AFRICAN CENTRE OF EXCELLENCE
IN DATA SCIENCE

Application of machine learning methods in analysis of infant mortality in Rwanda: Analysis of Rwanda Demographic Health Survey 2014-15 Dataset

By

Emmanuel MFATENEZA

Registration Number: **219014009**

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

University of Rwanda, College of Business and Economics

The African Centre of Excellence in Data Science (ACE-DS)

Supervisor: Dr. Claver Pierre RUTAYISIRE

September, 2020

DECLARATION

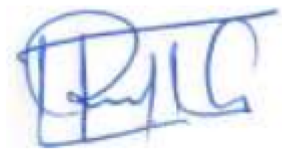
I declare that this dissertation entitled "Application of machine learning methods in analysis of infant mortality in Rwanda: Analysis of Rwanda Demographic Health Survey 2014-15 Dataset" is my original work and that to the best of my knowledge, it has not been presented for the award of a degree in any other University and all sources of materials used for this thesis have been properly acknowledged.

Student Name: Emmanuel MFATENEZA

Signature:

APPROVAL SHEET

This dissertation entitled "Application of machine learning methods in analysis of infant mortality in Rwanda: Analysis of Rwanda Demographic Health Survey 2014-15 Dataset " written and submitted by Emmanuel MFATENEZA in partial fulfilment of the requirements for the degree of Master of Science in Data Science majoring in Biostatistics is hereby accepted and approved. The rate of plagiarism tested using Turnitin is 18% which is less than 20% accepted by ACE-DS.



Dr. Pierre Claver RUTAYISIRE

Supervisor

Dr. Ignace KABANO

Head of Training

DEDICATION

I dedicated this thesis to parents, siblings and friends for their support and encouragement.

ACKNOWLEDGEMENTS

I want to express my sincere gratitude to Dr. Pierre Claver RUTAYISIRE for accepting to supervise this thesis and providing me all essential materials and moral supports that helped me to achieve my research goal. His patience, his methodological, theoretical coaching, feedback, encouragement and contributable guidance importantly enabled me to arrive to the end of this thesis. They frankly helped me in the development of this thesis into what it is today by giving constructive comments and overall guidance. I also thank for taking your time to advice without a schedule at any time when I need your hands.

I thank the African Centre of Excellence in Data Science for granting me an excellence bursary allowing therefore this research; and acknowledge the Department of Data Science at the University of Rwanda for its administrative supports.

My gratitude also goes to my parents who have supplied everything that is needed to get me to this academic level. A hallmark of my academic success has been their parental encouragement and moral support.

Special appreciation goes to the entire team of my classmates for crucial support and guidance provided to me during thesis development. Besides, I acknowledge Aimable MPAYIMANA who is the founder of Hope of Family for his crucial contribution and encouragement during my study.

Finally, I would also wish to recognize the administrative staff of University of Rwanda as well as the entire team of African Centre of Excellence in Data Science who have been supportive in one way or another throughout my stay at the University of Rwanda.

ABSTRACT

Extensive research on infant mortality (IM) exists in developing countries; however, most of the methods applied thus far relied on conventional regression analyses with limited prediction capability. Machine learning (ML) methods is new method used to provide accurate prediction of the factors associated with IM; however, there is no study conducted using ML methods in Rwanda. This study used ML methods to determine factors associated with IM and building its predictive models.

A cross-sectional study design was conducted using 2014-15 Rwanda Demographic and Health Survey. Python software was employed to apply ML methods through Logistic Regression, Random Forest, Decision Tree and Support Vector Machine. Multivariate logistic regression was employed as a traditional method. 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.

Marital status, maternal education, wealth index, sex of child and birth interval was statistically significant factors associated with infant mortality. By applying ML methods, our results revealed that random forest model was best predictive model of infant mortality with model accuracy (84.29%), recall (91.33%), precision (80.31%), F1 score (85.46%) and AUROC (84.20%); followed by decision tree model with model accuracy (83.02%), recall (90.97%), precision (78.96%), F1 score (84.67%) and AUROC(82.94%), followed by super vector machine with model accuracy (68.62%), recall (74.94%), precision(66.97%), F1 score (70.73%) and AUROC (68.55%) and last was logistic regression with low accuracy of prediction (61.49%), recall (61.05%), precision (62.15%), F1 score (61.59%) and AUROC (61.50%) compared to other predictive models.

In developing a predictive model, ML methods are used to classify certain hidden information that could not be detected by traditional statistical methods. Random forest was classified as the best classifier to be used for the predictive models of infant mortality.

Keywords*:* Infant mortality; machine learning; logistic regression; model accuracy

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ACRONYMS

AUROC	: Area under the Receiver Operating Characteristics.
EDHS	: Ethiopia Demographic and Health Surveys.
FN	: False Negative
FP	: False positive
IM	: Infant Mortality
GDP	: Gross Domestic Product
HIV/AIDS	: Human Immunodeficiency Virus/Acquired Immunodeficiency Syndrome
PTB	: Premature or Preterm Birth.
RDHS	: Rwanda Demographic Health Survey Dataset
ROC	: Receiver Operating Characteristic.
ML	: Machine Learning
SIDS	: Sudden infant death syndrome.
SSA	: Sub-Saharan Africa
TN	: True Negative
UNICEF	: United Nations Children’s Fund.
IRB	: Institutional Review Board
RNEC	: Rwanda National Ethics Committee

CHAPTER ONE

GENERAL INTRODUCTION

1.1. Background of study

Globally, there were 5.3 million infants and young teenagers in the first five years of life in 2018 and 2.3 million deaths occurred and 1.5 million deaths occurred at age 1 and 11 months, accounting for the highest number of deaths worldwide[1].

Infant mortality in Sub-Saharan Africa decreased from 182 to 58 deaths per 1000 live births from 1990 to 2017[2].Infant mortality rate is observed as essential national indicator of health because it's specifically sensitive to structural factors such as socioeconomic development and basic living conditions[3].

The national preterm birth rate was estimated at 14.9 million babies in 184 countries, of which more than one million died mostly from severe premature birth complications. Prematurity was the leading cause of death in the first month of life and was a factor in over 75% of early deaths in the neonatal period.[4].

It is possible to classify child mortality into three categories: perinatal, post-neonatal and neonatal. Late fetal death (22 weeks gestation to birth) or death of a child up to 1 week postpartum is perinatal mortality. The deaths of children aged 29 days to one year are post-neonatal mortality. Neonatal mortality is the death of newborns within 28 days of the postpartum period. Neonatal mortality is frequently due to insufficient access to essential medical services during and after pregnancy and childbirth[5]

One of the key health metrics for measuring the efficiency of the health system around the globe is the infant mortality rate. The study showed that marital status was a significant proxy indicator of post-neonatal mortality factors, such as socio-economic impact and other non-educational circumstances. Infant mortality was correlated with post-neonatal mortality, where maternal age mortality disparities were greater[6]

1.2. Rwanda’s background information

Rwanda is situated in the area of East Africa and its landlocked nation is 1,200 kilometers from the Indian Ocean and 2,000 kilometers from the Atlantic Ocean. It is bordered to the north by Uganda, to the east by Tanzania, to the west by the Democratic Republic of Congo and to the south by Burundi. It occupies a very varied landscape of 26,338 km2, ranging from the thick equatorial forest on the north-western volcanic slopes of the country to the tropical savannah to the east, along the Kagera River. The country is divided into five provinces: the Province of the West, the Province of the North, the Province of the South and East and Kigali city[7].

Since the year 2000, Rwanda has achieved sustained rates of economic growth. GDP grew at an average of 7.9 percent per year between then and 2018, and GDP per capita rose from USD 225 in 2000 to USD 787 in 2018, and the economy grew at 8.6 percent thanks to the strong performance of the agriculture sector, which grew at 6 percent, manufacturing, which grew at 10 percent, and the services sector, which grew at 9 percent in 2018[8].

Over the past 10 years, infant mortality has declined gradually, from 86 deaths per 1,000 live births in 2005 to 62 deaths per 1,000 live births in 2007-08, 50 deaths per 1,000 live births in 2010, and 32 deaths per 1,000 live births in 2014-15. Under-5 mortality also decreased during this period, from 152 deaths per 1,000 live births in 2005 to 103 per 1,000 live births in 2007-08, 76 per 1,000 live births in 2010 and 50 per 1,000 live births in 2014-15. Among other factors, reductions in infant and under-5 mortality are likely to be due to the implementation of systematic prevention of childhood diseases in all health facilities and neighborhood health systems and the introduction of new vaccines.[7]

The Ministry of Health of Rwanda stipulates different steps to improve the condition of health. This sector has earned increased government expenditure related to decentralization of the health sector by prioritizing reproduction, maternal, newborn and child health during health system policies and the incorporation of external health sectors, this combination of various factors has played an important role in improving reproduction, maternal, newborn and child health

1.3.Problem statement

Due to strong community engagement and the implementation of strong strategic management of children's diseases within health facilities, the introduction of new vaccines and the enhancement of nutrition for both mothers and their children, Rwanda has made improvements in infant mortality. Infant mortality, however, remains high relative to the Sustainable Development Goals (SDG) target of infant mortality of 12 deaths per 1000 live births.[9].

Various researchers have applied the various models and methods to analyze risk factors for infant mortality, mainly methods were logistic regression[10][11] ,survival analyses[12] and multivariate decomposition analyses[13]. Those approaches used in infant mortality research have been limited to prediction abilities and have not thoroughly clarified the factors that account for the difference in infant mortality [14]

New artificial intelligence approaches, including machine learning, present opportunities to boost mortality risk prediction and to classify risk factors for targeting particular interventions. This latest statistical application has been found to provide more detailed statistical analysis estimates and to use classification algorithms to describe the result variable in terms of the independent variables[15]. There is, however, a void in the implementation of machine-leaning techniques in Rwanda to examine infant mortality. There is no research performed in Rwanda using the techniques of machine learning to examine infant mortality. The use of machine learning techniques to evaluate infant mortality in Rwanda can help to recognize the strongest risk factors associated with infant mortality and to establish predictive models for infant mortality in Rwanda. Therefore, therefore, this research aims to apply machine learning methods based on the RDHS2014-15 dataset to analyze infant mortality in Rwanda

1.4. Objective of the study

1.4.1. General objective

The general objective was to apply machine learning methods based on the RDHS2014-15 dataset to analyze infant mortality in Rwanda.

