationships between variables. In Rwandan under-5 children mortality prediction, the Support Vector Machine (SVM) classifier shows potential due to its robust method for decoding complex information and deriving valuable insights.

According to Awad et al. [2015], once the SVM model is given a set of labeled training data, it is able to categorize the target outcome as yes or no depending on the classes of a target variable. A support vector machine outputs the hyperplane, which is in two-dimensional plane with a line separating the classes of the categorical target variable [Awad et al., 2015].
To demonstrate the dataset's labeling above, (x_i, y_i) where x_i represents the i^{th}feature vector and y_i represents the appropriate class, which in the case of binary classification is either 1 to represent the presence or -1 to represent the absence which in most studies was case is denoted by 0. The basis of SVM is finding the hyperplane that divides data points into the two classes with the highest margin; in the case of linear SVM

$$f(x) = Wx + b \tag{2.9}$$

where b is the bias term and W is the weight vector perpendicular to the hyperplane. This weight vector and

b should be determined through optimization so that every data point is accurately classified with the largest possible margin. To convert the input space of non-linearly separable data into a higher-dimensional feature space where the data can be separable linearly, SVM employs a kernel function. The function involved in making decisions is:

$$f(x) = \sum_{i=1}^{N} \alpha_i y_i K(x_i, x) + b \tag{2.10}$$

where α_i are the coefficients associated to y_i classes, $K(x_i, x)$ is the kernal function that goes with $\alpha_i y_i$ as directing vector to convert Equation 2.10 in linearly separable data like Equation 2.9 .
Using the above approach, Emanuel Mfateneza found that the support vector machine model placed second in terms of accuracy 68.6%, recall 74.9%, precision 67%, F1 score 70.73%, and AUROC 68.6% when it came to forecasting infant mortality in Rwanda [Mfateneza et al., 2022].

Briefly stated, it's still difficult to reliably predict death for children under-5 in Rwanda through traditional techniques. However, the emergence of Artificial Intelligence (AI) and Machine Learning (ML) offers a chance to improve health sciences forecasting capacities through the use of extensive datasets. So, there is a dearth of research expressly concentrating on the prediction of under-5 children mortality using more than four machine learning algorithms, given the wide range of variables in Rwanda, in order to potentially improve the accuracy.

Chapter 3

Methodology

This section outlines the methods used for data collection, variable selection, dataset transformation for machine learning, and the algorithm selection for accurate prediction of under-5 children mortality in Rwanda. The chosen algorithm and SHapley Additive exPlanations (SHAP) values were utilized to determine feature importance for predicting under-5 children mortality. Evaluation metrics were explored to assess the performance improvement of the chosen algorithm in predicting under-5 children mortality in Rwanda.

3.1 Study design

Secondary data from the 2019/2020 Rwandan Demographic and Health Survey (RDHS-2019/2020) dataset were obtained from the national institute of statistics of Rwanda (NISR) platform, after formal subscription which allow readers to download data from any Demographic and health survey in general. Out of 30,820 identity cases in the RDHS-2019/2020 dataset, only 8,065 were identified for individuals under 60 months old, forming the cohorts for this study.

3.2 Variables

Selected variables from the dataset describing risk factors influencing under-5 children mortality in Rwanda underwent machine learning algorithms for data preparation and analysis. Fourteen variables were selected, and split into two parts, namely dependent variables and independent variables. On the one hand, the variable:

- Child is alive (B5): is indicating the survival status of a live birth until the fifth birthday was chosen as the binary dependent variable. On the other hand, variables likely affecting under-5 children mortality were selected, including categorical variables like:
- Highest educational level of parents (S108): is the level of schooling attended and level of literacy mainly that of the mother.

- Exposure of the family (V625): is essential to our understanding of fertility since it is the main indicator of women's exposure to the risk being pregnant.

- Currently breastfeeding (V404): is breastfeeding status at the time of the survey, the kid may or may not be breastfeeding.

- Wealth index combined (V190):is a composite measure of a household's cumulative living standard.

- Desire for more children (V605): the question on desire for more children is rephrased to refer to desire for another child after the one that they are expecting.

- Sex of child (B4): as reported in the birth history, sex reported in the household schedule if the child was not reported in a birth history of an individual woman respondent.

- Number of antenatal visits during pregnancy (M14): is grouped into categories of no antenatal care, 1 visit, 2-3 visits, 4+ visits, and missing/"don't know" before calculating percentages. Timing of first antenatal visit is grouped into categories of no antenatal visit, $<=$ 3 months, 4-5 months, 6-7 months, 8+ months, and missing/"don't know" before calculating percentages.

- Had diarrhea recently (H11):this is observed among children under age 5 who had diarrhea in the 2 weeks preceding the survey.

- Source of drinking water (V113): the main source of drinking water for members of the household is classified as improved and unimproved sources, Typically, this is used as a characteristic related to prevalence of diarrhea and numerical variables like:

- Births in last five years (V208): are women of childbearing age about their reproductive history over the past five years. This helps in calculating fertility rates, understanding birth trends, and assessing maternal and child health.

- Number of living children (V218): The total number of children the individual woman respondent has given birth to, including any current pregnancy.

- Total children ever born (V201): is a demographic measure that refers to the total number of children a woman has given birth to over her lifetime.

- Birth order number (BORD): is the order number of the births from first to last. Twins are given the same birth order, but the birth order of a child born after twins will be the total number of births preceding plus one. For example, if a second birth resulted in twins the birth order will be 2 for both twins, and a birth following the twins will be birth order 4.

3.3 Analysis methods

After behaviors of the variables were analyzed using SPSS, RDHS-2019/2020 dataset was investigated to identify predictable feature importance among independent variables relating to under-5 children mortality using predictive modeling with machine learning algorithms. Machine learning is increasingly valuable for demographic and health surveys to enhance maternal and child health in Rwanda.

3.3.1 Data preparation and preprocessing

Pre-processing and data preparation techniques were used on the raw data to remove elements that could interfere with our machine learning models' ability to train and predict. The data that were deemed suitable for usage had been encoded, examined for data imbalance, and free of duplicates, outliers, and missing data.

3.3.2 Descriptive Analysis

Simple descriptive analysis were done to identify the relationship between independent variables and the dependent variable, using SPSS software for analysis.

3.3.3 Machine learning algorithm for predictive model

Seven machine learning algorithms were employed simultaneously with a fixed random state to determine the best predictor for the datasets, namely logistic regression, KNN classifier, Decision trees classifier, Random Forest, Linear discriminant analysis, Naïve Bayes, and Support vector machines. A brief explanation of each machine learning algorithm's application in predicting the mortality of under-5 children in Rwanda is provided below.

i. Logistic regression

The Logistic regression Machine learning technique is used to predict a categorical dependent variable given a set of independent variables. An S-shaped logistic function is fitted in place of a regression line in logistic regression, yielding a binary result with values between 0 and 1 [Rymarczyk et al., 2019].
Based on these predictor factors, the logistic regression model in this study calculates the probability of under-5 death. For every predictor variable x_i, the model estimates a coefficient that shows how much and in which direction the variable will affect the likelihood of children mortality y_i as shows Equation 2.1. The probability score it produces ranges from 0 to 1, with values closer to 1 denoting a higher likelihood of death and values closer to 0 denoting a lower likelihood. A logical and quantitative way to comprehending and forecasting

Rwanda's under-5 children mortality rate is provided by logistic regression.

ii. Linear discriminant Analysis

A statistical technique for dimensionality reduction and classification is called linear discriminant analysis (LDA). This last, can offer a strong methodology that makes use of the association between predictor factors and the categorical outcome (in this example, mortality status) to predict the mortality of under-5 children in Rwanda. The LDA model in Equation 2.2 is trained on large datasets to identify the linear combinations of predictor variables that provide the best discrimination between children who survived and those who did not. With the least amount of within-group variance, LDA determines the linear discriminant functions that optimize the separation between the mortality groups. The LDA technique is used to find a linear transformation that discriminants between different classes. If the classes are non-linearly separable, LDA cannot find a lower dimensional space [Tharwat et al., 2017]. The LDA model with suitable metrics including precision, sensitivity, specificity, and accuracy. To confirm the model's prediction power and guarantee that it can be applied to new data, cross-validation techniques can be used.

iii. KNN classifier

The KNN classifier, when used in data with high sample sizes such as those that approach infinity, the KNN classifier has demonstrated impressive performance, with its error rate roughly reaching the Bayes optimum under relatively moderate conditions [Hu et al., 2016]. However, a number of factors, including the choice of distance metrics and the k value, can have an impact on how well the KNN classification performs, and KNN classifiers classification accuracy can be impacted by the distance function that is selected [Zhang et al., 2017]. In this study Training a KNN involved keeping the complete dataset and figuring out how far apart each occurrence is. A new instance is classified by the method simply by looking at the majority class among its K closest neighbors being an user-defined number. Which cases are the closest neighbors is determined by the distance metric (such as the Euclidean distance) [Hu et al., 2016].
Considering the use and importance of KNN classifier, this machine learning technique is also worth being considered for this study to predict under-5 children mortality that uses a large dataset containing numerous variables.

iv. Decision trees classification algorithm

A decision tree (DT) is depicted inverted, with its root at the top and a node that corresponds to the con-

dition describing a characteristic. The decision rule is indicated by the branches of the tree that branch left or right from the node [Charbuty and Abdulazeez, 2021].
This study construct the Decision Tree by iteratively dividing the data into subsets and maximizing homogeneity within each subset using entropy or Gini impurity criteria. Understanding the hierarchical links between predictor variables and the target variable (survival or mortality) is made easier by the tree structure [Zhang et al., 2019].
Decision Tree learn more about the main causes of children mortality under the age of five. Nodes nearer the tree's base, for instance, indicate characteristics that have a bigger impact on mortality prediction.

