



ISSN: 2410-1397

Master Project in Biometry

# Supervised Machine Learning Approaches to Predict Infant Mortality: A case study of the 2014 Kenya Demographic and Health Survey

Research Report in Mathematics, Number .40, 2021

Caroline Kioko

July 2021





# **Supervised Machine Learning Approaches to Predict Infant Mortality: A case study of the 2014 Kenya Demographic and Health Survey**

**Research Report in Mathematics, Number 40, 2021**

Caroline Kioko

School of Mathematics  
College of Biological and Physical sciences  
Chiromo, off Riverside Drive  
30197-00100 Nairobi, Kenya

**Master Thesis**

Submitted to the School of Mathematics in partial fulfilment for a degree in Master of Science in Mathematics

Submitted to: The Graduate School, University of Nairobi, Kenya

## Abstract

The study used supervised machine learning approaches to predict infant mortality in Kenya and the 2014 Kenya Demographic and Health Survey. Different classification methods were used. The methods were Logistic regression, K-nearest neighbor and Random forest model. Random Forest performed well with an accuracy of approximately 97.1% followed by Logistic Regression model with 86.1% and K-nearest neighbor with 85.6%. The results concluded that random forest model was the best performing model.

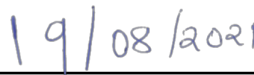


## Declaration and Approval

I the undersigned declare that this dissertation is my original work and to the best of my knowledge, it has not been submitted in support of an award of a degree in any other university or institution of learning.



Signature



Date

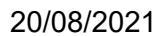
CAROLINE KEMUNTO KIOKO

Reg No. I56/34480/2019

In my capacity as a supervisor of the candidate's dissertation, I certify that this dissertation has my approval for submission.



Signature



Date

Dr. Linda Chaba.  
School of Mathematics,  
University of Nairobi,  
Box 30197, 00100 Nairobi, Kenya.  
E-mail: [lindachaba@gmail.com](mailto:lindachaba@gmail.com)



## Dedication

This project is dedicated to my father (John Kioko), mother (Esther Sani Kioko), sister (Linet Kioko) and brothers (Asher and Lameck).



# Contents

<b>Abstract</b> .....	<b>ii</b>
<b>Declaration and Approval</b> .....	<b>iv</b>
<b>Dedication</b> .....	<b>vi</b>
<b>List of Figures</b> .....	<b>ix</b>
<b>List of Tables</b> .....	<b>x</b>
<b>Acknowledgments</b> .....	<b>xi</b>
<b>1 CHAPTER 1: INTRODUCTION</b> .....	<b>1</b>
1.1 <b>Background of the Study</b> .....	<b>1</b>
1.1.1 <i>Global Mortality</i> .....	<b>1</b>
1.1.2 <i>Mortality in Sub-Saharan Africa</i> .....	<b>1</b>
1.1.3 <i>Mortality in Kenya</i> .....	<b>1</b>
1.1.4 <i>Machine Learning</i> .....	<b>2</b>
1.2 <b>PROBLEM STATEMENT</b> .....	<b>3</b>
1.3 <b>OBJECTIVES</b> .....	<b>4</b>
1.3.1 <b>General Objective</b> .....	<b>4</b>
1.3.2 <b>Specific Objectives</b> .....	<b>4</b>
1.4 <b>SIGNIFICANCE OF THE STUDY</b> .....	<b>4</b>
<b>2 CHAPTER 2: LITERATURE REVIEW</b> .....	<b>5</b>
2.1 <b>INTRODUCTION</b> .....	<b>5</b>
2.2 <b>EMPIRICAL LITERATURE</b> .....	<b>5</b>
2.3 <b>SUMMARY OF THE LITERATURE REVIEW</b> .....	<b>8</b>
2.4 <b>LITERATURE GAP</b> .....	<b>8</b>
<b>3 RESEARCH METHODOLOGY</b> .....	<b>9</b>
3.1 <b>DATA DESCRIPTION</b> .....	<b>9</b>
3.2 <b>Study Variables and Measurements</b> .....	<b>9</b>
3.3 <b>Supervised machine learning methods</b> .....	<b>10</b>
3.3.1 <b>Logistic Regression</b> .....	<b>10</b>
3.3.2 <b>K-Nearest Neighbor</b> .....	<b>11</b>
3.3.3 <b>Random Forest</b> .....	<b>12</b>
3.4 <b>Training and testing data</b> .....	<b>13</b>
3.5 <b>Performance Evaluation</b> .....	<b>13</b>
3.5.1 <b>Confusion Matrix</b> .....	<b>13</b>
3.5.2 <b>Receiver Operating Characteristics (ROC) curves and Area Under Curve (AUC)</b> .....	<b>15</b>
<b>4 CHAPTER 4: DATA ANALYSIS AND RESULTS</b> .....	<b>16</b>
4.1 <b>RESULTS</b> .....	<b>16</b>