1.4.2. Specific objectives

1. To identify risk factors that are strongest associated with infant mortality
2. To build the predictive models of infant mortality using different classification algorithms of machine learning
3. To compare the performance of different machine learning algorithms for predicting infant mortality

1.5. Research question

1. What are strongest risks factors associated with infant mortality?
2. What is performance of predictive models of infant mortality using different classification algorithms of machine learning?
3. Is any predictive model among machine learning algorithm performing better than others in predicting infant mortality?

1.6. Significance of the Study

The machine learning methods were applied in assessing infant mortality and its risk factor based on the RDHS 2014-15 dataset. The study provided information on risk factors for infant mortality and the best predictive model in Rwanda for predicting the risk of infant mortality. The Ministry of Health and other stakeholders could use the results of this study to plan and implement child health intervention programs in Rwanda in order to reduce infant mortality. Furthermore, the results of the study may help other researchers recognize the current evidence on infant mortality.

1.7. Scope and Limitation of the Study

The study focused on the prediction of the risk of infant mortality in Rwanda by applying machine learning algorithms based on the feature attributes selected from the All Births File (BR) RDHS2014-15 dataset. Four machine learning approaches, specifically logistic regression, Random forest, Decision tree and Support Vector Machine algorithms, were used to develop the predictive models of infant mortality. The performance of the predictive models was evaluated using evaluation metrics that were used to verify the performance of the predictive model, namely confusion matrix, accuracy, precision, recall, F1 score and Area under the Receiver Operating Characteristics (AUROC). The analysis did not take into account variables that in RDHS2014-15 had several missing values, and multivariate logistic regression was used as an example of conventional statistical methods.

1.8. Ethical consideration

In this study, Researcher used the secondary data analysis that was carried out after receiving ethical approval from the website of Demographic Health Survey.

1.9. Organization of study

This research is arranged into five chapters. The first chapter deals with the research history, aims and significance of the analysis. The related literature on the use of machine learning in infant mortality is discussed in the second chapter. The third chapter deals with the methods of research used to create a predictive infant mortality model. The fourth chapter deals with presentation and discussion of outcomes. The general conclusion and recommendation are provided in Chapter Five.

CHAPTER TWO

LITERATURE REVIEW

2.1. Theoretical review

Infant mortality is the number of deaths of children under one year of age per 1,000 live births. It is an important measure of the well-being of infants, girls and pregnant women. It is connected in any given geographical area to a number of factors, including women's health status, the quality of and access to medical care services, socio-economic conditions, and public health practices[16]. Infant mortality refers to the number of deaths below the age of one newborn infant. Newborn deaths are related to the probability of newborns prior to their first birthday [17].

Kotsadam et al. (2018) argued that living for newborns in the first year and beyond depends on various factors such as maternal health, diseases, access to immunizations, safe drinking water, access to food and sanitation, all of which are paramount to reducing infant and maternal mortality rates for a more successful quality of life and economic productivity [18].

The way to estimate the determinants of child health based on the role of health output and how households relate to health inputs and child health was suggested in his theoretical framework. Among other endogenous household variables that can be changed by economic and demographic choices, he showed how child survival can be perceived. Child survival, such as place of residence, public program, price, salaries, desires, marital status, individual biological and economic endowment, can be interpreted as directly conditioned on exogenous constraints and environmental conditions[19].

In their theoretical framework, Mosley and Chen (1984) categorized the determinants of infant mortality into five categories, such as environmental pollution, accidents, maternal causes, and treatment causes for nutritional status and infant disease. They find that 97% of born children are projected to live up to five years of age, but the influence of socioeconomic, environmental and biological factors is the driving force behind the decrease in probability of survival. The combined sequence of biological markers of proximate determinants of disease and nutritional deficiencies [20].

2.2. Conceptual Framework of infant mortality

Conceptually, one of the most sensitive and commonly used indicators of a population's social and economic growth is the infant mortality rate as defined as the likelihood of a living born child dying before its first birthday. The empirical structure of child mortality was studied by Mosley and Chen (1984), Stressing that differences in infant and child mortality are explained by environmental variables that are socio-economic, bio-demographic and household [20]. In this study, the below conceptual framework of infant mortality was adjusted based on analytical framework of Mosley and Chen (1984).

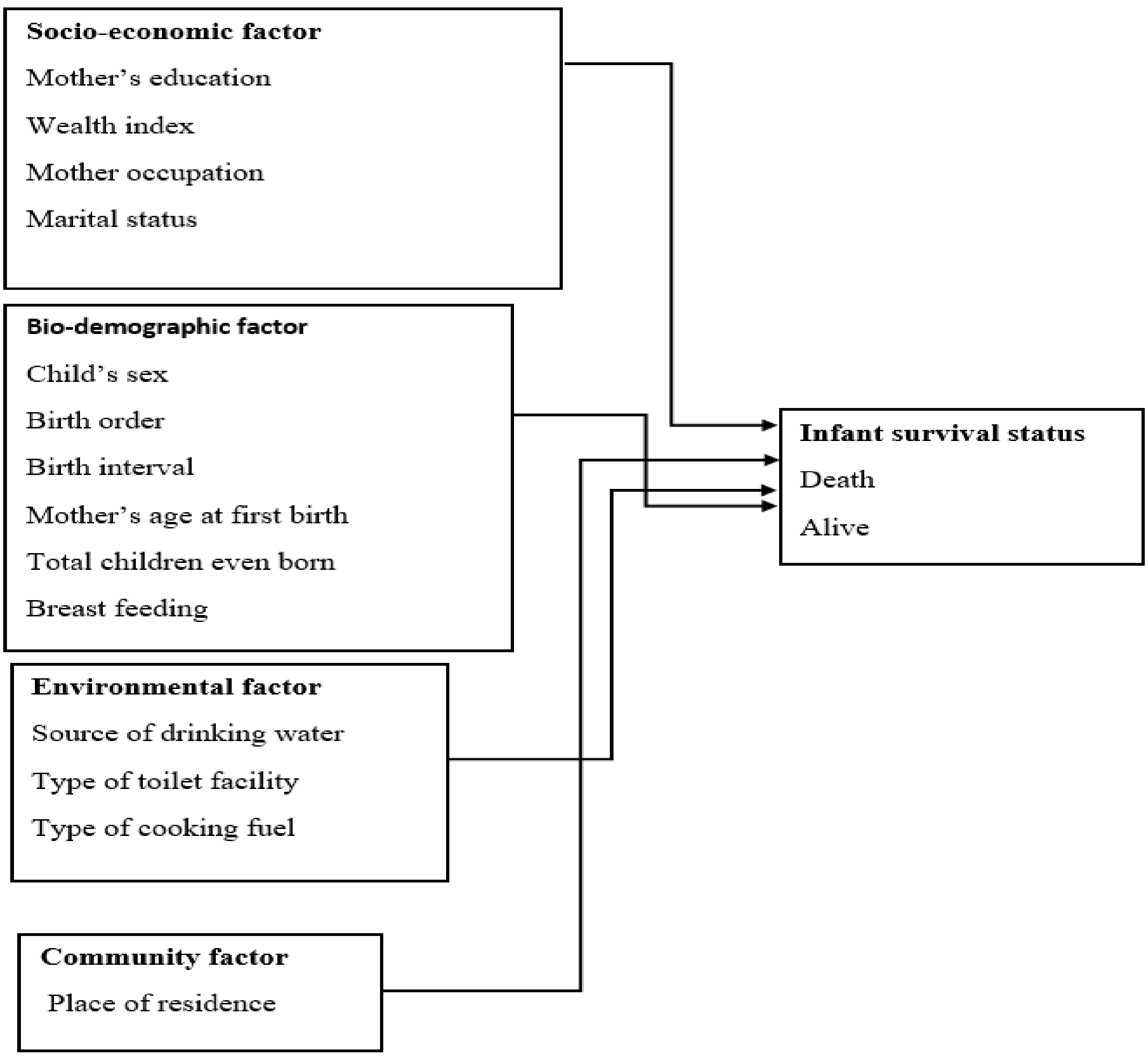


Figure 2.1: Conceptual framework

2.3. The determinants of infant mortality

The previous study showed that survival analysis was used in Kenya to model child survival and found that modern contraceptives are likely to be used by qualified women because this has a major impact on reducing the risk of mortality and high mortality rates in low-living households. Kids living in rural areas are more likely to die because they are more vulnerable to poverty than urban kids [21].

Ntenda et al.(2014) found that providing health services , improving maternal care, neonatal care, improving the quality of life in rural areas and improving food intake could reduce infant mortality in Malawi[22]. Socio-economic variables such as birth rate, national income, labor force for women, healthcare spending and literacy rates for women have been found to affect infant mortality [23].

The major proximate determinants for both infant and child mortality are bio-demographic variables such as marital status, birth order, form of birth and preceding birth interval. The socio-economic determinants of schooling, household size and household sex were also found to be the most significant determinants of infant and child mortality [24].The variety of socio-economic and demographic factors affect infant mortality, such as child sex, mother's age at first birth, birth order, pre-birth interval, among others. [25].

The variables affecting child mortality at the national level were the availability of health care, medical practitioners, poverty, fiscal implications, public sector health policy, national literacy status, population welfare programs and fertility. Factors influencing child mortality at the household level have been found to be linked to household activity and ability, such as socioeconomic status, family cultural and social values, parents ' educational level, fertility preferences, and socioeconomic autonomy of women[26].