v. Random Forest

Random Forest is a machine-learning algorithm widely used in classification problems. The algorithm contains a number of individual decision trees that operate as an ensemble on different subsets of the given data and each independent tree produces a class prediction and the class with the majority votes becomes the classifier's prediction. In a Random Forest, an ensemble means a combination of multiple models and therefore a collection of models as it is used to make predictions instead of an individual model [Louppe, 2014]. A certain number of decision trees must be grown in order to train the Random Forest model. At each split, each tree is trained individually using a bootstrapped sample of the dataset (a technique called bagging), and it bases its decisions on randomly chosen features. To get a final predicted, this technique combines the predictions from several trees.

vi. Support Vector Machine (SVM)

Support vector machines (SVM) is a supervised machine learning technique that is often used in classification problems. The SVMs work well in situations where it's necessary to divide data into discrete groups according to precise correlations between predictor variables, like survival and death. By optimizing the margin between the classes, the SVM algorithm creates a hyper-plane that precisely divides the training observations according to their class labels. SVM allocates a test observation to a class based on which side of the hyperplane it is on [Awad et al., 2015].
The goal of this work is to better find the ideal hyperplane that divides the classes (survival and mortality) in the feature space in order to train the SVM classifier. The distance between the hyperplane and the closest data points "support vectors" for each class is known as the margin, and SVMs seek to maximize it.

vii. Naïve Bayes classifier

This machine learning technique makes the assumption that a specific data point's properties are independent of one another, which is frequently not the case in practice. By calculating the odds of each class (survival and death) based on the observed feature values, you may train the Naïve Bayes classifier. Naïve Bayes classifiers determine the class with the highest probability as the prediction by applying Bayes' theorem to compute the posterior probability of each class given the data [Greiner et al., Technical report TR03-09, Department of Computing Science, University of Alberta, Canada, 2003].
Naïve Bayes has been proven to be very effective in a variety of applications, despite this oversimplifying assumption. When compared to more advanced techniques, Naïve Bayes classifiers and learners can operate incredibly quickly [Berrar, 2018]. Considering the benefits of the Naïve Bayes machine learning technique in past works of different authors, it is likely to offer similar benefits if used in this study.

3.3.4 Metrics for evaluating the performance of the predictive model deployed

Several metrics can be used to evaluate the effectiveness of a predictive model when it is used to predict the death of under-5 children, or any other predictive task. Evaluation metrics methods specifically confusion matrix, accuracy, precision, recall, F1 score, and Area under the Receiver Operating Characteristics (AUROC) were used to evaluate the performance of predictive models. Python 3 (ipykernel) packages like Pandas were the most used tools of analysis of variables associated with under-5 children mortality.

A. Confusion matrix

With evaluation metrics technique, a confusion matrix for risk factors influencing under-5 children mortality in Rwanda was constructed. Along the process, a supervised classification model that predicts the presence or absence of specific risk factors for individual children was needed. Also, it was assumed to have a dataset with labeled instances of children, where each instance had features representing various risk factors and a binary label indicating whether the child experienced under-5 children mortality expressed by 1 if the child is alive or 0 if not. The confusion matrix model can be expressed as in Figure 3.3-1.

		Predicted class	
		Positive	Negative
Actual Class	Positive	True Positive (TP)	False Negative(FN)
	Negative	False Positive(FP)	True Negative(TN)

Figure 3.3-1: Confusion matrix

From the matrix accuracy and sensitivity were defined as follows:
The sensitivity is given by the formula:

$$Sensitivity = \frac{TP}{TP + FN} \tag{3.1}$$

which was the percentage of all true cases of under-5 children that are alive in a population, the accuracy is defined as

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.2}$$

which is the probability of correct predictions of the test and the Equation 3.3 which shows how the good the model predict the the future under-5 children mortality data in Rwanda.

$$Specificity = \frac{TN}{TN + FP} \tag{3.3}$$

which is percentage of all under-5 children that are not alive in a population who are labeled correctly by the test in B [Naidu et al., 2023].

B. F1 score

Naidu et al. [2023] stated that F1 Score is the combination of precision and recall metrics, and the harmonic mean of precision (P) and recall (R). For this study:

$$F_1 = \frac{2PR}{P + R} \tag{3.4}$$

where

$$P = \frac{TP}{TP + FP} \tag{3.5}$$

and the true positive rate

$$R = \frac{TP}{TP + FN} \tag{3.6}$$

C. Area under the operating characteristic curve

Using the true positive rate and false positive rate from the confusion matrices, receiver operating characteristic (ROC) curves were constructed. A perfect classifier is indicated by a maximum value of 1, which falls between 0 and 1. This is known as the area under the curve (AUC). Nonetheless, the practical lower limit for random classification is 0.5, and classifiers with an AUC higher than 0.5 are at least capable of differentiating between the two target outcome classes. Receiver Operating Curves (ROC) were used to compare various models and identify the model with the best classification threshold. The Area Under the Curve (AUC) were used to calculate the interval within which the true Area under the ROC curve lies within 98% of confidence.

3.3.5 SHAP description

SHapley Additive exPlanations (SHAP) is also a method for interpreting the output of machine learning models. It leverages ideas from cooperative game theory to clarify how individual features contribute to the model's predictions. The SHAP values assign an importance score to each feature for a specific prediction, offering a thorough insight into the influence of each feature on the model's output. SHAP is based on Shapley values, a concept from cooperative game theory introduced by Lloyd Shapley in 1953. In a cooperative game, players work together to achieve a collective payoff. The Shapley value provides a fair distribution of this payoff among the players based on their contributions. In the context of machine learning, features are analogous to players, and the model's prediction is the payoff[JA, 2020].

Let $\hat{f}(x)$ be the original model to be explained and $g(z^{'})$ the explanation model. The purpose is to explain $\hat{f}(x)$ base on a single x. The Shapley values explanation in this study was represented as an additive feature attribution method or a linear model that connects LIME (Local Interpretable Model-agnostic Explanation) often use simplified inputs $x^{'}$ that map the original inputs through mapping $x = (h_x(x^{'}))$ and $g(z^{'}) \backsimeq \hat{f}(h_x(z^{'}))$ when $x^{'} \backsimeq z^{'}$ and Shapley values, as:

$$g(z^{'}) = \phi_0 + \sum_{j=1}^{M} \phi_j z_j^{'} \tag{3.7}$$

where g is the explanation model, $z^{'} \in \{0,1\}^M$ is the coalition binary vector, M is the maximum coalition size, $\phi_j \in R$ is the feature attribution for a feature j.

For x, the instance of interest the coalition vector $x^{'}$ is of all $1^{'}s$. So the simplified form is

$$g(x^{'}) = \phi_0 + \sum_{j=1}^{M} \phi_j \tag{3.8}$$

The shapley values were fulfilling three properties stated below:

a) local efficiency

The local efficiency $\hat{f}(x)$ is expressed by:

$$\hat{f}(x) = \phi_0 + \sum_{j=1}^{M} \phi_j x_j^{'} \tag{3.9}$$

local accuracy or efficiency requires the explanation model to at least match the output of $\hat{f}$ for the simplified input $x^{'}$.

b) Missingness

Drawn as $x_j^{'} = 0$ and $\phi_j = 0$ representing the absence of the feature values, then missingness requires features missing in the original input to have no impact.

c) consistency

Considering the fact that $\hat{f}_x(z^{'}) = \hat{f}(h_x(z^{'}))$ and $z_j^{'}$ indicating that $z_j^{'} = 0$ for f and $f^{'}$ which immplies that

$$\hat{f}_x(z^{'}) - \hat{f}^{'}(z_j^{'}) \geq \hat{f}_x(z^{'}) - \hat{f}_x(z_{-j}^{'}) \tag{3.10}$$

then

$$\phi_j(\hat{f}^{'}, x) \geq \phi_j(\hat{f}, x). \tag{3.11}$$

SHAP feature dependence

In this study part, SHAP dependence might be the simplest global interpretation plot where we can use feature and data instance mathematically and the plot is made by couples $\{x_j^i, \phi_j^i\}_{i=1}^n$ now the shapley interaction index

$$\phi_{ij} = \sum_{S \leq \{i,j\}} \frac{|S|!(M - |S| - 2)! S_{ij(S)}}{2(M-1)!}, \tag{3.12}$$

when $i \neq j$ so

$$S_{ij} = \hat{f}_x(S \cup \{i,j\}) - \hat{f}_x(S \cup \{i\}) - \hat{f}_x(S \cup \{j\}) + \hat{f}_x(S), \tag{3.13}$$

where M is the set of all features, S is the subset of N not containing feature i, $|S|$ is the number of features

in subset S, $\hat{f}_x(S)$ is the model's prediction using only the features in subset S, $\hat{f}_x((S) \cup i)$ is the model's prediction using only the features in subset S plus feature i. $\hat{f}_x((S) \cup \{i, j\})$ is the model's prediction using only the features in subset S plus feature i and j, $\hat{f}_x((S) \cup j)$ is the model's prediction using only the features in subset S plus feature j [Molnar, 2021].

Briefly; the SHAP computes shapley values with a theoretical foundation in game theory, the prediction is fairly distributed among the feature value it connects LIME (Local Interpretable Model-agnostic Explanation) and shapley values, for the tree based model each predict is given a local explanation by SHAP values, which show which features drove the prediction higher or lower in relation to the average prediction. Each predict is given a local explanation by SHAP values, which show which features drove the prediction higher or lower in relation to the average prediction [Chaudhuri et al., 2023].