---

4.1.1	Introduction.....	16
4.1.2	Descriptive results of the background characteristics .....	16
4.1.3	Predicting Infant Mortality .....	20
4.1.4	Receiver Operating Characteristics (ROC) Curve .....	21
4.1.5	Variable importance measures for the random forest model .....	22
<b>5</b>	<b>CHAPTER 5: DISCUSSION AND CONCLUSION .....</b>	<b>23</b>
5.1	Discussion .....	23
5.2	Conclusion .....	23
5.3	Future Research .....	23
5.4	Limitation of the Study .....	24
	<b>References .....</b>	<b>25</b>

## List of Figures

Figure 1. ROC Curves for the three models.....	21
Figure 2. Variable Importance for the random forest model .....	22

## List of Tables

Table 1. <i>Summary Statistics of study variables</i> .....	16
Table 2. <i>Description of continuous variables</i> .....	20
Table 3. <i>Results of the three machine learning models</i> .....	20

## Acknowledgments

First I thank my supervisor, the school of mathematics and the University of Nairobi.

Caroline Kioko

---

Nairobi, 2021.



---

# 1 CHAPTER 1: INTRODUCTION

## 1.1 Background of the Study

### 1.1.1 Global Mortality

Infant mortality is defined as the death of a child before attaining the age of one year [21]. Globally, there has been good progress in reducing child mortality. As compared to early 1990, millions of children have a better chance of surviving. However, the global number of child deaths is still high. According to the United Nations Inter-agency Group for Child Mortality Estimation (UN IGME), the estimate of infant mortality rate worldwide by the Sustainable Development Goal region has reduced from 48 deaths per 1,000 live births in 2003 to 28 deaths per 1,000 live births in 2019.

According to [38], the global under 5 mortality rate has reduced from 90.6 deaths per 1,000 live-births (90% uncertainty interval 89.3 – 92.2) in 1990 to 42.5 (40.9 – 45.6) in 2015, which represents a reduction of under-five mortality rate by 59 percent. At the same period, the annual number of under five deaths worldwide decreased from 12.7 million (12.6million – 13.0 million) to 5.9 million (5.7 million – 6.4 million). The global under five mortality rate reduced by 53% (50 – 55%) in the past 25 years and thus missed the MDG 4 target [38]. Based on point estimates, two regions East Asia and the Pacific, Latin America and Caribbean achieved the MDG 4 target [38].

Therefore, between 2016 and 2030, 94.4 million children are projected to die before the age of 5 years if the mortality rate remains constant in each country and 68.8 million would die if each country continues to reduce its mortality rate at the pace estimates from 2000 to 2015 [38].

### 1.1.2 Mortality in Sub-Saharan Africa

Sub-Saharan Africa remains to be the region with the highest under-five mortality rate. The under-five mortality rate was 78 deaths per 1,000 livebirths in 2018 which translates as 1 in every 12 children dies before the age of 5 [38]. According to [38], seven countries from sub-Saharan Africa region recorded mortality rates of more than 100 deaths per 1,000 live births. If all countries achieve the Sustainable Development Goal of an under five mortality rate of 25 or fewer deaths per 1,000 livebirths by 2030, we project 56.0 million deaths by 2030. The two thirds of all sub-Saharan African countries need to accelerate progress to achieve this target [38].

### 1.1.3 Mortality in Kenya

According to the United Nations International Children’s Emergency Fund (UNICEF), the infant mortality rate in Kenya was 31.87 per 1,000 livebirths in 2019, reducing by 32.97% from 1990. How-

ever, the country had not achieved Millennium Development Goal 4 (MDG 4) by 2015. Even though Kenya has made a huge progress on improving child survival since independence (1963), under-five mortality still remains high and heterogeneous with substantial differences between counties [19]. Historically, the regions which have recorded the highest child mortality rates includes the coastal region, arid and semi-arid areas around Lake Turkana and areas around the Lake Victoria region. The Coastal and Lake Victoria region have high child mortality rates due to intense malaria transmission. Also, the reason for high child mortality rates in arid and semi-arid areas around Lake Turkana is due to harsh arid conditions.