Kong et al,.(2016) identified several predictors of infant mortality and morbidity, finding several pre-term birth factors, apart from gestation and birth weight, that could be associated with risks of high mortality and morbidity outcomes [27]. Prematurity factors, congenital causes, injury, other illnesses, child illnesses, maternal disorders, Sudden Infant Death Syndrome, lack of oxygen during delivery have been shown to cause infant mortality[28]

In ECO countries, between 2005 and 2012, Rezaei et al, (2015) studied the key factors influencing infant mortality. He concluded that GDP per capita was the key determinant of child mortality, public expenditure as a proportion of overall health expenditures and total fertility rate[6]. Household-level factors influencing infant mortality are related to the behavior and ability of families, such as socioeconomic status, family cultural and social values, parents ' educational level, fertility preferences, and socioeconomic autonomy of women[26].

While researchers have devoted significant attention to the influence of individual-level variables on infant mortality, less is understood about how group characteristics impact children's health outcomes, even though they play a prominent role in theoretical models. This research uses multivariate and multilevel discrete-time event history analysis to systematically analyze the effect of contextual variables on the risk of dying before the age of five, and their relative significance in relation to individual variables, using data from the latest round of Demographic Health Surveys (DHS) for all countries in sub-Saharan Africa. The findings suggest that some of the community's features affect children's mortality risks, beyond the intermediate factors used in this investigation.[29].

2.4. Empirical review of study

Data mining techniques such as Decision tree, Random Forest, Help Vector Machine and Naïve Bayes algorithms were used to predict infant and child mortality based on 10,641 records from the Ethiopia Demographic and Health Survey dataset and the results showed that the random forest is a good classifier compared to others with 96.74 percent accuracy, 79.53 percent average accuracy in imbalanced train data[30]

In classifying infant mortality rate, life expectancy at birth, annual population growth, and gross domestic product, the decision tree algorithm was applied, and the results showed that annual population growth is strongly correlated with the prediction of child mortality. With 97.4 percent ROC curve result of the three classes, the model produced has high acceptability in predicting child mortality under five years of age[31].

In particular, Random forest and Naive Bayes machine learning techniques were used to prevent infant mortality in the Brazilian Northeast and the results showed that the Naive Bayes classifier has better performance than the other predictive classifiers with 98.2 percent accuracy performance and 92.1 percent receiver operating characteristic (ROC) region[32]. In order to build a web-based child mortality prediction model based on EDHS 2011 data, the classifier decision tree and rule induction (using the PART algorithm) were applied and the results found that breastfeeding, maternal education, pre-birth interval, low birth weight, family planning, paternal education, mother's age at first birth and diarrhea were correlated with child mortality. With 94.3 percent accuracy, 93.8 percent sensitivity, 94.3 percent precision and 94.8 percent region under ROC, the decision tree model had better results[33].

Based on n community-based epidemiological data sets, neural network and decision tree data mining techniques were used to predict the risk of child mortality and results showed that child mortality was correlated with the climate, household literacy, household health , child age, household windows, household water, and even household ethnicity [34]. In addition, machine learning techniques have been used to build models that estimate the probability of adverse birth outcomes by assessing associated risk factors as a function of the available data. The Self Organization Map algorithm has been found to show that child deaths are associated with low birth weight, preterm birth, access to prenatal care and other variables [35].

CHAPTER THREE

RESEARCH METHODOLOGY

3.1. Research design

This study used a cross-sectional study design where machine learning methods were used to construct predictive models of infant mortality in Rwanda based on the RDHS 2014-15 dataset, namely logistic, decision tree, random forest and support vector machine. Computer leaning techniques are typically versatile, non-parametric methods for data predictions or classifications, using algorithms to detect data patterns using variable and model selection techniques[36]. By first conducting data analysis, the methods construct analytical models to find hidden insights and use algorithms that iteratively learn from historical data and help predict unknown data and provide the advantages over statistical methods used for forecasts[37]. The algorithm that describes how the predictions are made using the raw data is typically defined and can allow for a larger number of predictors and high-dimensional data. The basic principle of creating a model that is capable of making predictions is predictive modelling, in which such a model involves a machine learning algorithm that seeks unique properties to obtain certain predictions from a training and testing dataset.

3.2. Source of data

The data set used for the study comes from the Rwanda Demographic and Health Survey (RDHS 2014-15), conducted in Rwanda as a cross-sectional, nationally representative survey. In order to obtain existing demographic and health measures, including family planning, maternal mortality, infant and child mortality, maternal and child nutrition status, prenatal care, delivery and postnatal care, childhood diseases and paediatric immunization, this national survey was performed.
Domestic violence indicators, the prevalence of malaria and anaemia among women and children and the prevalence of HIV infection in Rwanda have also been assessed. The target population was females aged 15-49 years from sampled households. The RDHS2014-15 data collection fieldwork was conducted from 9th November 2014 to 8th April 2015.

All 492 of the selected clusters were surveyed and the Household Questionnaire was completed by all 30,058 households, including 6,069 (20 percent) urban and 23,989 (80 percent) rural. The composite sample analysis was used in the surveys to achieve reliable estimates of standard errors and confidence intervals, which compensated for the sampling weight due to multi-stage stratified sampling.

In this analysis, the RDHS2014-15 dataset of all birth files (BR) obtained using the Woman's Questionnaire and collected data on mother's birth histories was used[7]

3.3. Study variables

The All Births File (BR) RDHS 2014-15 dataset contained 1078 variables with 30058 records. The variables were chosen in this analysis based on the study's current literature review. Infant mortality was the dependent variable in this analysis, which was described among all children ever born as the number of deaths occurring during the first year of life. The independent variables were the variables correlated with infant mortality, and 15 variables were reported to be included in this analysis, as shown in Table 3.1.

Table 3.1: List of variables selected to be included in study

Names of variable	Variable coding	Type of variable
Infant Survival Status	1=Death 2=Alive	Dependent
Place of Residence	1=Urban 2=Rural	Independent
Marital status	1=Single 2=Married 3= Living with partner 4=Widowed 5=Divorced/separated	Independent
Maternal Education	1= No formal education 2=Primary 3= Secondary and over	Independent
Mother Occupation	1=Working 2=Not working	Independent
Wealth Index	1=Low 2=Middle 3=High	Independent

Maternal age at first birth	1= Below 20year old 2=20-34 years old 3=35 years old and over	Independent
Sex of child	1=Male 2=Female	Independent
Birth order	1=1 or 2 birth 2= 3 or 4 birth 3=5 births and over	Independent
Birth interval	1= Less 24 month 2=24 months and over 3= First births	Independent
Children ever born	1=1-3children 2=4-6 children 3= Over 6 children	Independent
Breastfeeding	1=Yes 2=No	Independent
Source of drinking water	1= Improved 2=Not improved	Independent
Type of toilet Facility	1= Improved 2= Not improved	Independent
Type of cook fuel	1= Improved 2= Not improved	Independent

3.4 .Study sample size

The total sample size of this study was 30058 babies born from women aged 15-49 years, based on the RDHS2014-15 dataset of all births file (BR). This research included all births. The overall number of child deaths was 1952, while 28,106 were live births.

3.5. Data Pre-processing

Data processing is a technique for data mining which transforms raw data into a comprehensible format. Raw data (real world data) is often incomplete and data cannot be sent through a model.

In order to produce the best performance, data preparation is necessary for what is done in machine learning algorithms on data[38]. The missing values were substituted by the media and they did not include those that had many missing values in this study.

3.6. Features selection

The method of reducing the number of variables during the creation of a predictive model is feature selection. In data pre-processing, the methods of feature selection may be used to achieve successful data reduction. In this analysis, the Random Forest Algorithm was used to determine the best features contributing to infant mortality.

3.7. Imbalanced Data Handling

The imbalanced data is a special case for the classification problem where the class distribution between the classes is not standardized. In this analysis, there was disproportion in the binary classification of infant survival variable (Death / Alive). The number of infant mortality was 1952 deaths whereas there were 28,106 births in living births. We noticed that the dependent variable has an imbalance of "Death" and "Alive" proportions. To balance the distribution of groups, the Random oversampling approach was used. This is the non-heuristic form of re-balancing class allocation by randomly replicating instances in the positive class[39].Random Oversampling involves supplementing the training data with multiple copies of some of the minority classes. In this study, random over sampling was applied to increase the number of infant mortality in to balance with children alive.

3.8. Methods of analysis

3.8.1. Multivariate analysis

Multivariate logistic regression analysis was used to examine the relationship between independent variables and infant mortality at a meaningful level of 5% and a confidence interval of 95%. The risk factors associated with infant mortality have also been used to classify.

3.8.2 Methods of Building Predictive Models

The basic principle of creating a model that is capable of making predictions is predictive modelling, in which such a model involves a machine learning algorithm that seeks unique properties to obtain certain predictions from a training and testing dataset.

Machine learning algorithms, namely Logistic regression, Decision tree, Random Forest and Support Vector Machine algorithms were used in this study to construct a predictive model for infant mortality based on the RDHS2014-15 dataset called all births (BR) register. These algorithms were suitable for this study because the binary answer was the dependent variable infant survival status (Death/ Alive). The dataset was split into two elements, such as training and research datasets, 80% of which were for training datasets and 20% for test datasets.

3.8.2.1. Logistic regression

Logistic regression is a classification algorithm used to assign a discrete set of classes to observations and is a method borrowed from the field of statistics through machine learning. For binary classification issues (problems with two class values), it is the go-to technique. It is a classification function that uses class for construction and with a single estimator uses a single multinomial logistic regression model. Using the logistic sigmoid equation, the logistic regression transforms its output to return a probability value that can then be mapped to two or more different groups. It is a particular case of generalized linear modelling, also called a logistic model or logit model, and is commonly used in medical, social sciences, marketing applications for binary classification methods[40]. The logistic regression algorithm was used in this analysis to construct the logistic predictive model of infant mortality

3.8.2.2. Decision Trees

The Decision tree is a type of algorithm for supervised learning that is often used in problems with classification. It works for input and output variables, both categorical and continuous. It is one of the methods to predictive modelling used in machine learning. As a predictive model, it utilizes a decision tree to go from assumptions about an item (represented in the branches) to conclusions about the target value of the item (represented in the leaves).