Chapter 4

Data analysis and main results

4.1 Introduction

The dataset for children under-5 in this study was sourced from the 2019/2020 Rwanda Demographic Health Survey (RDHS), selecting data for case identification between 2015 and 2019. By reviewing many works of literature to create the working dataset, 14 variables that were linked with mortality among children under-5 were selected. In actuality, one dependent variable and thirteen independent variables were employed.

This large dataset, which included both category and numerical data, was analyzed using Python 3.12 and a number of packages, including Pandas, NumPy, Matplotlib, seaborn, warnings and SHAP. Seven machine learning techniques were employed concurrently to train and test the under-5 children dataset, which allowed for the ranking of significant predictors of under-5 children mortality.

Research findings are presented in forms of tables, graphs, or diagrams as extracted from the research instruments used then a kind of explanation or interpretation follows after each presented item. All items are presented in connection with prior research objectives and research questions. The chapter therefore serves as empirical evidence in verification of hypotheses implied in this study, and serves as a basis for grounded conclusions.

4.2 Selected variables for under-5 children mortality

In this study, thirteen independent variables (Highest educational level, Births in last five years, Exposure, currently breastfeeding, number of living children, Wealth index combined, Total children ever born, Desire for more children, Sex of child, Birth order number, Number of antenatal visits during pregnancy, had diarrhea recently, source of drinking water) and the dependent variable “child is alive”, were considered selecting under-5 variables related to children mortality. Before more analysis, each variable included for the study is briefly described in the series of tables below.

Table 4.2- 1 : Selected variables for Under-5 children mortality

Number	Identification	Type	Label of case identification	Coded values
1	HE	numeric	Highest education level	1, pre-primary
2	BLFY	numeric	Births in the last five years	0, no birth
3	EXP	numeric	Exposure of the mother	0, Fecund
4	CBF	numeric	Currently brestfeeding	0, no
5	NLC	numeric	Number of living children	none
6	WI	numeric	Wealth index combined	1, poorest
7	TCEB	numeric	Total children ever born	none
8	DMC	numeric	Desire for more children	1, wants within 2 years
9	Sex	numeric	Sex of child	1,male
10	BO	numeric	Birth order number	1, none
11	D/A	numeric	child is alive	0, no
12	NAV	numeric	Number of antenatal visits during pregnancy	0,no antenatal visits
13	HD	numeric	Had diarrhea recently	0,no
14	SDW	numeric	Source of drinking water	10, piped water

As it can be seen in the Table 4.2- 1 , fourteen variables were selected as represented and lebaled in RDHS-2019/2020 as follows; HE (S108 or Highest educational level of parents), EXP (V625 or Exposure of the mother), CBF(V404 or Currently breastfeeding, WI (V190 or Wealth index of the family combined), DMC (V605 or Desire for more children), Sex (B4 or Sex of child), NAV (M14 or Number of antenatal visits during pregnancy), HD (H11 or Had diarrhea recently), SDW (V113 or Source of drinking water), and numerical variables like BLFY (V208 or Births in last five years), NLC (V218 or Number of living children), TCEB (V201 or Total children ever born), and BO (BORD or Birth order number). The dependent variable was coded by B5 (child is alive).

Table 4.2- 2 : The first 10 rows of the working dataset

CaseId	HE	BLFY	EXP	CBF	NLC	WI	TCEB	DMC	Sex	BO	D/A	NAV	HD	SDW
403 5 02	2.0	3	2	0	2	1	3	1	2	3	0	1.0	NaN	41
7 7 02	2.0	4	2	1	7	5	7	2	1	7	1	4.0	0.0	42
20 15 03	2.0	1	2	1	1	2	1	3	2	1	1	4.0	0.0	41
36 23 02	4.0	2	2	1	2	2	2	2	1	2	1	2.0	0.0	42
46 7 02	2.0	1	2	1	1	1	1	2	1	1	1	3.0	0.0	43
56 8 02	2.0	2	2	1	3	2	3	2	2	3	1	4.0	0.0	14
56 23 02	4.0	2	2	1	3	3	3	5	2	3	1	2.0	0.0	14
60 11 02	4.0	1	2	1	1	4	1	2	2	1	1	3.0	0.0	14
82 27 01	4.0	1	2	1	1	5	1	2	1	1	1	4.0	0.0	14
85 15 02	2.0	2	2	1	3	1	3	7	2	3	1	3.0	0.0	42

Table 4.2- 2 displays the first 10 rows of the working dataset of RDHS-2019/2020. This dataset initially contained too much information that was assigned to each of the variables, both dependent and independent; then the NaN data stands for missing data. In this study, 8065 children were represented by case identifications (CaseId) and were used for further analysis.

4.3 Statistical analysis of the dataset

The next string of tables gives the description for each of the 14 variables that were considered to examine the likelihood of the mortality of under-5 children in Rwanda as per the RDHS-2019/2020 dataset.

Table 4.3- 3 : Births in Last Five Years

Births in Last Five Years	Frequency	Percent	Cumulative Percent
1	4390	54.4	54.4
2	3219	39.9	94.3
3	424	5.3	99.6
4	27	.3	99.9
5	5	.1	**100.0**
Total	**8065**	**100.0**	

Table 4.3- 3 displays the status of mothers who were pregnant during the five years prior to the survey. Most of them had one child, but a tiny fraction had five throughout a five-year period, which could have an effect on the health of their progeny.

Table 4.3- 4 : Exposure of the mother

Exposure	Frequency	Percent	cummulative Percent
Fecund	4037	50.1	50.1
Pregnant	521	6.5	56.6
Postpartum amenorrheic	3055	37.9	94.4
Infecund, menopausal	452	5.6	**100.0**
Total	**8065**	**100.0**	

Table 4.3- 4 lists the mothers' exposures, which may have an impact on how they raised their kids.

Table 4.3- 5 : Currently breastfeeding

Currently breastfeeding	Frequency	Percent	Cumulative percent
No	2952	36.6	36.6
Yes	5113	63.4	**100.0**
Total	**8065**	**100.0**	

Table 4.3- 5 lists the mothers' breastfeeding status, which may have an impact on the under-5 children's welfare at some point.

Table 4.3- 6 : Number of living children

Number of living children	Frequency	Percent	Cumulative percentage
0	38	.5	.5
1	1589	19.7	20.2
2	2030	25.2	45.3
3	1592	19.7	65.1
4	1144	14.2	79.3
5	772	9.6	88.8
6	458	5.7	94.5
7	264	3.3	97.8
8	121	1.5	99.3
9	35	.4	99.7
10	14	.2	99.9
11	8	.1	**100.0**
Total	**8065**	**100.0**	

Table 4.3- 6 lists all of the children that were still residing in a home. Family expansion affects every member of the family, but it most directly affects the children.

Table 4.3- 7 : Wealth index combined

Wealth index combined	Frequency	Percent	Cumulative percentage
Poorest	1993	24.7	24.7
Poorer	1566	19.4	44.1
Middle	1514	18.8	62.9
Richer	1508	18.7	81.6
Richest	1484	18.4	**100.0**
Total	**8065**	**100.0**	

Table 4.3- 7 displays the families' wealth levels. It is evident that families with lower incomes were disproportionately affected by a variety of health issues that might eventually cause the death of their children.

Table 4.3- 8 : Total children ever born

Total children ever born	Frequency	Percent	Cumulative Percent
1	1477	18.3	18.3
2	1922	23.8	42.1
3	1566	19.4	61.6
4	1120	13.9	75.4
5	741	9.2	84.6
6	533	6.6	91.2
7	359	4.5	95.7
8	190	2.4	98.1
9	78	1.0	99.0
10	52	.6	99.7
11	19	.2	99.9
12	8	.1	**100.0**
Total	**8065**	**100.0**	

Table 4.3- 8 details all of the children that have been born into the family. The likelihood that mothers will eventually become insufficiently healthy to take care of themselves increases with the number of children they have.

Table 4.3- 9 : Desire for more children

Desire for more children	Frequency	Percent	Cumulative Percent
Wants within 2 years	367	4.6	4.6
Wants after 2+ years	3156	39.1	43.7
Wants, unsure timing	240	3.0	46.7
Undecided	78	1.0	47.6
Wants no more	3935	48.8	96.4
Sterilized	128	1.6	98.0
Declared infecund	161	2.0	**100.0**
Total	**8065**	**100.0**	

Table 4.3- 9 shows the families surveyed's desire for more children. The desire for close deliveries would make it harder to care for them than for individuals who need more time to give birth.

Table 4.3- 10 : Sex of child

Sex of Child	Frequency	Percent	Cumulative Percent
Male	4080	50.6	50.6
Female	3985	49.4	**100.0**
Total	**8065**	**100.0**	

Table 4.3- 10 presents the surveyed children's sex repartition, which is relevant to their future in their families.

Table 4.3- 11 : Birth order number

Number	Frequency	Percent	Cumulative percentage
1	2043	25.3	25.3
2	1817	22.5	47.9
3	1465	18.2	66.0
4	1017	12.6	78.6
5	670	8.3	86.9
6	462	5.7	92.7
7	313	3.9	96.6
8	149	1.8	98.4
9	70	0.9	99.3
10	40	0.5	99.8
11	15	0.2	100.0
12	4	0.0	**100.0**
Total	**8065**	**100.0**	

Table 4.3- 11 contains the birth order number. The number of years between births provides insight into a person's parenting style. In order to advise children's health, it is still necessary to consider the order of following births, even if most families consider financial considerations.