#### 1.1.4 Machine Learning

Machine learning (ML) is one of the most advanced concepts of artificial intelligence (AI), and provides a strategic approach to developing automated, complex and objective algorithmic techniques for multimodal and dimensional biomedical or mathematical data analysis [27]. The ML algorithms are able to read and modify its structure based on a set of observed data with adaptation done by optimizing over a cost function or an objective [14]. ML techniques can be classified in four ways:

##### a) Supervised learning techniques

Supervised learning techniques are ML learning techniques or algorithms that bind previous and current dataset with the help of labeled data to predict future events [20]. For supervised learning, the learning process starts with a dataset training and develops targeted activity to predict output values and the techniques are able to give results in input data with an adequate training process, compare results with actual results, identify errors and modify the model according to the results [26].

##### b) Unsupervised learning techniques

According to [25], these techniques are used when the training data set is non-classified or non labeled. The learning techniques deduce a function to extract hidden knowledge from unlabeled data-set. This technique does not identify the proper output but rather extracts observations from the dataset's to find hidden patterns from the unlabeled data set [25].

##### c) Semi-supervised learning techniques

According to [17], semi-supervised learning techniques lie between supervised learning techniques and unsupervised learning techniques, where labeled and unlabeled datasets are used in the training process. These learning techniques consider a smaller labeled data set and larger unlabeled data set [29].

##### d) Reinforcement learning techniques

According to [17], reinforcement learning techniques interact with the learning environment by actions to identify errors. Some of the common features of the reinforcement learning techniques are delayed rewards, trial and error searches and the techniques are used to identify



the ideal behavior in a specific context to increase the performance of the model [29], [28] and [30].

The machine learning process begins with collecting data from a variety of resources [12]. The next step is to fix the pre-processed data to fix data-related issues and reduce space size by deleting invalid file data to select interesting data [37] and sometimes the value of the dataset might be very hard for the system to make decision, therefore, machine learning algorithms are designed using others concept such as statistics, theory control and probability to analyze data and extract useful information from past experiences [12]. The next step is the performance evaluation of the models and finally is model optimization improving the model using new dataset and rules [25]). Machine learning techniques have been used in a variety of areas such as medicine, engineering, education, manufacturing and production, forecast, traffic management and robot among others [12].

In this work, supervised machine learning techniques are used to predict infant mortality in Kenya, using the 2014 Kenya Demographic and Health Survey data with common supervised learning algorithms which are random forests, logistic regressions and K-nearest neighbors. This approach is substantially faster and easier than manual classification [9].

## 1.2 PROBLEM STATEMENT

According to the World Health Organization, infant mortality remains to be a major concern in many parts of the world. In developing countries, high levels of infant deaths has been a serious problem.

The Kenyan Government's vision in reducing infant mortality has been remarkable for the past 18 years. Government programs in place such as Malezi Bora Strategy, Child Survival and Integrated Management of Childhood Illness Program have been approved to be effective in improving child health. Although infant mortality rates have decreased from 53 per 1,000 live births in 2003 to 32 per 1,000 live births in 2019, it is still high. Also, in 2015, Kenya did not achieve its Millennium Development Goal (MDG) target for reducing under-five mortality. Therefore, for the country to accelerate progress to 2030, she needs to understand what impacted mortality during the MDG period. By identifying factors that affect infant mortality, the current government efforts in place can be further enhanced and optimized.

Previous studies have investigated several predictors of infant mortality in Kenya. [16] reported the socioeconomic determinants of infant mortality in Kenya using the 2003 Kenya Demographic and Health Survey (KDHS). Similarly [2] analyzed regional variations of infant mortality in Kenya using the 2009 Kenya Demographic and Health Survey. Although many studies have been carried out previously to identify factors resulting in infant mortality in Kenya using KDHS datasets of different surveys, to the best of our knowledge, no studies have been done in Kenya to predict infant mortality risks using the supervised machine learning methods and the 2014 KDHS data.