The decision tree is a structure of a tree that classifies an input sample into one of its possible classes and utilizes the vast amount of information available by making decision rules to extract knowledge. Each node in a decision tree represents a role in an instance to be classified, and each branch represents a value the node may assume. Instances are listed and sorted, starting at the root node, based on their feature values. Post-pruning methods are widely used by decision tree classifiers to test the efficacy of decision trees, as a validation array is used to prune them. The most popular class of training instances that are sorted to it can be removed from any node and assigned [41]. One of the simplest data structures to understand in machine learning is the decision tree. Rules are first extracted from the training data set to form the decision tree that is then used for the research dataset classification. This research used the Decision Tree algorithm to construct a predictive model for infant mortality in the decision tree.

3.8.2.3. Random forest

Random forest is a sequence of decision trees and provided with some controlled alteration independently. This algorithm was based on classification trees, an ensemble learning technique that can solve both kinds of problems. The trees and the results of the Random Forest are based on the majority of precise development. Random forest is an ensemble classifier that can be used to address classification and regression problems, such as decision trees. It utilizes the idea of creating multiple random trees with training dataset bootstrap, sample bagging, voting scheme and randomly selected features in each decision split, which enhances the predictive ability and results in greater performance[42]. When a large proportion of the data is incomplete, it is useful to deal with missing values, outliers and preserve accuracy. Among common machine learning techniques, the model interpretability and prediction precision given by Random Forest is very special. [43].This research used the Random Forest Algorithm to create a predictive model of infant mortality in the Random Forest.

3.8.2.4. Support vector Machine

Support vector machines (SVM) are implemented by Cortes and are based on statistical learning and are an efficient tool for binary classification, regression or ranking functions. Due to many attractive features, it is very common to use the health care researcher for classification, handling complex non-linear data points. Its accuracy is good and less likely than other well-known classifiers to be more suitable than other classifier[44]. It is a method for prediction of classification and regression that uses machine learning theory to optimize predictive accuracy while preventing over-fit to the data automatically. Support Vector machines can be defined as systems that use a linear function hypothesis space in a high dimensional feature space, trained with an optimization theory learning algorithm that implements a learning bias derived from statistical learning theory [45]. This classifier was used in this analysis to construct a predictive model of infant mortality.

3.9. Methods of Performance Evaluation for Predictive Models

In this study, to determine the best model to predict infant mortality, the performance of the predictive models was evaluated. During the model evaluation phase, evaluation metrics such as confusion matrix, accuracy, precision, recall, F1 score and Area Under the Receiver Operating Characteristics (AUROC) were used in this study

.3.9.1. Confusion Matrix

The confusion matrix is a matrix of N X N, where N is the number of predicted classes and displays the number of correct and incorrect predictions made by the classification model relative to the data's actual results (target value). It is a method used to evaluate a classification model 's output that is recognized on a collection of test data for which the true values are known in table format. A fast understanding of model accuracy, precision, recall and F1 scores for predictive model construction is given by the confusion matrix. All data instances of a test dataset are predicted as positive or negative by the binary classifier[46] . The following table displays a 2×2 matrix for two classes (Positive and Negative).

Table 3.2: Confusion matrix

	Predicted values		
Actual values		**Positive (1)**	**Negative (0)**
	Positive (1)	TP	FP
	Negative (0)	FN	TN

- **True positive (TP)**: This shows that a model correctly predicted Positive cases as Positive.
- **False positive (FP)**: This shows that a model incorrectly predicted Negative cases as Positive.
- **False Negative (FN):** This shows that an incorrectly model predicted Positive cases as Negative.
- **True Negative (TN):** This shows that a model correctly predicted Negative cases as Positive

3.9.2. Accuracy

Accuracy is the percentage of true outcomes among the total number of cases tested. In this analysis, it was used to determine model efficiency and measure from the confusion matrix.

Accuracy = TP + TN/TP+TN+FP+FN

3.9.3. Precision

It is the number of right positive outcomes divided by the number of positive outcomes that the classifier predicts. In this analysis, it was used to test model output and it was estimated from the confusion matrix.

Precision=TP/TP+FP

3.9.4. Recall

The number of accurate positive outcomes, separated by the number of all related samples, is recalled. In this analysis, it was used to determine model efficiency and measure from the confusion matrix.

Recall=TP/TP+FN

3.9.5. F1 score

The inverse relationship between accuracy and recall is the F1 score or the F test. The harmonic mean of recall and accuracy is determined as

F1 score=2TP/2TP+FN+FP

3.9.6. Area under curve

One of the most commonly used metrics for the predictive model of assessment is Region under Curve (AUC). It is used for problems related to binary classification. The Field Under the Receiver Operating Characteristics (AUROC) shows how well the probabilities are segregated from the negative classes by the positive classes. Regardless of what classification threshold is selected, it tests the consistency of the model 's predictions and is an output measurement for the classification problem at different thresholds. When it has a value close to 1, the predictive model indicates goodness, while the value near 0 indicates bad model efficiency. In this analysis, this assessment metric was used to assess the efficiency of the predictive model.

CHAPTER FOUR

RESULTS PRESENTATION AND DISCUSSION

4.1. Introduction

This chapter presents the results were obtained by applying machine learning methods in analysis of infant mortality in Rwanda based on RDHS2014-15 dataset. The three machine learning methods such as decision tree, random forest and super vector machine algorithms were to build predictive models of infant mortality. The multivariate logistic regression was used in this study as example traditional statistical methods used to build model. The evaluation metrics methods such as confusion matrix and the area under the receiver operating characteristic (AUROC) were used to evaluated the performance of predictive models of infant mortality. The discussion of results was introduced in this chapter.

4.2. The Characteristics of Study Population

Around 6.5% of infants died before their first birthday out of the 30058 infants of the sample. Majority (79.8%) of infants were from rural area. The majority 58.9% of infants born from married women, 20.7% born to women lived with their partners,8.4% born to divorced/separated women, 8.2%born to widowed women and only 3.8% born to single women. The majority 67.9% of infants were born to mothers completed primary level, 22.5% born to mothers who had no formal education and 9.6% born from mothers who have completed at least secondary level. Over 95.0% of infants born to mothers were working while 5.0% of them born to mothers who were not working. The wealth index difference among families which participated in the survey was 42% low, 20% middle, and 38.1% high. The majority 67.3% of the infants were born to mothers aged between 20 and 34 years compared to 32.5% and 0.2% of infants born respectively to mothers aged below 20-year-old and 35 years old and over. Majority (50.6%) of the infants were males. The majority (51.7%) of infants were born to mothers had 1 or 2 births followed by those who were born to mothers had 3-4 births (28.6%). Most (52.3%) of the births occurred to mothers whose preceding birth interval was 24 months and above compared to 18.2% of births occurring to mothers with preceding birth intervals of less than 24 months. Majority of infants (42.2%) were born to mothers who had between 4 and 6 children. The majority (60.5%) of births were breastfed. Majority (60.7%) were from the families that had from un-unimproved sources of water. Over

95.3% of infants occurred in households with improved toilet facility while. More than 86% of infants were from the household with improved types of cook fuel (**Table 4. 3**).

Table 4. 3: Characteristics of study population

Characteristics	Frequency(N)	Percentage(%)
Infant Survival Status		
Death	1952	6.5
Alive	28106	93.5
Place of Residence		
Urban	6069	20.2
Rural	23989	79.8
Marital status		
Single	1146	3.8
Married	17696	58.9
Living with partner	6215	20.7
Widowed	2467	8.2
Divorced/separated	2534	8.4
Maternal Education		
No formal education	6778	22.5
Primary	20409	67.9
Secondary and over	2871	9.6
Mother Occupation		
Working	28546	95.0
Not working	1512	5.0
Wealth Index		
Low	12610	42.0
Middle	6008	20.0
High	11440	38.1
Maternal age at first birth		
Below 20year old	9770	32.5
20-34 years old	20217	67.3

35 years old and over	71	0.2
Sex of child		
Male	15216	50.6
Female	14842	49.4
Birth order		
1 or 2 birth	15532	51.7
3 or 4 birth	8601	28.6
5 births and over	5925	19.7
Birth interval		
Less 24 month	5479	18.2
24 months and over	15720	52.3
First births	8859	29.5
Birth interval	5479	18.2
Children ever born		
1-3children	9701	32.3
4-6 children	12698	42.2
Over 6 children	7659	25.5
Breastfeeding		
Yes	11862	39.5
No	18196	60.5
Source of drinking water		
Improved	11802	39.3
Not improved	18256	60.7
Type of toilet Facility		
Improved	28638	95.3
Not improved	1420	4.7
Type of cook fuel		
Improved	4146	13.8
Not improved	25912	86.2

4.3. Bivariate logistic regression for the associated factors of infant mortality

The results showed that place of residence ($x^2 = 24.06, p < 001$), marital status ($x^2 = 34.19, p < 001$), maternal education ($x^2 = 95, p < 001$), wealth index ($x^2 = 42.2, p < 001$), mother's age at first birth ($x^2 = 39.67, p < 001$), sex of child ($x^2 = 16.92, p < 001$), birth interval ($x^2 = 249.6, p < 001$), children ever born ($x^2 = 325.87, p < 001$), breastfeeding ($x^2 = 79.62, p < 001$), source of drinking water ($x^2 = 11.4, p < 001$), type of toilet facility ($x^2 = 15.59, p < 001$), and type of cook fuel ($x^2 = 29.67, p < 001$) were significantly associated with infant mortality. Infant deaths were relatively higher in rural 6.8% than in urban areas (5.1%). The divorced mothers had higher rate of infant mortality (8.2%) than other categories of mother's marital status. Mother with no formal education reported the highest level of infant mortality (8.6%) when compared to other categories of education. Those with low wealth index reported the highest number of deaths of their infants with 7.4% in relation compared to 6.7% and 5.4% of infant deaths distributed to middle and high wealth index respectively. Mothers who had their first birth below 20 years registered more deaths with 7.8% compared to 5.9% and 4.2% of infant deaths were registered with mothers aged 20-34 years and 35 years or over respectively. Infant mortality was highly prevalent in male children (7.1%). The preceding birth interval of less than 24 months was associated with higher risks of infant deaths with 10.8% than those with an interval of 24 months and above with 4.7 %. It was observed that mothers with above 6 children experienced the highest number of deaths 10.3% while 3.5% and 6.5% of infant deaths observed to mother had between one child or 3 children and 4 children or 6 children respectively. Infant deaths was higher associated in households did not provide breast feeding to their children (7.5%) than their counterparts. Mothers whose household had unimproved source of drinking water had presented more infant deaths with 6.9% than those in households that access improved source of drinking water. More deaths were reported in households with no improved toilet facility with 9.0% compared to households with improved toilets facility with 6.4% .