Table 4.3- 12 : Number of antenatal visits during pregnancy

Number of Antenatal VisitsDuring Pregnancy	Frequency	Percent	Valid Percent	Cumulative Percent
No antenatal visits	128	1.6	2.1	2.1
1	247	3.1	4.0	6.1
2	699	8.7	11.4	17.4
3	2151	26.7	34.9	52.4
4	2787	34.6	45.3	97.7
5	99	1.2	1.6	99.3
6	22	.3	.4	99.6
7	6	.1	.1	99.7
8	6	.1	.1	99.8
9	3	.0	.0	99.9
10	6	.1	.1	100.0
Don't know	1	.0	.0	**100.0**
Total	**6155**	**76.3**	**100.0**	

Table 4.3- 12 shows the frequency of antenatal visits during pregnancy. The greater frequency of prenatal visits that parents make throughout pregnancy has an effect on the newborn's future health.The valid percent shows the portion of all 8065 situations where missing values are present.

Table 4.3- 13 : Had diarrhea recently

Had Diarrhea Recently	Frequency	Percent	Valid Percent	Cumulative Percent
No	6627	82.2	85.3	85.3
Yes, last two weeks	1103	13.7	14.2	99.5
Don't know	39	.5	.5	**100.0**
Total	**7769**	**96.3**	**100.0**	

Table 4.3- 13 lists the number of children who had diarrhea a few days before the survey. Diarrhea may have had an impact on the affected children's living conditions.

Table 4.3- 14 : Source of drinking water

Source	Frequency	Percent	Cumulative Percent
Piped into dwelling	46	.6	.6
Piped to yard/plot	855	10.6	11.2
Piped to neighbour	270	3.3	14.5
Public tap/standpipe	2295	28.5	43.0
Tube well or borehole	121	1.5	44.5
Protected well	188	2.3	46.8
Unprotected well	122	1.5	48.3
Protected spring	2274	28.2	76.5
Unprotected spring	964	12.0	88.5
River/dam/lake/irrigation channel	615	7.6	96.1
Rainwater	37	.5	96.6
Tanker truck	1	.0	96.6
Cart with small tank	7	.1	96.7
Bottled water	157	1.9	98.6
Other	7	.1	98.7
Not a dejure resident	106	1.3	**100.0**
Total	**8065**	**100.0**	

Table 4.3- 14 displays the source of the drinking water. Depending on how they get it, the availability of drinking water may have an impact on children's lives, and the cleanliness of the water may have an affect on their quality of life.

Table 4.3- 15 : Highest educational level of the mother

Level	Frequency	Percent	Valid Percent	Cumulative Percent
Primary	5195	64.4	72.6	72.6
Post-primary/vocational	66	.8	.9	73.5
Secondary	1503	18.6	21.0	94.5
Higher	389	4.8	5.5	**100.0**
Total	**7153**	**88.7**	**100.0**	

Table 4.3- 15 shows the mother's highest degree of education. Having moms with a certain degree of education was crucial to raising their children; mothers' education has an effect on the children's way of life in one way or another.

Table 4.3- 16 : child is alive

child is alive?	Frequency	Percent	Cumulative Percent
No	296	3.7	3.7
Yes	7769	96.3	**100.0**
Total	**8065**	**100.0**	

Table 4.3- 16 presents the binary dependent variable in the study, "child is alive" in the context of the DHS. The under-5 Children mortality rate was seen in the 296 (3.7%) deaths of the under-5 children.

4.4 Data preprocessing

In data analysis, missing data is an issue to handle before any other activity then continues to plague data analysis methods. Since analysis methods gain sophistication, it is natural to encounter missing values in fields, but also all variables being equal, more data is always better.

4.4.1 Data cleaning

In this study, data cleaning was performed by removing all missing values and duplicates to have the final cleaned working dataset ready to explore some useful visualization and statistical analysis. H11 was eliminated because there was no data in it at the time. It would be better to verify if there are any additional missing values,it requires the removal of the case identification containing the missing data. The sample of a cleaned dataset are presented in Table 4.4- 17 .

Table 4.4- 17 : Cleaned dataset

Id	CaseId	HE	BLFY	EXP	CBF	NLC	WI	TCEB	DMC	Sex	BO	D/A	NAV	SDW
0	403 5 02	2.0	3	2	0	2	1	3	1	2	3	0	1.0	41
1	7 7 02	2.0	4	2	1	7	5	7	2	1	7	1	4.0	42
2	20 15 03	2.0	1	2	1	1	2	1	3	2	1	1	4.0	41
3	36 23 02	4.0	2	2	1	2	2	2	2	1	2	1	2.0	42
4	46 7 02	2.0	1	2	1	1	1	1	2	1	1	1	3.0	43
–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
8056	463 5 02	2.0	1	0	0	2	5	2	4	1	2	1	3.0	14
8057	464 3 02	2.0	1	0	0	4	1	4	5	2	4	1	3.0	42
8058	464 4 03	2.0	1	0	0	2	1	2	5	1	2	1	3.0	43
8059	464 12 02	2.0	1	0	0	6	2	6	5	1	6	1	3.0	42
8063	495 24 02	2.0	2	1	0	2	1	3	2	1	3	1	2.0	14

After removing the missing data as Table 4.4- 17 shows it, the dataset has changed to a cleaner one containing 5486 cases with 13 variables. Now, Table 4.4- 18 is the newly generated dependent variable.

Table 4.4- 18 : New child is alive

child is alive?	Frequency	Percent Valid	Cumulative Percent
No	128	2.3	2.3
Yes	5358	97.7	**100.0**
Total	**5486**	**100.0**	

In the working dataset, out of 5,486 rows, only 128 rows (2.33%) are associated with deceased children as shows the above Table 4.4- 18 . Thus, the elaborated datasets were the ones that were subjected to machine learning analysis.

4.4.2 Data encoding

The under-5 children dataset extracted from RDHS-2019/2020 contained continuous and categorical variables. As the machine learning tools could not treat categorical data in the same way as numerical data, some handlings were required to convert non continuous variables into numerical data which are readable by machine learning algorithms. The encoding process can be easily shown in Table 4.4- 19 .

Table 4.4- 19 : New dataset ready for machine learning

Id	HE	EXP	CBF	WI	DMC	SEX	SDW	BLFY	NLC	TCEB	Bo	NAV	D/A
0	0	2	0	0	0	1	7	3	2	3	3	1.0	0
1	0	2	1	4	1	0	8	4	7	7	7	4.0	1
2	0	2	1	1	2	1	7	1	1	1	1	4.0	1
3	2	2	1	1	1	0	8	2	2	2	2	2.0	1
4	0	2	1	0	1	0	9	1	1	1	1	3.0	1
–	–	–	–	–	–	–	–	–	–	–	–	–	—
5481	0	0	0	4	3	0	3	1	2	2	2	3.0	1
5482	0	0	0	0	4	1	8	1	4	4	4	3.0	1
5483	0	0	0	0	4	0	9	1	2	2	2	3.0	1
5484	0	0	0	1	4	0	8	1	6	6	6	3.0	1
5485	0	1	0	0	1	0	3	2	2	3	3	2.0	1

The modified new data in Table 4.4- 19 consists of numerical values that machine learning technologies can easily comprehend. For the purpose of training and testing the under-5 children mortality dataset in question, the encoding method proved to be a very helpful tool in converting the categorical data into a continuous dataset ready for the machine learning comprehension.

4.5 Machine Learning (Classification Problem)

Algorithms for machine learning are those that can identify hidden patterns in a dataset, forecast results, and enhance performance via independent experience. The used classification is a supervised machine learning procedure that predicts the class of input data points. To forecast the mortality of children under-5, seven machine learning algorithms were used: Random Forest, K-Nearest Neighbor, Decision Tree, Support Vector Machine, Linear discriminant Analysis, Naïve Bayes, and Logistic Regression Algorithm.

4.5.1 The training and testing the dataset under treatment

Assigning x as the independent variable and y as the dependent variable is the first step in training the machine on the available dataset. Only 20% of the data is used for testing; the remaining 80% is used for training through Stratified K Fold and Grid Search Cross Validation approach. The fixed random state 42 was used to guarantee repeatability. This division aims to evaluate which algorithms are more predictive of under-5 children mortality than others. The next subsection contains the testing results.

4.5.2 Selection of the machine learning algorithm for prediction

In this section, seven classification models were utilized, comprising linear and nonlinear models: Linear Discriminant Analysis (LDA), Logistic Regression (LR), Decision Trees (DT), Naïve Bayes (NB), Random Forest

(RF), Support Vector Machines (SVM), and K-Nearest Neighbors (KNN). The cross-validation outcomes for the fixed random state are shown in the following Table 4.5- 20 as averages (Average-cv-results) and standard deviations (Std-cv-results), which represent the performance levels in forecasting the dataset through training and testing sets.

Table 4.5- 20 : Results for chosen machine algorithm

ML_Algo	Average_cv_ results	Std_cv_ results
LR	0.979	0.005
LDA	0.963	0.009
KNN	0.976	0.001
DT	0.977	0.005
NB	0.599	0.016
SVM	0.976	0.002
RF	0.982	0.003

Table 4.5- 20 displas the average cross validation results where Random Forest (RF) scored an average of 0.982, followed by logistic regression, which scored 0.979. Random Forest emerged as the top classifier for predicting Under-5 children mortality with average cross validation score of 98.2%.