According to [4] machine learning provides solutions for all possible problems in vision, speech, health and robotics. Several studies have applied machine learning to solve health related problems.

[7] used a machine learning approach to predict under-five mortality determinants in Ethiopia using the 2016 Ethiopian Demographic and Health Survey data. [6] also used a machine learning approach for confirmation of Covid-19 cases (positive, negative, death and release). This means that there is no optimal machine learning technique that fits all situations. For instance, the method that works best for Ethiopian data or Israel Covid-19 data may not perform best for Kenyan data. This study aims to determine risk factors of infant mortality using the best performing supervised machine learning models and the 2014 Kenya Demographic and Health Survey (KDHS) data.

### 1.3 OBJECTIVES

The following are the study's objectives:

#### 1.3.1 General Objective

To evaluate supervised machine learning approach for predicting infant mortality.

#### 1.3.2 Specific Objectives

- a. To explore the 2014 KDHS dataset with reference to infant mortality
- b. To compare the performance of the supervised machine learning method for predicting infant mortality using 2014 KDHS dataset
- c. To determine risk factors of infant mortality in Kenya using the best performing Machine Learning approach identified in objective 2

### 1.4 SIGNIFICANCE OF THE STUDY

The identified risk factors of infant mortality will help the Government of Kenya and non-government institutions access if the implemented health programs in place are useful in increasing the survival rate of infants. This will lead to affordable infant programs such as vaccines and intensive care of newborns and their mothers.

---

## 2 CHAPTER 2: LITERATURE REVIEW

### 2.1 INTRODUCTION

Several studies have been done in developing countries using survey data and census data with the aim of investigating key predictors of infant mortality.

### 2.2 EMPIRICAL LITERATURE

[35] compared different supervised machine learning algorithms for disease prediction to identify key trends among different types of supervised machine learning algorithms, their performances and usage for disease risk prediction. Two databases (Scopus and PubMed) were searched for different types of search items and 48 articles were selected in total for the comparison among variants of supervised machine learning algorithms for disease prediction. The results of the study found that the Support Vector Machine (SVM) algorithm was applied most frequently (in 29 studies), followed by the Naïve Bayes algorithm (in 23 studies). The study also determined that among all the algorithms, the Random Forest (RF) algorithm had superior accuracy comparatively. Of the 17 studies it was applied, RF showed the highest accuracy in 9 of them followed by SVM. The study concluded that the relative performance of different variants of supervised machine learning algorithms for disease prediction can be used by researchers in the selection of an appropriate algorithm for their studies.

[32] applied five supervised machine learning techniques (Support Vector Machine, Random Forest, K-Nearest Neighbor, Naïve Bayes and Softmax) to predict stock market trends. The results showed that the Random Forest algorithm performs the best for large datasets and the Naïve Bayesian classifier performs best for small datasets. The results also revealed that reduction in the number of technical predictors reduces the accuracies of each algorithm.

[33] used Systematic Literature Review (SLR) method to gain a thorough insight into different algorithms used in supervised machine learning (SML) and their categories, and also to compare the best performance measures of the SML algorithm. Various algorithms under SML were Naïve Bayes, Logistic Regression, Random Forest, J48, CART, Artificial Neural Network, Multi-Layer Perceptron and Support Vector Machine (SVM). SLR was performed on the existing research works from the year 2015 – 2019, sorting of the papers according to selection criteria and data extraction was done, and 61 final studies were selected. The study found that SVM and Artificial Neural Network (ANN) were the top two performing algorithms in classification. The study concluded that research study should include more assessment measures of SML algorithms and unsupervised machine learning.

[31] applied machine learning in the analysis of infant mortality and its factors using the United States dataset. The study used the Birth Data Files of the year 2013 and important factors were identified using Kendall rank correlation coefficient. A randomized split of 80% to 20% was used to build a training dataset of 32432 cases and a testing dataset of 8081 cases out of 40541 cases. Three different classification models (Logistic Regression, Naïve Bayes, and Lagrangian Support Vector Machine) were used for the binary classification problem and the three models were fitted with the training data and cross validated in 10 folds using the test data. The evaluation metrics used were accuracy, precision, recall and F1-score. The results determined that the performance of the Logistic Regression Model was the best with a high precision score followed by Naïve Bayes and the Lagrangian Support Vector Machine. The study also indicated that identifying a mother at high risk is more important than misclassification of a low risk mother given by recall. The results suggested that the best model with high precision can be used to predict key factors and high risk mothers and thus they can be given proper medical care to mitigate the risk, hence reducing the infant mortality rate.