Table 4.4: Prevalence of infant mortality by independent variables and its associations

Characteristics	Infants deaths N(%)	Live births N(%)	Chi-square	p-value
Place of residence				
Urban	310(5.1%)	5759(94.9%)	24.06	<0.001**
Rural	1642(6.8%)	22347(93.2%)		
Marital status				
Single	61(5.3%)	1085(94.7%)	34.19	<0.001**
Married	1053(6.0%)	16643(94.0%)		
Living with partner	434(7.0%)	5781(93.0%)		
Widowed	196(7.9%)	2271(92.1%)		
Divorced/separated	208(8.2%)	2326(91.8%)		
Maternal education				
No formal education	583(8.6%)	6195(91.4%)	95	<0.001**
Primary	1269(6.2%)	19140(93.8%)		
Secondary and over	100(3.5%)	2771(96.5%)		
Mother occupation				
Working	1873(6.6%)	26673(93.4%)	4.22	0.40
Not working	79(5.2%)	1433(94.8%)		
Wealth Index				
Low	934(7.4%)	1176(92.6%)	42.2	<0.001**
Middle	405(6.7%)	5603(93.3%)		
High	613(5.4%)	10827(94.6%)		
Age of mother at first birth				
Below 20year old	760(7.8%)	9010(92.2%)	39.67	<0.001**
20-34 years old	1189(5.9%)	19028(94.1%)		
35 years old and over	3(4.2%)	68(95.8%)		
Sex of child				
Male	1076(7.1%)	14140(92.9%)	16.92	<0.001**
Female	876(5.9%)	13966(94.1%)		

Birth order				
1 or 2 birth	999(6.4%)	14533(93.6%)	229	0.892
3 or 4 birth	562(6.5%)	8039(93.4%)		
5 births and over	391(6.6%)	5534(93.4%)		
Birth interval				
Less 24 month	592(10.8%)	4887(89.2%)	249.6	<0.001**
24 months and over	746(4.7%)	14974(95.3%)		
First births	614(6.9%)	8245(93.1%)		
Children ever born				
1-3children	336(3.5%)	9365(96.5%)	325.87	<0.001**
4-6 children	830(6.5%)	11868(93.5%)		
Over 6 children	786(10.3%)	6873(89.7%)		
Breastfeeding				
Yes	584(4.9%)	11278(95.1%)	79.62	<0.001**
No	1368(7.5%)	16828(92.5%)		
Source of drinking water				
Improved	696(5.9%)	11106(94.1%)	11.4	<0.001**
Not improved	1256(6.9%)	17000(93.1%)		
Type of toilet Facility				
Improved	1824(6.4%)	26814(93.6%)	15.59	<0.001**
Not improved	128(9.0%)	1292(91.0%)		
Type of cook fuel				
Improved	189(4.6%)	3957(95.4%)	29.67	<0.001**
Not improved	1763(6.8%)	24149(93.2%)		

***Notes**: ** Statistical significance at p<0.001*

4.4 .Multivariate analysis of risks factors for infant mortality

Multivariate analysis of risk factors for infant mortality was performed by using a multivariate logistic regression models based full and reduce models. The results showed that marital status, maternal education, wealth index, sex of child, birth interval, children ever born, breastfeeding and type of toilet facility were the risk factor of infant mortality. The infants whose mothers were married [aOR=0.61, 95% CI (0.46-0.81),p<0.001] and widowed [aOR=0.7;95%CI(0.51-0.96), p<0.001] had less odds to die than those whose mothers were single. The infants whose mothers studied secondary and over had a 0.63 lower risk to die [aOR=0.63, 95% CI (0.50-0.80),p<0.001] than the infants whose mothers had no formal education. The births from households with a high wealth index had 0.77 lower risk of dying [aOR=0.77, 95% CI (0.69-0.87), p<0.001] than those births born from households with a low wealth index. Female infants had lower likelihoods of dying [aOR=0.80; 95%CI (0.73-0.88),p<0.001] than male infants. The infants born to mothers who had 5^{th} born or more were 0.67 times [aOR=0.67; 95% CI (0.57-0.79), p<0.001] less likely to die compared to infants born to mothers who had less than 3 children. The infants from mothers whose preceding birth interval was 24 months and above [aOR=0.46; 95% CI (0.41, 0.51),p<001] and those who had first births [aOR=0.86; 95% CI(0.74-1),p=0.047] had lower odds of dying of births than those from the mothers with preceding birth intervals of less than 24 months [aOR=0.46; 95% CI(0.41, 0.51),p<001]. The infants who were not breastfeed were 1.45 times [aOR=1.45; 95% CI (1.31-1.61),p<0.001] more likely to die when compared to those who were breastfed. The infants born in families with no improved toilet facility were 1.35 times more prone to die than those who from the families with improved toilet facility [aOR= 1.35, 95% CI (1.11-1.65),p=0.003] (**Table 4. *5***).

Table 4. 5: Multivariate analysis of risk factors for infant mortality

Variables	Full model				Reduced model			
	aOR	95 % CI		*p-value*	aOR	95 % CI		*p-value*
		Lower	Upper			Lower	Upper	
Place of Residence								
Urban	1							
Rural	1.10	0.94	1.29	0.243				
Marital status								
Single	1				1			
Married	0.62	0.46	0.82	<0.001**	0.61	0.46	0.81	<0.001**
Living with partner	0.83	0.62	1.11	0.202	0.82	0.61	1.10	0.181
Widowed	0.70	0.51	0.96	0.028*	0.70	0.51	0.96	0.027*
Divorced/separated	0.88	0.64	1.20	0.408	0.87	0.64	1.18	0.366
Maternal Education								
No formal education	1				1			
Primary	0.91	0.82	1.02	0.093	0.91	0.81	1.01	0.071
Secondary and over	0.66	0.52	0.85	<0.001**	0.63	0.50	0.80	<0.001**
Mother Occupation								
Working	1							
Not working	0.87	0.69	1.11	0.268				
Wealth Index								
Low	1				1			
Middle	0.89	0.79	1.01	0.076	0.89	0.79	1.01	0.078
High	0.80	0.70	0.90	<0.001**	0.77	0.69	0.87	<0.001**
Age of mother at first birth								
Below 20year old	1							
20-34 years old	0.91	0.83	1.01	0.066				
35 years old and over	1.16	0.36	3.74	0.802				
Sex of child								
Male	1				1			

Female	0.80	0.73	0.88	<0.001**	0.80	0.73	0.88	<0.001**
Birth order								
1 or 2 birth	1				1			
3 or 4 birth	0.93	0.81	1.07	0.299	0.93	0.81	1.07	0.297
5 births and over	0.67	0.57	0.79	<0.001**	0.67	0.57	0.79	<0.001**
Birth interval								
Less 24 month	1				1			
24 months and over	0.45	0.40	0.51	<0.001**	0.46	0.41	0.51	<0.001**
First births	0.86	0.74	0.99	0.042*	0.86	0.74	1.00	0.047*
Children ever born								
1-3children	1				1			
4-6 children	2.34	2.02	2.70	<0.001**	2.37	2.05	2.73	<0.001**
Over 6 children	4.15	3.53	4.88	<0.001**	4.26	3.63	4.99	<0.001**
Breastfeeding								
Yes	1				1			
No	1.45	1.31	1.61	<0.001**	1.45	1.31	1.61	<0.001**
Source of drinking water								
Improved	1							
Not improved	1.04	0.94	1.16	0.433				
Type of toilet Facility								
Improved	1				1			
Not improved	1.34	1.10	1.63	0.003*	1.35	1.11	1.65	0.003*
Type of cook fuel								
Improved	1							
Not improved	0.98	0.80	1.21	0.862				

Notes: aOR**: Adjusted odds ratio; **CI**: Confidence intervals; ** statistical significance levels at 0.05; ****** high statistical significance level at 0.001*

4.5. Importance features selection

Random forest classifier was used to identify important features that are associated with infant mortality as it shown in **Figure 2**. It indicated the top 10 best features contribute to infant mortality were Marital status, children ever born ,birth order, wealth index, preceding birth interval ,maternal education, source of drinking water, sex of child, mother age at first birth and breast feeding.