4.5.3 Comparison of machine learning algorithms

In comparison of the highly-ranked model to others, the following Figure 4.5-1 shows it graphically.

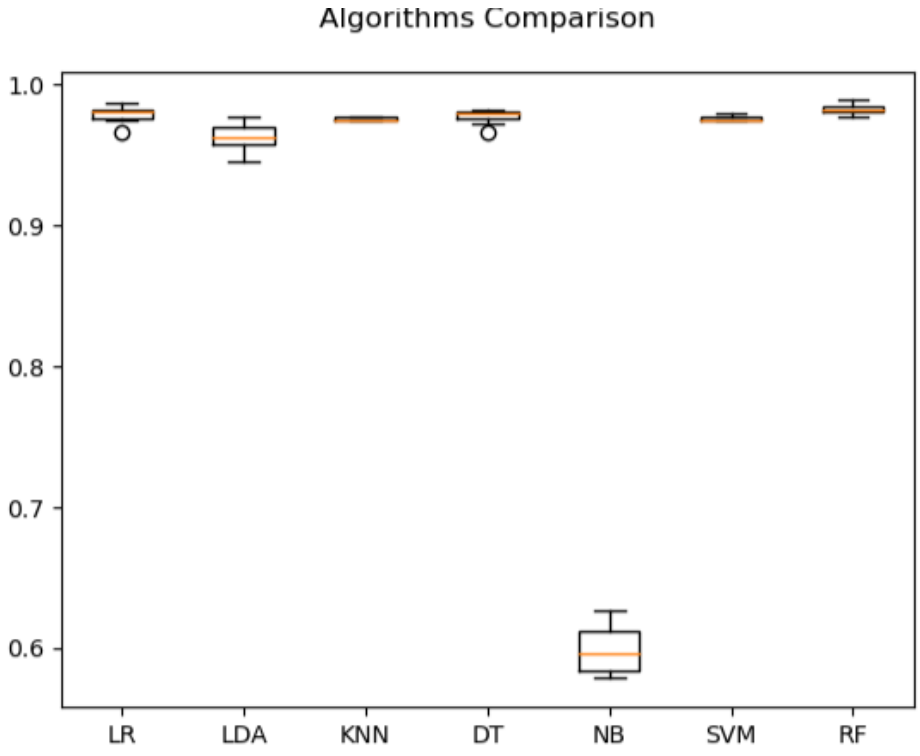


Figure 4.5-1: algorithm-comparison

As shown in the Figure 4.5-1, there is no big difference among most of the classifiers but the Random

Forest model emerged as the most accurate predictor of under-5 children mortality in Rwanda. Despite some unexpected results, the Random Forest performed better in categorizing children under-5 as alive. Therefore, further model evaluation was necessary to enhance the process.

4.5.4 Model Predictions

The under-5 children mortality rate in Rwanda could be most accurately predicted by the Random Forest model, Hence it was critical to confirm that the predicted and expected values agreed without checking whether the expected values are the same as the predicted ones in table 4.5- 20

Table 4.5- 21 : Model prediction

Id	Expected	Predicted
568	0	1
526	1	1
180	1	1
879	1	1
845	1	1
96	1	1
144	1	1
1031	1	1
364	1	1
229	1	1

As shows Table 4.5- 21 , overall the Random Forest has done a better job of categorizing the children under-5 as alive, but some expected death (0) are predicted to be alive (1). Therefore further model evaluation was required to improve the process.

4.5.5 Model evaluation metrics

Model evaluation metrics were utilized to evaluate the effectiveness of the selected model in predicting mortality or identifying cases accurately. Various techniques were employed, all centering around the confusion matrix: F1 scores, the confusion matrix itself, the ROC Curve before parameter tuning, and hyperparameter tuning to optimize the performance of Random Forest$'$s prediction.

- **Confusion matrix**

The confusion matrix was utilized to examine the predictions and identify any potential biases in the model. Every feature in the study may have a prediction made by the selected machine learning method that is related to the predicted one, and the Figure 4.5-2 below is the best tool to avail it.

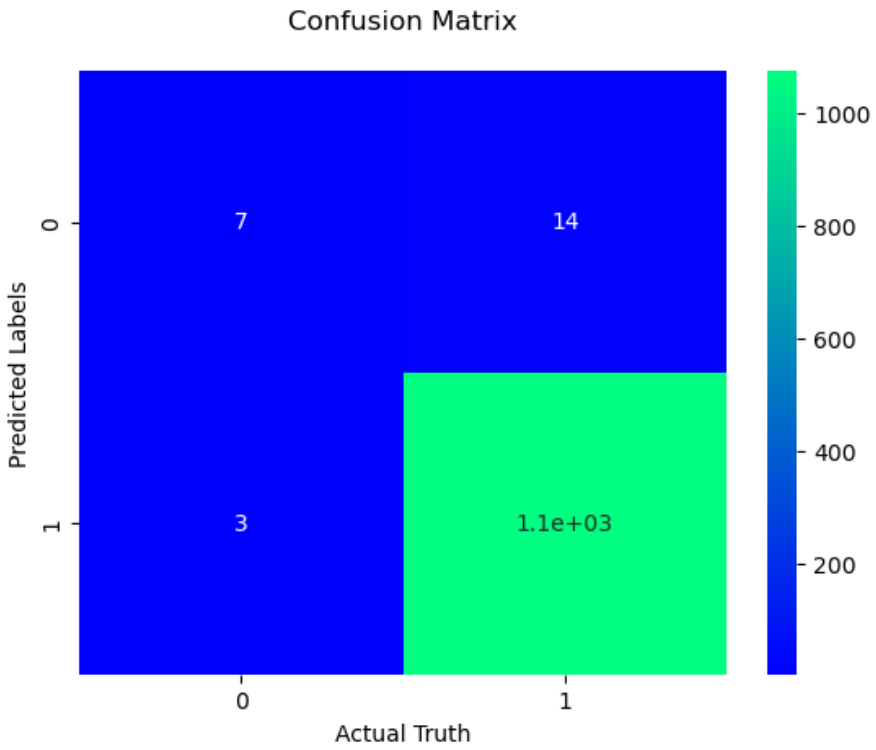


Figure 4.5-2: confusion matrix A

In Figure 4.5-2 it is evident that the Random Forest model incorrectly predicted 14 children to die (False Positive) when they were actually alive. Conversely, it predicted 3 children to be alive (False Negative) when they had already passed away. However, the last diagonal portions were correctly predicted as expected.

- **F1 scores**

Traditionally, F1 scores were calculated using the confusion matrix data shown above. The F1 score is equal to twice the product of recall and precision divided by their sum. The outcomes discovered are presented in the Table 4.5- 22 below.

Table 4.5- 22 : F1 scores

D/A	Precision	recall	f1-score	support
0	0.70	0.33	0.45	21
1	0.99	1.00	0.99	1077
Accuracy			0.98	1098
macro avg	0.84	0.67	0.72	1098
Weighted avg	0.98	0.98	0.98	1098

The aforementioned claim states that the accuracy of the F1 scores as a good predictor was 98% evaluated by applying Equation 3.2 . Since the dataset was imbalanced the accuracy should be supported by Area Under the Curve (AUC) probability values to accept the success of the chosen model.

- **ROC Curve before tuning parameters**

Receiver Operating Characteristics (ROC) was used to determine if the model fits the dataset using the confusion matrix shown above, with 0 denoting falsehood and 1 denoting truth. The Figure 4.5-3 demonstrates the area under the curve.

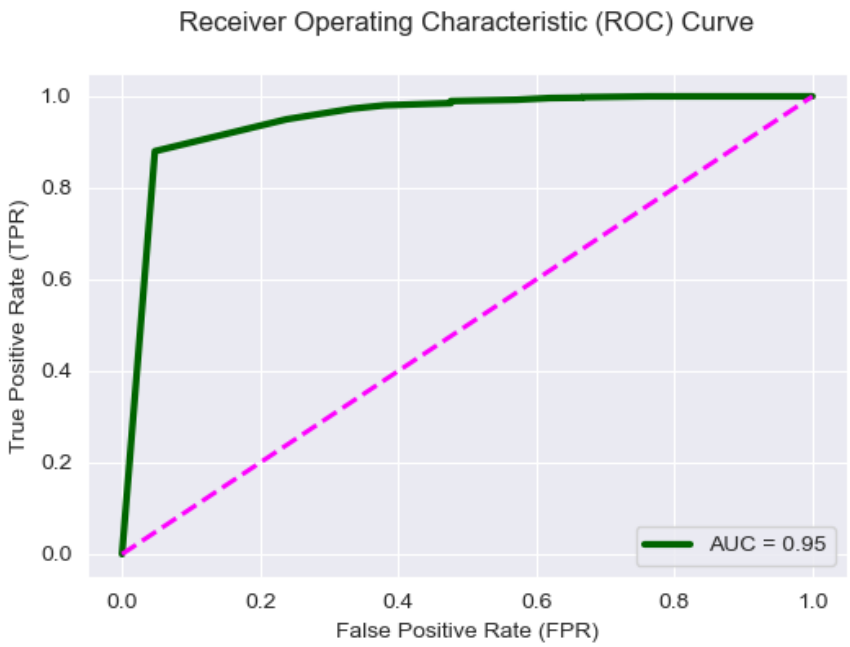


Figure 4.5-3: ROC/Area Under the Curve

The area under the curve was fitted at 95%, which is acceptable. However, it would be advisable to explore whether it can be further improved by utilizing a different model, for instance the one in the next subsection.