[34] used multiple logistic regression and cross tabulation analysis to investigate the predictors of infant mortality in Bangladesh. Predictors of infant mortality were categorized into two sections: Neonatal Mortality and Post-neonatal Mortality. The study used data from the Bangladesh Demographic and Health Survey (BDHS) 1999-2000. The study considered all births and deaths that occurred during 5 years prior to the survey and the sample children were then divided into two cohorts: neonatal (deaths within 0 – 28 days of age) and post neonatal (deaths between 1 – 11 months of age). Results from the study indicated that parents' education had a significant negative effect on infant mortality while occupation of parents had a significant influence on post-neonatal mortality only. The study also determined that mother's education, family size, breastfeeding status, mother's age at birth, birth spacing, complication during birth, type of birth, timing of first antenatal check and Tetanus Toxoid (TT) during pregnancy had significant effect on neonatal mortality while post-neonatal mortality varied significantly by education and occupation of father, family size, breastfeeding status, mother's age at birth, type of birth and TT during pregnancy. The study suggested that further research is needed to assess the impact of several variables on neonates and post-neonates so that policy formulation will be easier for the people involved in planning purposes.

[11] conducted a research study to determine the comparison of three classes of Marginal Risk Model in predicting infant mortality among newborn babies at Kigali University Teaching Hospital (KUTH), Rwanda, 2016. The three models were the Bootstrap Marginal Risk Set Model (BMRS), Jackknife Marginal Risk Set Model and Marginal Risk Set Model. The study used 2117 newborns at the KUTH recorded from the 1st January to the 31st December 2016. The models revealed that female babies survive better than male babies, and the risk is higher for babies whose parents are under 20 years old as compared to other parents' age groups. The results also indicated that the risk is lower for underweight babies than babies with normal weight and overweight and even lower for babies with normal circumference of head as compared to those with relatively small heads. The results concluded that being abnormal in weight and head increased the risk of infant

mortality and avoidance of early pregnancy with proper clinical care would reduce infant mortality in Kigali.

[13] used chi-square analysis to determine the effect of migration on infant mortality between migrants and non-migrants in Lagos State. The target population for the survey was women ages 15-49 who had given birth to at least two children, and the study population consisted of both migrants and non-migrants in Lagos State Southern western part of Nigeria. A sample of 2000 was used, in which 1000 were migrants and 1000 were non-migrants. Results indicated that the current age of the mother had an impact on infant mortality and it was higher among migrants than non-migrants. The study also determined that there was significant difference in the effect of “sex of first dead child” on “the number of children who died with one year of birth” either migrants or non-migrants, and there was a significant difference in the effect of residence on infant mortality between the rural and the urban migrant dwellers ( $p < 0.05$ ) whereas there was no significant difference in the effect of place of residence on the infant mortality between the rural and urban-migrant dwellers ( $p > 0.05$ ). The results suggested that, concerning mother’ health and the child, there is need for specific programs such as family planning services to minimize the incident of high risk of births that occurs below the age of 18 years and over the age of 35 years. The study also concluded that medical services should be made available and affordable to encourage women to seek modern curative measures for themselves and their children.

[18] used three models (forced-entry, forward-selection, and backward selection) of multivariate logistic regression to identify trends and risk factors for infant mortality in the Lao People’s Democratic Republic. The study used 53,727 live births and 2189 women from the 2017 Lao Social Indicator Survey. Results indicated that the estimated infant mortality rate decreased from 191 per 1000 live births in 1978-1987 to 39 in 2017 and the factors that were associated with the high infant mortality in all three models of multivariate logistic regression were auxiliary nurses as birth attendants compared to doctors, male infants, and small birth size compared to average in all 2189 women; and 1-3 antenatal care visits compared to four visits, auxiliary nurses as birth attendants compared to doctors, male infants, postnatal baby checks, and being pregnant at the interview in 1950 women whose infants’ birth size was average or large. The study concluded that maternal, child healthcare and family planning should be strengthened for instance upgrading auxiliary nurses to mid-level nurses and antenatal care quality should be improved.