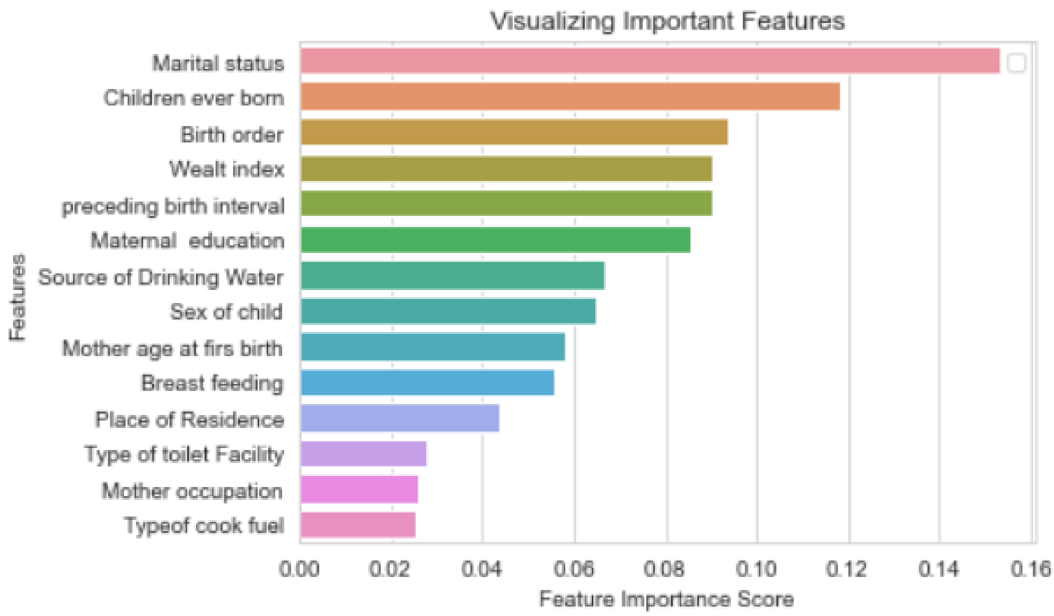


Figure 2: Importance features

4.6 .Predicting Infant mortality

The ML methods approach models namely Logistic Regression, Random Forests, Decision Tree, and the Support Vector Machine classifiers were applied to build a predictive model of IM. All predictive models of IM were trained on training data of 80% and tested on a test dataset of 20%. The performance predictive models were evaluated and compared using evaluation metrics namely Confusion matrix, Accuracy, Precision, Recall and F1 score, and Area Under receiver operating characteristics AUROC. Our results showed that the logistic regression models predicted correctly 3442 infants died before completing their first year while 3472 infants were still alive. It has wrongly predicted 2114 births died before completing their first years and 2215 births were still alive. Our findings found that logistic regression model has generally predicted IM at 61.4 % of accuracy with recall (61.0%), precision (62.1%), F1 score (61.5 %), and AUROC (61.5%). Random forest model was predicted correctly 4283 infants as died and 5194 infants as alive. It has wrongly predicted 1273 infants as died and 493 births as still alive. It was found that the random forest model has generally predicted infant deaths correctly at 84.2% of accuracy with recall (91.3%), precision (80.3%), F1 score (85.4%), and AUROC (84.2%). We found that Decision tree model was predicted correctly 4184 infants died and 5151 infants were alive. It has wrongly predicted 1372 births as died and 536 births as still alive. It was generally predicted infant death correctly at 83 %) of accuracy with recall (90.97 %), precision (78.9%), F1 score (84.6%), and AUROC (82.9%). It was found that the Support Vector Machine model was predicted correctly 3454 infants died and 4262 infants were alive. It has wrongly predicted 2102 births as died and 1425 births as still alive. It was generally predicted infant death correctly at 68.6 % of accuracy with recall (74.9%), precision (66.9%), F1 score (70.7%), and AUROC (68.5%). Based on the predictive model of performance results above, Random forest was the best predictive model of Infant mortality compared to other models applied in this study (**Table 4. 6**).

Table 4. 6: Predictive models' performance of Infant mortality

			Predictive Models							
Evaluation Matrix			**Logistic Regression**		**Random Forest**		**Decision Tree**		**Support Vector Machine**	
Confusion matrix			**Predicted**		**Predicted**		**Predicted**		**Predicted**	
			Dead	**Alive**	**Dead**	**Alive**	**Dead**	**Alive**	**Dead**	**Alive**
		Dead	3442	2114	4283	1273	4184	1372	3454	2102
	Observed	**Alive**	2215	3472	493	5194	536	5151	1425	4262
			%		**%**		**%**		**%**	
Accuracy			61.49		84.29		83.02		68.62	
Recall			61.05		91.33		90.97		74.94	
Precision			62.15		80.31		78.96		66.97	
F1 score			61.59		85.46		84.67		70.73	
AUROC			61.50		84.20		82.94		68.55	

Notes: AUROC: *Area under the Receiver Operating Characteristics*

4.7. Discussion of findings

This study described the application of ML methods in the analysis of infant mortality in Rwanda. This study shows that ML methods predict the risk factors for infant mortality is better than the logistic regression models or traditional methods. This result is not surprising, since ML methods are documented to outperform logistic methods in several fields of medicine. These findings were relevant to the findings of prior studies [28–31]. By using Random forest classifier methods, the findings revealed that province of residence, household wealth index, sex of children, maternal education, source of drinking water, maternal age at first birth, birth order, marital status, child twin, breastfeeding status, and form of residence and number of children ever born were all important risk factors associated with infant mortality. These results collaborated with findings of a prior study [32].

Regarding the predictive analysis, the prediction accuracies and AUC statistics revealed the highest for the Random Forest model. Our results confirmed a higher predictive power compared to the other ML models included in this study. Random forest model was predicted correctly 4160 infants as died and 1396 infants as alive. It has wrongly predicted 815 infants as died and 1,396 births as still alive. It was found that the random forest model has generally predicted infant deaths correctly at 80.3% of accuracy with recall: 85.6%, precision: 77.7%, F1 score: 81.5%, and AUROC: 80.3 %. These results are in congruence with the previous studies [18,29]. Decision tree model was predicted correctly 4134 infants died and 4833 infants as alive. It has wrongly predicted 854 births as died and 1422 births as still alive. It was generally predicted infant death correctly at 79.8% of accuracy with recall: 85%, precision: 77.3%, F1 score: 80.9%, and AUROC: 79.7%. These findings were similar to the earlier studies [29,33]. Support Vector Machine model was predicted correctly 3528 infants as died and 3918 infants as alive. It has wrongly predicted 1,769 births as died and 2,082 births as still alive. It was generally predicted infant death correctly at 66.2% of accuracy with recall: 68.9%, precision: 65.9%, F1 score: 67.4%, and AUROC: 66.2%. These results were analogous to prior studies [30,34].

Logistic regression models predicted correctly 3710 infants were died before completing their first years while 3,369 infants were still alive. It has wrongly predicted 2318 births will die before completing the first years and 1846 births were still alive. In congruence with prior studies [9,35], our results from logistic regression model has generally predicted infant mortality at 62.3% of accuracy with recall: 59.24%, precision: 64.6%, F1 score: 61.8%, and AUROC: 63%. Based on the predictive model's performance, Random forest was the best predictive model of infant mortality compared to other model results since it had the highest scores different evaluation metrics used in this study. In similar vein with earlier studies [9,13], our results revealed that ML methods or deep learning models are better than traditional analytical approach. Therefore, our overall results confirmed that ML methods significantly provide better discrimination than the traditional models in assessing the factors associated with infant mortality. These results are in similar vein with past studies that documented that the ML methods are more appropriate methods to determine factors associated with infant mortality and it presents a better goodness of fit in most critical groups [31,35].

CHAPTER FIVE

CONCLUSION AND RECOMMENDATIONS

5.1. General Conclusion

In developing a predictive model, ML approaches are strong and can be used to classify certain secret knowledge that could not be detected by conventional statistical methods. ML techniques can improve the accuracy of the algorithm and use training data for the training model and use unseen test data to make predictions. The general goals of this study were to apply methods of ML, namely logistic regression, random forest, decision tree, and support vector machine in the analysis of infant mortality in Rwanda. As an example of conventional statistical approaches that can be used to construct predictive models, logistic regression was used. The residence, wealth index, sex of children, maternal education, source of drinking water, age of the mother at first birth, birth order, marital status, child twin, breastfeeding, place of residence, and number of children ever born were the main factors by applying random forest methods to select best features associated with infant mortality. ML approaches have high output accuracy compared to conventional statistical methods. Among the four ML algorithms used in this study, the random forest was classified as the best classifier to be used for the predictive model of infant mortality in Rwanda compared to other ML models used in this study. Our approaches, ML methods, are recommended to be adapted to tackle other health outcomes such as survival very preterm, neonatal mortality, stunting, and low birth weight infants that remain public health concerns in Rwanda.

5.2. General recommendation

The ministry of health and other stakeholders can be used the findings of this study to design and implement child health intervention programs in Rwanda in order to reduce risk of dying among infants by applying Random forest model in prediction of infant mortality. I would like to suggest also University of Rwanda and other higher institution to teach machine learning program for find out the solution of community problem by applying machine learning methods.

5.3. Further work

This study was only used the RDHS2014-15 and focused on the four machine learning methods namely logistic regression, random forest, decision tree, and support vector machine. We would like to suggest other researchers to applying machine learning by taking into consideration of DHS datasets from different countries and compare the predictive models of infant mortality.

REFERENCES

[1] UNICEF: WHO: World Bank: UN DESA, "Levels & Trends in Child Mortality 2019," 2019.

[2] D. Trevisanuto *et al.*, "Effect of a neonatal resuscitation course on healthcare providers' performances assessed by video recording in a low-resource setting," *PLoS One*, vol. 10, no. 12, pp. 1–32, 2015, doi: 10.1371/journal.pone.0144443.

[3] D. D. Reidpath and P. Allotey, "Infant mortality rate as an indicator of population health," *J. Epidemiol. Community Health*, vol. 57, no. 5, pp. 344–346, 2003, doi: 10.1136/jech.57.5.344.

[4] H. Blencowe *et al.*, "Born Too Soon: The global epidemiology of 15 million preterm births," *Reprod. Health*, vol. 10, no. SUPPL. 1, pp. 1–14, 2013, doi: 10.1186/1742-4755-10-S1-S2.

[5] A. K. Singha, "Application of Machine Learning in Analysis of Infant Mortality and its Factors," no. February, 2016, doi: 10.13140/RG.2.1.3857.3687.

[6] S. Rezaei, K. Jamshidi, and K. Moradi, "Impact of socioeconomic and health system factors on infant mortality rate in Organization of the Petroleum Exporting Countries (OPEC): Evidence from 2004 to 2013," *Int. J. Pediatr.*, vol. 3, no. 2, pp. 493–499, 2015, doi: 10.22038/ijp.2015.4087.

[7] NISR, *Rwanda Demographic and Health Survey 2014-15 - Final Report*. 2015.

[8] Government of Rwanda, "2019 Rwanda Voluntary National Review (VNR) Report," *Sustain. Dev. Goals*, no. June, pp. 1–109, 2019.

[9] I. Monetary *et al.*, "Cameroon : Poverty Reduction Strategy Paper International Monetary Fund Washington , D . C .," no. 10, 2010.