- **Hyperparameter tuning to improve the performance of RF**

In this phase, hyperparameter tuning was employed to enhance the model's ability to predict under-5 children mortality. It was suggested to train the model using optimal parameters across the entire training set by introducing new splits and estimator size for testing, what led to the following adjustments in the confusion matrix:

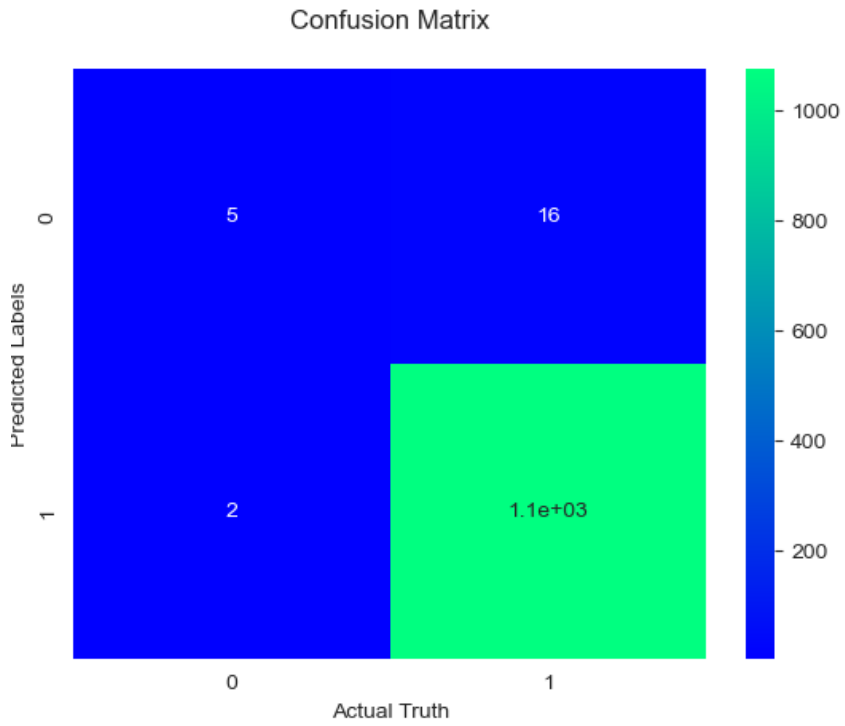


Figure 4.5-4: confusion tuned

As for the case of the confusion matrix explored earlier, this model predicted 16 children to have died (False Positive) when they were still alive; and predicted 2 children to be alive (False Negative), while they died.

- **ROC Curve after tuning parameters**

After hyperparameter adjustment, the receiver operating characteristic was enhanced for the most current confusion matrix. Below is what the diagram demonstrates:

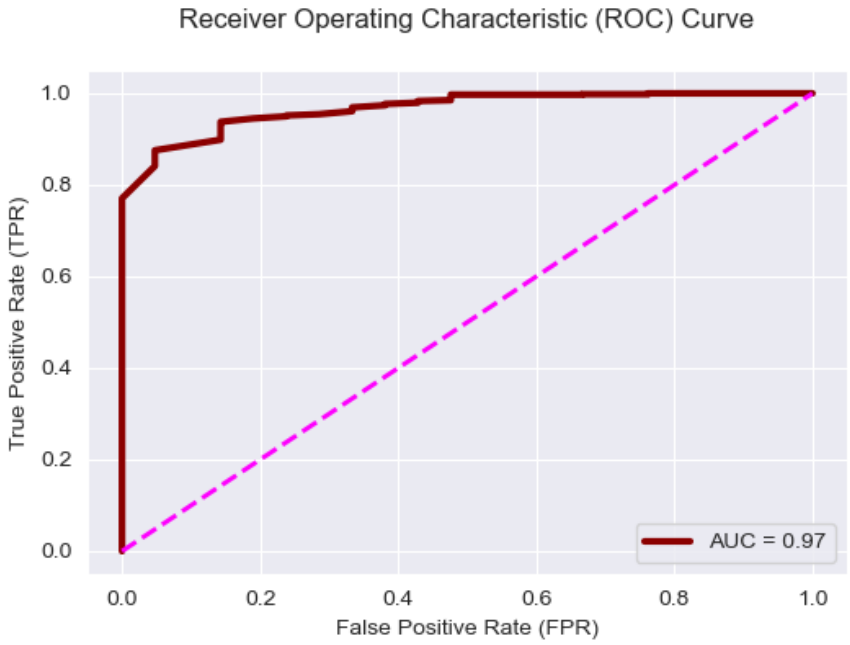


Figure 4.5-5: ROC/Area Under the Curve

As shown by the Figure 4.5.5 the area under the curve was showing a bright correspondence of true positive with the false positive rate at 97% level, which was better than the on before hyperparameter tuning.

Table 4.5- 23 : New F1 scores

D/A	Precision	recall	f1-score	support
0	0.71	0.24	0.36	21
1	0.99	1.00	0.99	1077
Accuracy			0.98	1098
macro avg	0.85	0.62	0.67	1098
Weighted avg	0.98	0.98	0.98	1098

Table 4.5- 23 provide a slight increase in the accuracy score, the model's ability to classify the label of under-5 children as died has improved after hyperparameter adjustment. In the ROC-AUC representation, hyperparameter adjustment has elevated the Area Under the Curve (AUC) from 95% to 97%.

4.5.6 Feature importance

The main performance metric of the model was feature importance, enabling it to identify the features that were primarily thought to predict the under-5 children mortality in Rwanda. Since the Random Forest algorithm proved to be the most effective in predicting under-5 children mortality data from the RDHS-2019/2020. That is through the operation of Classification and Regression Trees (CART), where the trait is more significant the higher the rise in leaf purity. This is completed for every tree, averaged over all the trees, and then normalized to 1. Therefore, a Random Forest's complete importance score is 1, and the following outcomes were observed:

Table 4.5- 24 : Feature importance levels

Id	Features	Feature importance
0	HE	0.019
1	EXP	0.038
2	CBF	0.045
3	WI	0.066
4	DMC	0.056
5	Sex	0.037
6	SDW	0.078
7	BLFY	0.055
8	NLC	0.409
9	TCEB	0.062
10	B0	0.060
11	NAV	0.074

As seen in the Table 4.5- 24 , the number of living children (NLC) and the source of drinking water (SDW) emerged as the most important features for classification respectively. The Random Forest algorithm employed here facilitated the classification of feature importance. Thus, this can be well depicted in Figure 4.5.6 which shaws it more clearly.

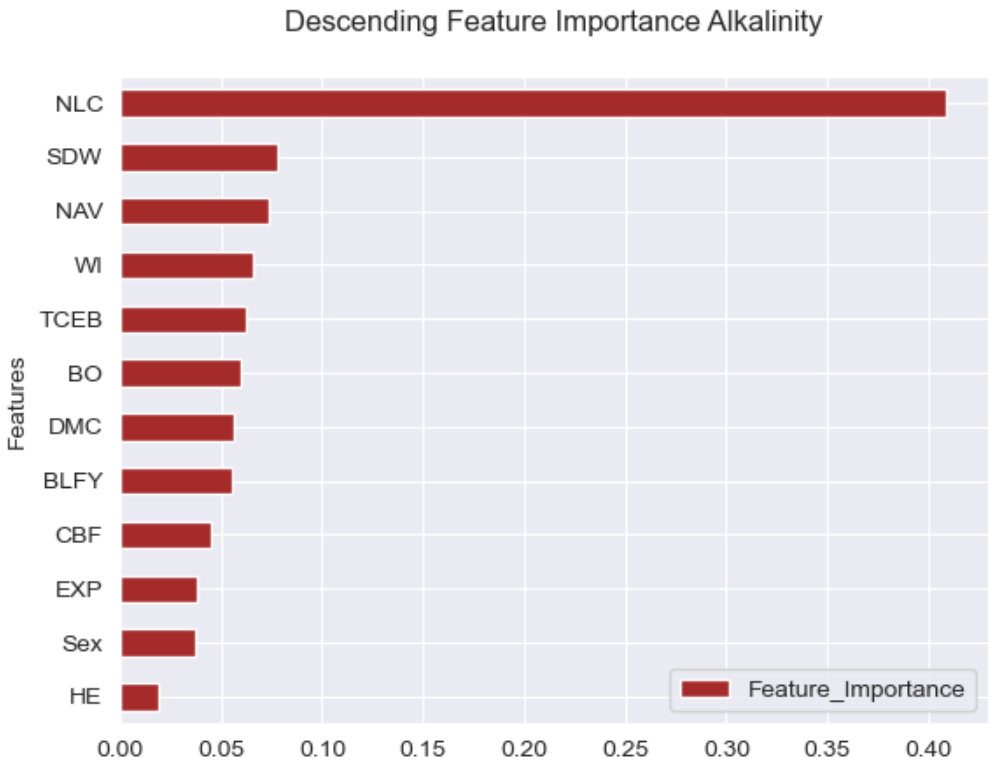


Figure 4.5-6: Descending feature importance levels

In contrast, the children′s sex and education level of parents showed no influence on the mortality rate for children under-5. In summary, the Random Forest algorithm emerges as the most reliable predictor of under-5 children mortality in Rwanda, surpassing other algorithms. Data from RDHS-2019/2020 highlights the

number of living children as the most important factor to consider when forecasting under-5 children mortality in Rwanda as observed in the outcomes from the Figure 4.5.6.

4.5.7 Important features using SHAP values

The SHAP values are presented because the SHapley Additive exPlanation (SHAP) is a potent model that can effectively identify and rank the relevance of features in predicting data as a machine learning approach.