[1] used the Kaplan-Meir method and Cox proportional hazards regression model to identify the survival status and the proximate determinants associated with the infant mortality. The study used the 2016 Ethiopian Demographic and Health Survey where records of all 10,641 live births and survival data of all 2826 infants born 5 years before the survey were reviewed. The results of Kaplan-Meir estimation determined that about 65% infant deaths occurred in the early months of life immediately after birth and reduced in the later months of follow-up time, and the results of Cox proportional hazard model indicated that mothers’ level of education, preceding birth interval, plurality, size of child at birth and sex of child were significant predictors of infant mortality. The study also showed that the risk of dying in infancy was lower for babies whose mothers had

secondary education level (RR = 0.68, 95% CI: 0.56-0.98), higher education level (RR = 0.51, 95% CI: 0.45-0.80), for preceding birth interval longer than 47 months (RR = 0.51, 95% CI: 0.27-0.92) and higher for birth interval shorter than 24 months (RR = 2.02, 95% CI: 1.40-2.92), for multiple births (RR = 4.07, 95% CI: 1.14-14.50), for very small size of infants (RR = 3.74, 95% CI: 1.73-8.12), for smaller than average size of infants (RR = 3.23, 95% CI: 1.40-7.41) and for female infants (RR = 1.26, 95% CI: 1.01-1.56) compared to the reference category. The study recommended that close monitoring and supporting reproductive age mothers to increase the uptakes of family planning and antenatal care follow-ups will increase the survival of infants.

## **2.3 SUMMARY OF THE LITERATURE REVIEW**

From the literature review, it is clear that infant mortality remains to be a major concern in various developing countries and several studies have been done to investigate infant mortality risks. Male infants were found to have high infant mortality compared to female infants. Children born to uneducated mothers were found to have high infant mortality compared to educated mothers. It was also identified that babies born to young mothers (less than 18 years) were at high risk of experiencing high mortality rates compared to babies born to mothers above 18 years.

## **2.4 LITERATURE GAP**

From the literature review, only few studies have applied machine learning methods to predict infant mortality risks using the best performing algorithm. In Kenya, most researchers have used various traditional methods to predict infant mortality using KDHS data, yet machine learning techniques are available and need to be applied for better predictions. All classification techniques in Machine learning and Deep learning algorithms need to be implemented in health studies especially when dealing with child mortality predictions, and then compare results from all the techniques to obtain better results.

## 3 RESEARCH METHODOLOGY

### 3.1 DATA DESCRIPTION

The data source for this study is secondary data from the 2014 Kenya Demographic and Health Survey (KDHS). The 2014 Kenya Demographic and Health Survey was a nationally probability sample survey of approximately 40,300 households (KDHS, 2014). The previous surveys were conducted in 1989, 1993, 1998, 2003, and 2008-09. The survey used a two-stage sample design based on the 2009 Kenya Population and Housing Census and was designed to produce representative estimates for most of the survey indicators at the national level, urban and rural areas separately, at the regional level (former provinces), and for selected indicators at the county level (KDHS, 2014). In the first stage, 1,612 clusters (995 rural and 617 urban) were selected from the master frame and in the second stage of selection households were selected systematically from an updated list of households, in which 25 households were selected from each cluster (KDHS, 2014).

The eligible women for the interview were all women aged 15-49 years who were either usual residents or visitors present in the selected household on the night before the survey. Out of 15,317 women who were identified as eligible for the full Woman's Questionnaire interview, 14,741 were successively interviewed with a response rate of 96%, and out of 16,855 women who were identified as eligible for the short Woman's Questionnaire interview, 16,338 were successively interviewed with a response rate of 97% (KDHS, 2014). The response rates were generally lower in urban areas than rural areas and the reason being both eligible men and women were not available at home despite repeated visits to the household (KDHS, 2014).

The unit of analysis includes infants with a total sample size of 40,300 selected from 1,612 clusters across Kenya. This is based on children's data obtained from retrospective information from mothers about their children that died before attaining age 1 within the 5 years preceding the survey (2008-09 to 2014).

The 2014 KDHS dataset was downloaded from the Kenya National Bureau of Statistics (KNBS) website, Excel 2016 was used to open the data set, and R programming language (version 4.0.5) and the caret package [15] will be used to perform data processing and analysis.

### 3.2 Study Variables and Measurements

In this study, the outcome of interest was infant mortality measured as a binary outcome. Infant mortality was measured as being alive (coded as 0) or dead (coded as 1).