[10] E. Biracyaza and S. Habimana, "Modeling the Risks Factors Associated with Infant Mortality in Rwanda from 2011 to 2015: Analysis of Rwanda Demographic and Health Survey (RDHS) 2014/2015," *J. Pediartrics Child Heal. Care*, vol. 4, no. 1, p. 1026, 2019.

[11] J. Dusabe, “Determinants of Under-5 Mortality in Rwanda,” 2016.

[12] F. Munyamahoro, “An Empirical Analysis of Death of ChildrenUnder Five Years in Rwanda,” *J. Med. Res. Heal. Educ.*, vol. 1, no. 2, pp. 2015–2018, 2017, [Online]. Available: https://www.imedpub.com/articles/an-empirical-analysis-of-death-of-childrenunder-five-years-in-rwanda.php?aid=20095.

[13] D. Dwomoh, S. Amuasi, K. Agyabeng, G. Incoom, Y. Alhassan, and A. E. Yawson, “Understanding the determinants of infant and under-five mortality rates: A multivariate decomposition analysis of Demographic and Health Surveys in Ghana, 2003, 2008 and 2014,” *BMJ Glob. Heal.*, vol. 4, no. 4, pp. 1–20, 2019, doi: 10.1136/bmjgh-2019-001658.

[14] M. Podda, D. Bacciu, A. Micheli, R. Bellù, G. Placidi, and L. Gagliardi, “A machine learning approach to estimating preterm infants survival: development of the Preterm Infants Survival Assessment (PISA) predictor,” *Sci. Rep.*, vol. 8, no. 1, 2018, doi: 10.1038/s41598-018-31920-6.

[15] S. Sakr *et al.*, “Comparison of machine learning techniques to predict all-cause mortality using fitness data: The Henry Ford exercIse testing (FIT) project,” *BMC Med. Inform. Decis. Mak.*, vol. 17, no. 1, pp. 1–15, 2017, doi: 10.1186/s12911-017-0566-6.

[16] O. Wellington, “Determinant of Infant Mortality Rate: A Panel Data Analysis of African Countries,” vol. 4, no. 18, 2014.

[17] O. T. POPOOLA, “An Empirical Analysis of Infant Mortality,” *J. Econ. Manag. Sci.*, vol. 2, no. 1, p. p27, 2019, doi: 10.30560/jems.v2n1p27.

[18] A. Kotsadam, G. Østby, S. A. Rustad, A. F. Tollefsen, and H. Urdal, “Development aid and infant mortality. Micro-level evidence from Nigeria,” *World Dev.*, vol. 105, pp. 59–69, 2018, doi: 10.1016/j.worlddev.2017.12.022.

[19] T. P. Schultz, “Studying the impact of household economic and community variables on child mortality.,” *Child Surviv. Strateg. Res.*, vol. 10, no. 1984, pp. 215–235, 1984, doi: 10.2307/2807962.

[20] C. Mosley,W,H and Chen, L, “Extracted from : Population and Development Review 1984 ; 10 Suppl : 25 – 45 .,” *Bull World Heal. Organ*, vol. 81, no. 2, pp. 140–145, 2003.

[21] Kabubo-Mariara, “Child Survival and Policy Options in Kenya: Evidence from Demographic and Health Surveys,” *J. Rev. Glob. Econ.*, pp. 13–26, 2012, doi: 10.6000/1929-7092.2012.01.2.

[22] P. A. M. Ntenda, K. Y. Chuang, F. N. Tiruneh, and Y. C. Chuang, “Factors Associated with Infant Mortality in Malawi,” *J. Exp. Clin. Med.*, vol. 6, no. 4, pp. 125–131, 2014, doi: 10.1016/j.jecm.2014.06.005.

[23] S. Vijayakumar, M. G. Deepika, A. Jyotishi, and D. Gupta, “Factors affecting infant mortality rate in India: An analysis of Indian states,” *Adv. Intell. Syst. Comput.*, vol. 530, pp. 707–719, 2016, doi: 10.1007/978-3-319-47952-1_57.

[24] D. Mekonnen, “Infant and Child Mortality in Ethiopia,” *Masters thesis spring 2011*, 2011.

[25] R. Bello and A. Joseph, “Determinants of Child Mortality in Oyo State, Nigeria,” *African Res. Rev.*, vol. 8, no. 1, p. 252, 2014, doi: 10.4314/afrrev.v8i1.17.

[26] Rana Ejaz Ali Khan, Khadija Malik Bari, and Muhammad Ali Raza, “Socioeconomic determinants of child mortality: Evidence from Pakistan Demographic and Health Survey,” *Bus. Rev.*, vol. 13, no. 2, pp. 34–50, 2018.

[27] X. Y. Kong *et al.*, “Neonatal mortality and morbidity among infants between 24 to 31 complete weeks: A multicenter survey in China from 2013 to 2014,” *BMC Pediatr.*, vol. 16, no. 1, pp. 1–8, 2016, doi: 10.1186/s12887-016-0716-5.

[28] R. Lamichhane, Y. Zhao, S. Paudel, and E. O. Adewuyi, “Factors associated with infant mortality in Nepal: A comparative analysis of Nepal demographic and health surveys (NDHS) 2006 and 2011,” *BMC Public Health*, vol. 17, no. 1, pp. 1–18, 2017, doi: 10.1186/s12889-016-3922-z.

[29] C. Mugeni *et al.*, “Nationwide implementation of integrated community case management of childhood illness in Rwanda,” *Glob. Heal. Sci. Pract.*, vol. 2, no. 3, pp. 328–341, 2014, doi: 10.9745/GHSP-D-14-00080.

[30] S. Sakr *et al.*, “Comparison of machine learning techniques to predict all-cause mortality using fitness data: The Henry Ford exercIse testing (FIT) project,” *BMC Med. Inform. Decis. Mak.*, vol. 17, no. 1, pp. 1–15, 2017, doi: 10.1186/s12911-017-0566-6.

[31] L. J. B. Caluza, “Machine Learning Algorithm Application in Predicting Children Mortality : A Model Development,” *Int. J. Inf. Sci. Appl.*, vol. 10, no. 1, pp. 1–6, 2018.

[32] R. Ramos, C. Silva, M. W. L. Moreira, J. J. P. C. Rodrigues, M. Oliveira, and O. Monteiro, “Using predictive classifiers to prevent infant mortality in the Brazilian northeast,” *2017 IEEE 19th Int. Conf. e-Health Networking, Appl. Serv. Heal. 2017*, vol. 2017-Decem, pp. 1–6, 2017, doi: 10.1109/HealthCom.2017.8210811.

[33] B. Tesfaye, S. Atique, N. Elias, L. Dibaba, S. A. Shabbir, and M. Kebede, “Determinants and development of a web-based child mortality prediction model in resource-limited settings: A data mining approach,” *Comput. Methods Programs Biomed.*, vol. 140, no. November, pp. 45–51, 2017, doi: 10.1016/j.cmpb.2016.11.013.

[34] A. SHEGAW, “Application of data mining technology to predict child mortality patterns: the case of Butajira Rural Health Project (BRHP),” 2002, [Online]. Available: http://etd.aau.edu.et/handle/123456789/2801.

[35] R. R. Beeman, “List of Figures,” *Evol. South. Backcountry*, no. December 2014, pp. ix–x, 2014, doi: 10.9783/9780812200874.ix.

[36] T. D. Buskirk, A. Kirchner, A. Eck, and C. S. Signorino, “An Introduction to Machine Learning Methods for Survey Researchers,” *Surv. Pract.*, vol. 11, no. 1, pp. 1–10, 2018, doi: 10.29115/sp-2018-0004.

[37] D. Pietersma, R. Lacroix, D. Lefebvre, and K. M. Wade, “Performance analysis for machine-learning experiments using small data sets,” *Comput. Electron. Agric.*, vol. 38, no. 1, pp. 1–17, 2003, doi: 10.1016/S0168-1699(02)00104-7.

[38] K. A. Gaurav and L. Patel, *Machine Learning With R*. 2020.

[39] R. Barandela, R. M. Valdovinos, J. Salvador Sánchez, and F. J. Ferri, “The imbalanced training sample problem: under or over sampling?,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 3138, pp. 806–814, 2004, doi: 10.1007/978-3-540-27868-9_88.

[40] L. Guo, “Applying data mining techniques in property/casualty insurance,” *Casualty Actuar. Soc. Forum*, pp. 1–25, 2003, [Online]. Available:

http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Applying+Data+Mining+Techniques+in+Property/Casualty+Insurance#0.

[41] R. Konieczny and R. Idczak, "Mössbauer study of Fe-Re alloys prepared by mechanical alloying," *Hyperfine Interact.*, vol. 237, no. 1, pp. 1–8, 2016, doi: 10.1007/s10751-016-1232-6.

[42] M. H. H. Parmar and G. H. Shah, "Experimental and Comparative Analysis of Machine Learning Classifiers," *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 3, no. 10, 2013, [Online]. Available: https://www.researchgate.net/publication/262337213_Experimental_and_Comparative_Analysis_of_Machine_Learning_Classifiers.

[43] J. Ali, R. Khan, N. Ahmad, and I. Maqsood, "Random Forests and Decision Trees," *Int. J. Comput. Sci. Issues*, vol. 9, no. 5, pp. 272–278, 2012.

[44] H. Bhavsar and A. Ganatra, "A Comparative Study of Training Algorithms for Supervised Machine Learning," *Int. J. Soft Comput. Eng.*, vol. 2, no. 4, pp. 74–81, 2012, doi: 10.1.1.492.6088.

[45] V. Jakkula, "Tutorial on Support Vector Machine (SVM)," *Sch. EECS, Washingt. State Univ.*, pp. 1–13, 2011, [Online]. Available: http://www.ccs.neu.edu/course/cs5100f11/resources/jakkula.pdf.

[46] A. Arivarasan, "Classification based Performance analysis using Naïve-Bayes J48 and Random forest algorithms," *Int. J. Appl. Res. 2017*, vol. 3, no. 6, pp. 174–178, 2017.