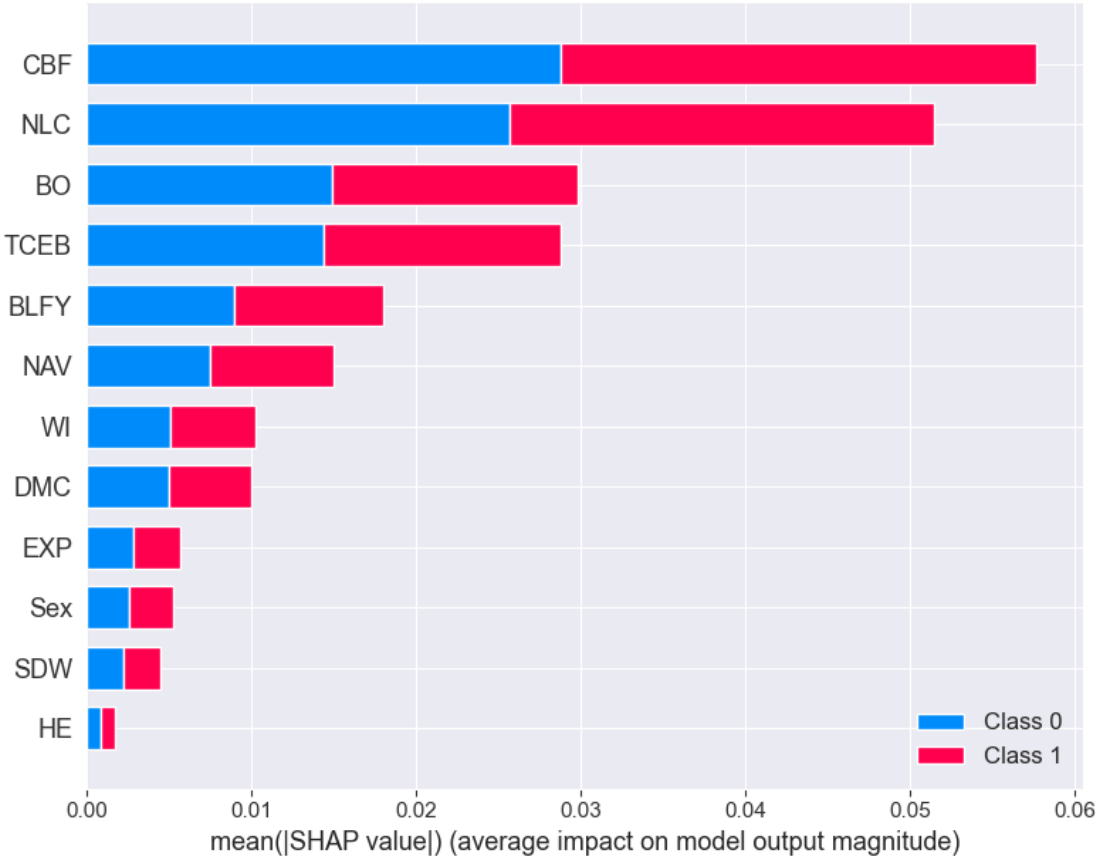


Figure 4.5-7: Mean SHAP values

Currently breastfeeding ranks the top factor predicting the mortality of children under-5, based on the SHAP values' average influence on model output magnitude. The number of living children also proves to be a significant predictor of mortality as shown in Figure 4.5.7. However, there exists a slight disparity between the traditional feature importance provided by the RF classifier built-in feature and the SHAP values' feature importance. Higher education level (HE) is identified as the least significant feature by both techniques. While SHAP values attribute contributions to each feature for a specific label prediction, the feature importance from the Random Forest function indicates a feature's importance in terms of the model's predictive performance. The number of living children (NLC) and the total number of children ever born (TCEB) rank among the top five powerful features for prediction, irrespective of the technique used to represent feature relevance. Lastly, exposure (EXP), higher education (HE), and child sex (Sex) rank as the least effective variables for this classification, according to both approaches.

4.5.8 Discussion of the results

The present study conducted using the Rwandan Demographic and Health Survey (RDHS-2019/2020) dataset resulted in successful and reliable findings. Regarding the mortality rate, 37 children out of 1000 live births were found to have passed away before reaching 60 months, reflecting a mortality rate consistent with the global under-5 children mortality rate of 37 deaths per 1000 live births in 2020. Concerning machine learning training and testing, seven machine learning algorithms were trained and tested with a fixed random state, including Linear Discriminant Analysis (LDA), Logistic Regression (LR), Decision Trees (DT), Naïve Bayes (NB), Random Forest (RF), Support Vector Machines (SVM), and K-Nearest Neighbors (KNN). Among these, Random Forest emerged as the best predictor, as indicated by its highest cross validation average's results (0.982). These findings are consistent with earlier research that found that Random Forest is the best machine learning technique for more accurately predicting under-5 children mortality [I, 2007, Mfateneza et al., 2022, Bitew et al., 2020]. Nevertheless, some studies contradicted this by finding that the same machine learning algorithm had a low predictive accuracy. Comparing the results from Random Forest and SHAP values, a slight disparity in ranking feature importance was observed [Shukla et al., 2020].

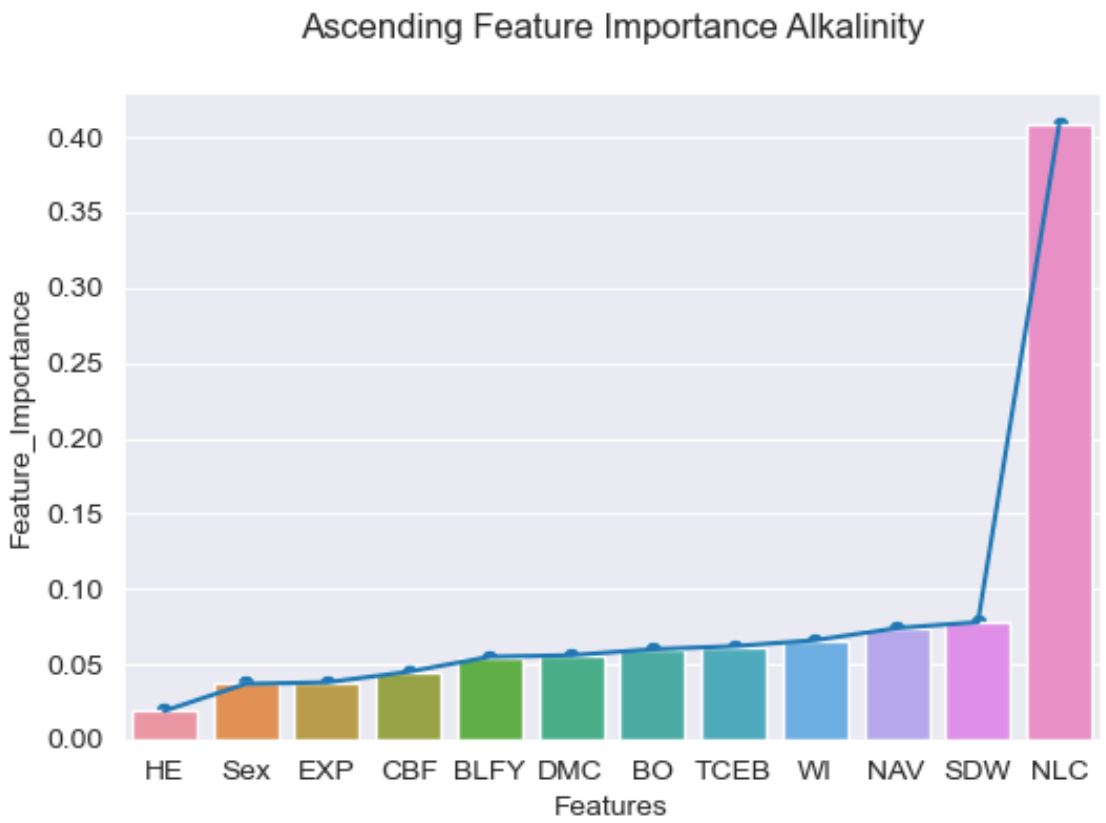


Figure 4.5-8: Asc feature importance

However, the number of living children ranked among the top two features, highlighting its significance in reflecting under-5 children mortality in Rwanda as shown in Figure 4.5.8. Conversely, variables such as sex

and level of education showed less impact on under-5 children mortality.

In developing countries like Rwanda, larger families have limited resources to allocate per child and short birth intervals, which are more common in larger families, are associated with higher risks of preterm births, low birth weight and neonatal mortality.

Regarding the top five features chosen from both SHAP values and Random Forest based on all features used to train the models, the Area Under the Curve and F1 score performance of all classifiers were essentially the same, with very little variation. When compared to the findings in Table 4.5- 24 , parameter adjustment enhanced the models' performance as shown in Figure 3.3-1, the Random Forest's performance increased significantly when all features were taken into account, with parameter tweaking reaching 99% as area under the curve as opposed to 98%.

These findings do not contradict the conclusions of another researcher who made an effort to arrive at nearly identical conclusions [Bitew et al., 2020]. However, several other research found different characteristics based on the variables they selected to take into account when estimating the death rate of children under-5 [I, 2007]. Further discussing, it is important to note that there was a meaningful correlation between SHAP values and ML algorithms, suggesting that under-5 children mortality was predominantly influenced by factors like the source of drinking water and the number of living children. Despite this, many researchers overlook the importance of the number of alive children, which has been shown to positively impact under-5 children mortality rates [JA, 2020, Molnar, 2021].