For independent variables (features), the modified version of Mosley and Chen's [23] conceptual framework was used. Factors related to infant mortality were grouped into three levels, namely community factors, socio-economic factors and proximate factors [23]. The community level factors consisted of geographical region (Nairobi, Central, Coast, Eastern, North Eastern, Nyanza, Rift Valley and Western) and place of residence (urban/rural). The socio-economic factor is mother's education level (No education, primary, secondary and higher), marital status (Divorced/separated, Married/Living with Partner, Widowed, never in Union), family size, time to water source and wealth index (Poorer, Poorest, Middle, Richer, Richest). The proximate factors consist of mother's age group (15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49), age of the mother at first birth, number of births in the last three years, sex of child (male or female), source of drinking water (Borehole, Rainwater, Waterbodies, Public, Protected, Unprotected, Piped, other), type of toilet facility (Flush, latrine, No facility, other) and the place of delivery (Government, Private, Mission, Home, Other).

### 3.3 Supervised machine learning methods

The three widely used machine learning algorithms - Logistic Regression (LR), K Nearest-Neighbor (KNN) and Random Forest (RF) models will be used to predict infant mortality in Kenya using the 2014 KDHS data.

#### 3.3.1 Logistic Regression

Logistic regression is the analysis to conduct when the dependent variable is binary (0 or 1) and it is widely used in population health research as an inferential tool. According to [5], independent variables ( $x$ ) are combined linearly using weights or coefficient values to predict the dependent variable ( $y$ ).

The logistic regression equation is given as:

$$\text{Logistic regression}(p) = \ln(p/1 - p) \quad (1)$$

#### Steps performed by logistic regression algorithm

**Step 1:** Input: Set of (input, output) training pair samples; call the input sample features  $x_1, x_2$  to  $x_n$  and the output results be  $y$ . There can be lots of input features  $x_i$ .

**Step 2:** Let  $p(x)$  be a linear function of  $x$ . Every increment of a component of  $x$  would add or subtract so much to the probability.

**Step 3:** Calculate odds ratio which is odds in favor of a particular event.



$$\text{Odds} = p/1 - p$$

**Where:**

*p* stands for the probability of the positive event.

**Step 4:** Define the logit function to calculate the logarithm of the odds ratio.

$$\text{logit}(p) = \log p/(1-p)$$

**Step 5:** Logit function takes input values in the range 0 to 1 and transforms them to values over the entire real number range, which express the linear relationship between feature values and the log-odds.

**Step 6:** Now to predict the probability in order to classify the class, use logistic function /sigmoid function.

**Step 7:** Output: Set of weights  $w$  (or  $W_i$ ), one for each feature, whose linear combination predicts the value of  $y$ .

### 3.3.2 K-Nearest Neighbor

According to [3] k-NN classifier is one of the simplest and most widely used in such classification algorithms and was proposed in 1951 by Fix and Hodges, and modified by Cover and Hart. k-NN technique can be used in both classification and regression problems. k-NN depends on calculating the distance between the tested, and the training data samples in order to identify its nearest neighbors, and the tested sample is then assigned to the class of its nearest neighbor [3].

According to [3], the k-value in k-NN stands for the number of nearest neighbors. When  $k = 1$ , the new data object is assigned to the class of its nearest neighbor, and the neighbors are taken from a set of training data objects where the classification is already known. K-NN works best with numerical data and various numerical measures have been used such as Euclidean, Manhattan, Minkowsky, City-block, and Chebyshev distances. The most widely used distance function is Euclidean distance.

The Euclidean distance function is given as:

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (2)$$

### Steps performed by k-Nearest Neighbor algorithm

**Step 1:** Finding the number of nearest neighbors (k-values).

**Step 2:** Calculating the distance between the test sample and all the training samples.

**Step 3:** Sorting the distance and finding the nearest neighbors based on the  $k^{th}$  minimum distance.

**Step 4:** Assembling the categories of the nearest neighbors.

**Step 5:** Utilizing the simple majority of the category of nearest neighbors as the prediction value of the new data object.

### 3.3.3 Random Forest

According to [5], random forest creates multiple decision trees and makes it random. Random forest builds multiple decision trees and merges them to get a more accurate and stable prediction [5].