[47] P. J. G. Lisboa, "A review of evidence of health benefit from artificial neural networks in medical intervention," *Neural Networks*, vol. 15, no. 1, pp. 11–39, 2002, doi: 10.1016/S0893-6080(01)00111-3.

[48] J. Jaskari *et al.*, "Machine Learning Methods for Neonatal Mortality and Morbidity Classification," *IEEE Access*, vol. 8, pp. 123347–123358, 2020, doi: https://doi.org/10.1109/ACCESS.2020.3006710.

[49] C. Lipton, *The Mythos of Model Interpretability*. ICML Workshop on Human Interpretability in Machine Learning (WHI 2016). New York, NY, USA, 2016.

[50] E. D. Peet, D. Schultz, and S. L. Lovejoy, *Using an Innovative Database and Machine Reduce Infant Mortality*. Santa Monica, CA: RAND Corporation, 2021.

[51] F. H. Bitew, S. H. Nyarko, L. Potter, and C. S. Sparks, "Machine learning approach for predicting under-five mortality determinants in Ethiopia: evidence from the 2016 Ethiopian Demographic and Health Survey," *Genus*, vol. 76, no. 1, 2020, doi: 10.1186/s41118-020-00106-2.

[52] S. Sakr, R. Elshawi, A. Ahmed, W. Qureshi, C. Brawner, and S. Keteyian, "Comparison of machine learning techniques to predict all-cause mortality using fitness data: the Henry ford exercIse testing (FIT) project," pp. 1–15, 2017.

[53] Z. Kefi, K. Aloui, and M. S. Naceur, "New approach based on machine learning for short-term mortality prediction in neonatal intensive care unit," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 7, pp. 583–591, 2019, doi: 10.14569/ijacsa.2019.0100778.

[54] A. K. Singha, D. Phukan, S. Bhasin, and R. Santhanam, "Application of Machine Learning in Analysis of Infant Mortality and its Factors," no. February, 2016, doi: 10.13140/RG.2.1.3857.3687.

[55] M. Podda, D. Bacciu, A. Micheli, R. Bellù, G. Placidi, and L. Gagliardi, "A machine learning approach to estimating preterm infants survival: development of the Preterm Infants Survival Assessment (PISA) predictor," *Sci Rep*, vol. 8, p. 13743., 2018.

[56] A. E. Adegbosin, B. Stantic, and J. Sun, "Efficacy of deep learning methods for five mortality in 34 income countries," *BMJ Open*, vol. 2020, no. 10, p. e034524, 2020, doi: 10.1136/bmjopen-2019-034524.

Application of machine learning methods in analysis of Infant Mortality in Rwanda:Analysis of RDHS2014-15

ORIGINALITY REPORT

18%	14%	6%	9%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.rhsupplies.org Internet Source	1%
2	www.measuredhs.com Internet Source	1%
3	irphouse.com Internet Source	1%
4	oro.open.ac.uk Internet Source	1%
5	sustainabledevelopment.un.org Internet Source	1%
6	Submitted to University of Rwanda Student Paper	1%
7	mpra.ub.uni-muenchen.de Internet Source	<1%
8	Hasina Rakotomanana, Gail E. Gates, Deana Hildebrand, Barbara J. Stoecker. "Situation and determinants of the infant and young child	<1%

feeding (IYCF) indicators in Madagascar: analysis of the 2009 Demographic and Health Survey", BMC Public Health, 2017
Publication

9	Submitted to University of Melbourne Student Paper	<1%
10	erepository.uonbi.ac.ke Internet Source	<1%
11	Submitted to Universiti Teknologi MARA Student Paper	<1%
12	Submitted to Heriot-Watt University Student Paper	<1%
13	bmcpediatr.biomedcentral.com Internet Source	<1%
14	scholar.mzumbe.ac.tz Internet Source	<1%
15	www.coursehero.com Internet Source	<1%
16	Submitted to Liverpool John Moores University Student Paper	<1%
17	Brook Tesfaye, Suleman Atique, Noah Elias, Legesse Dibaba, Syed-Abdul Shabbir, Mihiretu Kebede. "Determinants and development of a web-based child mortality prediction model in resource-limited settings: A data mining	<1%

Childbirth, 2020
Publication

24	www.unpei.org Internet Source	<1%
25	Submitted to Mahidol University Student Paper	<1%
26	Submitted to University of Sheffield Student Paper	<1%
27	mafiadoc.com Internet Source	<1%
28	Submitted to Sheffield Hallam University Student Paper	<1%
29	www.rikidscount.org Internet Source	<1%
30	www.statista.com Internet Source	<1%
31	opendocs.ids.ac.uk Internet Source	<1%
32	publications.waset.org Internet Source	<1%
33	Submitted to Australian National University Student Paper	<1%
34	Submitted to Mount Kenya University Student Paper	<1%

35	etd.aau.edu.et Internet Source	<1%
36	Submitted to University of Bradford Student Paper	<1%
37	Submitted to University of Sydney Student Paper	<1%
38	meddic.jp Internet Source	<1%
39	Submitted to The University of Memphis Student Paper	<1%
40	www.tandfonline.com Internet Source	<1%
41	Tadesse Sebsibe, Alemu Yinges. "Urban-rural differentials in child undernutrition in Ethiopia", International Journal of Nutrition and Metabolism, 2015 Publication	<1%
42	www.ijcaonline.org Internet Source	<1%
43	dspace.daffodilvarsity.edu.bd:8080 Internet Source	<1%
44	studentsrepo.um.edu.my Internet Source	<1%
45	Jing-Jing Wan, Bo-Lun Chen, Yi-Xiu Kong, Xing-	

	Gang Ma, Yong-Tao Yu. "An Early Intestinal Cancer Prediction Algorithm Based on Deep Belief Network", Scientific Reports, 2019 Publication	<1%
46	Submitted to Eiffel Corporation Student Paper	<1%
47	repository.out.ac.tz Internet Source	<1%
48	pdfs.semanticscholar.org Internet Source	<1%
49	kebijakankesehatanindonesia.net Internet Source	<1%
50	Submitted to Indian Institute of Technology, Madras Student Paper	<1%
51	towardsdatascience.com Internet Source	<1%
52	Submitted to essex Student Paper	<1%
53	issuu.com Internet Source	<1%
54	Submitted to African Population Health Research Centre Student Paper	<1%
	Submitted to Campbell University	

55	Student Paper	<1%
56	www.inderscience.com Internet Source	<1%
57	Tom Lutalo, Ron Gray, John Santelli, David Guwatudde et al. "Unfulfilled need for contraception among women with unmet need but with the intention to use contraception in Rakai, Uganda: a longitudinal study", BMC Women's Health, 2018 Publication	<1%
58	Submitted to University of Computer Studies Student Paper	<1%
59	www.neliti.com Internet Source	<1%
60	nmsac.unm.edu Internet Source	<1%
61	www.koreascience.or.kr Internet Source	<1%
62	Wingston Ng'ambi, Tara Mangal, Andrew Phillips, Tim Colbourn, Joseph Mfutso-Bengo, Paul Revill, Timothy B. Hallett. "Factors associated with healthcare seeking behaviour for children in Malawi: 2016", Tropical Medicine & International Health, 2020 Publication	<1%

63	Submitted to University of Witwatersrand Student Paper	<1%
64	www.repository.cam.ac.uk Internet Source	<1%
65	www.ijert.org Internet Source	<1%
66	dottorati.unica.it Internet Source	<1%
67	researchspace.ukzn.ac.za Internet Source	<1%
68	I. Luoma, K. Puura, T. Tamminen, P. Kaukonen, J. Piha, E. Räsänen, K. Kumpulainen, I. Moilanen, A. M. Koivisto, F. Almqvist. "Emotional and behavioural symptoms in 8–9-year-old children in relation to family structure", European Child & Adolescent Psychiatry, 1999 Publication	<1%
69	"Death at the Opposite Ends of the Eurasian Continent", Walter de Gruyter GmbH, 2011 Publication	<1%
70	hub.hku.hk Internet Source	<1%
71	diva-portal.org Internet Source	<1%

72	mhealth.jmir.org Internet Source	<1%
73	www.informatica.si Internet Source	<1%
74	LAWRENCE IKAMARI. "SIBLING MORTALITY CORRELATION IN KENYA", Journal of Biosocial Science, 2000 Publication	<1%
75	S. B. Kotsiantis, D. Kanellopoulos, V. Karioti, V. Tampakas. "An ontology-based portal for credit risk analysis", 2009 2nd IEEE International Conference on Computer Science and Information Technology, 2009 Publication	<1%
76	allafrica.com Internet Source	<1%
77	www.slideshare.net Internet Source	<1%
78	Submitted to National University of Ireland, Galway Student Paper	<1%
79	ueaeprints.uea.ac.uk Internet Source	<1%
80	hdl.handle.net Internet Source	<1%

81	worldwidescience.org Internet Source	<1%
82	etd.ohiolink.edu Internet Source	<1%
83	archive.org Internet Source	<1%
84	MMTE Huynen, P Martens, RS De Groot. "Linkages between biodiversity loss and human health: a global indicator analysis", International Journal of Environmental Health Research, 2004 Publication	<1%
85	spheru.ca Internet Source	<1%
86	www.unicef.org Internet Source	<1%
87	repository.smuc.edu.et Internet Source	<1%
88	uir.unisa.ac.za Internet Source	<1%
89	eprints.nottingham.ac.uk Internet Source	<1%
90	Predictive Analytics with Microsoft Azure Machine Learning, 2015. Publication	<1%

91	Muralikrishna Iyyanki, Prisilla Jayanthi, Valli Manickam. "chapter 10 Machine Learning for Health Data Analytics", IGI Global, 2020 Publication	<1%
92	Muhammad Abu Bakar, Rahma Fiaz, Eustasius Musenge. "Factors influencing malaria infection in Rwanda 2010: a cross-sectional survey study using generalized structural equation modeling", Research Square, 2019 Publication	<1%
93	"Program Abstracts from the GSA 2019 Annual Scientific Meeting “Strength in Age—Harnessing the Power of Networks”", Innovation in Aging, 2019 Publication	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		