Regarding mortality in general, Rwanda's under-5 children mortality rate has decreased over time, from 228 deaths per 1000 live births in 1972 to 39.4 deaths per 1000 live births in 2021 [NISR and ICF, 2021].

This study's findings, including the identification of new influential features and the effectiveness of ML algorithms, contribute significantly to understanding and predicting under-5 children mortality in Rwanda. Specifically, the Random Forest algorithm emerged as the most effective predictor, with the number of living children being a key feature.

Chapter 5

Conclusion and recommendations

5.1 Conclusion

The evolution of new software and algorithms has brought significant advancements in machine learning intelligence, enabling efficient analysis of healthcare data. With the complexity of datasets like RDHS-2019/2020, sophisticated techniques are necessary to extract crucial information and inform decision-making. In the context of under-5 children mortality in Rwanda as per the Ministry of Health's RDHS-2019/2020 data, in contrast with other various studies that had been utilizing other different tools and variables, this study utilized machine learning, more importantly Random Forest, to predict mortality rates. Regarding variables treated, certain variables were purposedly sampled with the idea that they might be related to under-5 children mortality in Rwanda where, among 8065 cases, 14 variables were selected and, as result, 37 out of 1000 live births were found to have died before the age of 60 months. This demonstrated the improvement of above the 5% found in the previous Rwandan Demographic and Health Survey, conducted in 2014/2015.

With the use of machine learning classifier, Random Forest emerged as the best predictor of under-5 children mortality in Rwanda followed by the logistic regression. the random forest predicted the number of living children for each householder, and in terms of features ranked by SHAP model number of living children comes after "currently breastfeeding" which was ranked the top feature to consider when predicting the under-5 children mortality in Rwanda. Else, the findings highlighted key factors such as the "number of living children" and the "quality of the family's water source" as significant contributors to under-5 children mortality.

Moving forward and wrapping up, integrating machine learning algorithms like Random Forest into healthcare analyses, with a focus on variables like the number of living children and breastfeeding practices, could help reduce under-5 children mortality rates in Rwanda. Also, a significant correlation between SHAP values and machine learning algorithms was noted indicating that the two variables that most influence under-5 children mortality, namely the number of living children (more children implies that some death may occur

occasionally due to different aspects) and the dirtiness of the family's water source. This suggests that any family with a large number of difficult-to-feed children or a distant, dirtier water source will likely experience their children's death.

5.2 Recommendation

Based on the findings of this research, the Ministry of Health or healthcare systems in Rwanda are encouraged to adopt the Random Forest, or the other machine learning methods which showed a potential level of prediction of the feature since they scored above 95%, for analyzing datasets from demographic and health surveys. Emphasizing on measures which help the families in limiting births and breastfeeding practices can be instrumental in lowering under-5 children mortality rates. Additionally, this study underscores the potential of machine learning in public health research and information science, offering valuable insights for future studies. It is hoped that the findings will inspire the healthcare sector to reinforce machine learning technologies so as to improve under-5 children's survival rates, and thereby gaining a competitive edge in addressing demographic, socio-economic, parental, environmental, and epidemiological factors. Generally, researchers in Rwanda's healthcare field are encouraged to utilize machine learning algorithms and SHAP values for handling large datasets effectively, given their diverse nature and continuous generation. For the large datasets generated by the majority of health services, machine learning algorithms ought to be the most effective in general. Households are urged to continue increasing the number of prenatal appointments throughout and after their pregnancy as well as breastfeeding their infants promptly and safely. Lastly, households are warned to be careful with the source of the water they use, especially getting rid of them if dirty and looking for cleaner sources of water as this emerged among top features of importance to decline under-5 children mortality rates among Rwandan households.

Bibliography

Mariette Awad, Rahul Khanna, Mariette Awad, and Rahul Khanna. Support vector machines for classification. *Efficient learning machines: Theories, concepts, and applications for engineers and system designers*, pages 39–66, 2015.

Daniel Berrar. *Bayes' Theorem and Naive Bayes Classifier*. Encyclopedia of Bioinformatics and Computational Biology, 01 2018. ISBN 9780128096338. doi: Volume1,2019,Pages403-412.

Fikrewold H Bitew, Samuel H Nyarko, Lloyd Potter, and Corey S Sparks. Machine learning approach for predicting under-five mortality determinants in ethiopia: evidence from the 2016 ethiopian demographic and health survey. *Genus*, 76:1–16, 2020.

Bahzad Charbuty and Adnan Abdulazeez. Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends*, 2(01):20–28, 2021.

Sheetal Chaudhuri, John Larkin, Murilo Guedes, Yue Jiao, Peter Kotanko, Yuedong Wang, Len Usvyat, and Jeroen P Kooman. Predicting mortality risk in dialysis: Assessment of risk factors using traditional and advanced modeling techniques within the monitoring dialysis outcomes initiative. *Hemodialysis International*, 27(1):62–73, 2023.

Albert Dukuzumuremyi. *Machine learning based prediction of malaria outbreak using environment data in Rwanda*. PhD thesis, University of Rwanda, 2020.

Russ Greiner, B Poulin, Paul Lu, J Anvik, Z Lu, Cam Macdonell, David Wishart, Roman Eisner, and Duane Szafron. Explaining naive bayes classifications. Technical report TR03-09, Department of Computing Science, University of Alberta, Canada, 2003.

Neil Gupta, Lisa R Hirschhorn, Felix C Rwabukwisi, Peter Drobac, Felix Sayinzoga, Cathy Mugeni, Fulgence Nkikabahizi, Tatien Bucyana, Hema Magge, and Daniel M Kagabo. Causes of death and predictors of childhood mortality in rwanda: a matched case-control study using verbal social autopsy. *BMC public health*, 18:1–9, 2018.

Li-Yu Hu, Min-Wei Huang, Shih-Wen Ke, and Chih-Fong Tsai. The distance function effect on k-nearest neighbor classification for medical datasets. *SpringerPlus*, 5:1–9, 2016.

Habimana I. *Determinants of under-five mortality in Rwanda*. PhD thesis, University of Groningen, 2007.

Jomah JA. Factors affecting under-five mortality. *International Journal of Development Research*, 10(01): 33558–33561, 2020.

Rornald Muhumuza Kananura. Machine learning predictive modelling for identification of predictors of acute respiratory infection and diarrhoea in ugandaś rural and urban settings. *PLOS Global Public Health*, 2(5): e000430, 2022.

Gilles Louppe. Understanding random forests: From theory to practice. *arXiv preprint arXiv:1407.7502*, 2014.

Iqbal Madakkatel, Ang Zhou, Mark D McDonnell, and Elina Hyppönen. Combining machine learning and conventional statistical approaches for risk factor discovery in a large cohort study. *Scientific reports*, 11(1): 22997, 2021.

Emmanuel Mfateneza, Pierre Claver Rutayisire, Emmanuel Biracyaza, Sanctus Musafiri, and Willy Gasafari Mpabuka. Application of machine learning methods for predicting infant mortality in rwanda: analysis of Rwanda demographic health survey 2014–15 dataset. *BMC Pregnancy and Childbirth*, 22(1):388, 2022.

Christoph Molnar. *Interpretable Machine Learning*. 2 edition, 2021. URL `https://christophm.github.io/interpretable-ml-book`.

Gireen Naidu, Tranos Zuva, and Elias Mmbongeni Sibanda. A review of evaluation metrics in machine learning algorithms. In *Computer Science On-line Conference*, pages 15–25. Springer, 2023.

Ministry of Health (MOH)[Rwanda] NISR, National Institute of Statistics of Rwanda (NISR)[Rwanda] and ICF. Rwanda demographic and health survey 2019-20 final report. *Kigali, Rwanda and Rockville*, 2021.

World Health Organization. From mdgs, millennium development goals to sdgs, sustainable development goals. *World Health Organization*, 204, 2015.

Tomasz Rymarczyk, Edward Kozłowski, Grzegorz Kłosowski, and Konrad Niderla. Logistic regression for machine learning in process tomography. *Sensors*, 19(15):3400, 2019.

Hemo SA and Rayhan MI. Classification tree and random forest model to predict under-five malnutrition in bangladesh. *Biom Biostat Int J*, 10(3):116–123, 2021.

Rakesh Kumar Saroj, Pawan Kumar Yadav, Rajneesh Singh, and Obvious N Chilyabanyama. Machine learning algorithms for understanding the determinants of under-five mortality. *BioData mining*, 15(1):20, 2022.

Vivek V Shukla, Barry Eggleston, Namasivayam Ambalavanan, Elizabeth M McClure, Musaku Mwenechanya, Elwyn Chomba, Carl Bose, Melissa Bauserman, Antoinette Tshefu, and Shivaprasad S Goudar. Predictive modeling for perinatal mortality in resource-limited settings. *JAMA network open*, 3 (11):e2026750–e2026750, 2020.

Alaa Tharwat, Tarek Gaber, Abdelhameed Ibrahim, and Aboul Ella Hassanien. Linear discriminant analysis: A detailed tutorial. *AI communications*, 30(2):169–190, 2017.

Shichao Zhang, Xuelong Li, Ming Zong, Xiaofeng Zhu, and Ruili Wang. Efficient knn classification with different numbers of nearest neighbors. *IEEE transactions on neural networks and learning systems*, 29(5): 1774–1785, 2017.

Yupu Zhang, Jinhai Liu, Zhihang Zhang, and Junnan Huang. Prediction of daily smoking behavior based on decision tree machine learning algorithm. In *2019 IEEE 9th international conference on electronics information and emergency communication (ICEIEC)*, pages 330–333. IEEE, 2019.