According to [5], random forest is a meta estimator that fits several decision tree classifiers on various sub-samples of the data set and it uses averaging to improve the predictive accuracy and control over-fitting. The main advantage of the random forest is that it can be used for both classification and regression problems. Also, it is easy to use random forest algorithms because hyper-parameters often produce a good prediction result and the number of hyper-parameters is not that high hence easy to understand.

### Steps performed by random forest algorithm

Input: Set of (input, output) training pair samples; call the input sample features  $x_1, x_2$  to  $x_n$  and the output result as  $y$ .

**Step 1:** Randomly select “k” features from total “m” features of data set.

**Where  $k < M$**

**Step 2:** Among the “k” features, calculate the node “d” using the best split point.

**Step 3:** Split the node into daughter nodes using the best split.

**Step 4:** Repeat 1 to 3 steps until “l” number of nodes has been reached.

**Step 5:** Build forest by repeating steps 1 to 4 for “n” number times to create “n” number of trees.

**Output:** On average it takes all the predictions, which cancels out the biases, and attains performance by selecting the best feature from decisions instead of most important feature [5].

### 3.4 Training and testing data

Randomized split of 80% to 20% was done; where 80% of the total sample ( $0.80 * 4833 = 3866$ ) was used as trained data to prepare the models and the remaining 20% of the random sample ( $0.20 * 4833 = 967$ ) was used as a test data to predict the measures of model performance [7].

### 3.5 Performance Evaluation

The algorithm evaluation is mostly judged by prediction accuracy [24] and according to [3], the most widely used technique for summarizing the performance of supervised machine learning models is the confusion matrix.

#### 3.5.1 Confusion Matrix

According to [5] a confusion matrix is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known and it is a table with 4 different combinations of predicted and actual values in the case for a binary classifier.

According to [5], a true positive is an outcome where the model correctly predicts the positive class while a true negative is an outcome where the model correctly predicts the negative class.

#### Metrics computed from a confusion matrix

Model accuracy metrics are metrics that show how well the model performs in predicting the dead and alive cases. For this study, the most popular performance metrics were calculated from a confusion matrix. The metrics were sensitivity, specificity, positive predictive values and negative predictive values. According to [7], sensitivity refers to the proportion of subjects who have dead cases and give positive test results, specificity refers to the proportion of subjects who are alive and give negative test results, positive predictive value refers to the proportion of results that are true positives (truly dead) and negative predictive value refers to the proportion of negative results that are true negatives (truly alive).

## **Accuracy**

Defined as the fraction of predictions that the model got correct.

According to [8], accuracy equation is given as;

$$\mathbf{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

### **Where;**

TP = True Positive

TN = True Negative

FP = False Positive

FN = False Negative

### ***The formula for sensitivity is given as:***

$$\text{Sensitivity} = TP / (TP + FN)$$

Specificity is calculated as:

$$\text{Specificity} = TN / (TN + FP)$$

Positive predictive value is calculated as:

$$\text{Positive predictive value (PPV)} = TP / (TP + FP)$$

Negative predictive value is calculated as:

$$\text{Negative predictive value (NPV)} = TN / (TN + FN)$$

---

### 3.5.2 Receiver Operating Characteristics (ROC) curves and Area Under Curve (AUC)

Receiver Operating Characteristics (ROC) and Area Under Curve (AUC) metrics were used to evaluate model performance in differentiating between dead and alive cases. According to [10] receiver operating characteristic curves compares sensitivity versus specificity across a range of values for the ability to predict a dichotomous outcome and the overall accuracy is expressed as area under the ROC curve (AUC) and provides a useful parameter for comparing test performance between dead and alive cases.

AUC gives a summary of the ROC curve, hence the higher the AUC, the better the performance of the model at differentiating between positive classes and negative classes [10]. For the best performing model, a measure of variable importance for the model (Mean Decrease in Gini) was calculated for each variable.

## 4 CHAPTER 4: DATA ANALYSIS AND RESULTS

### 4.1 RESULTS

#### 4.1.1 Introduction

This section reports the descriptive of the data for the several variables under study, presenting the results of the three models (logistic, random forest and K-nearest neighbors) and interpretations.

#### 4.1.2 Descriptive results of the background characteristics

##### 4.1.2.1 Categorical Variables

**Table 1. Summary Statistics of study variables**

<b>Variable</b>	<b>Frequency</b>	<b>Percentage</b>	<b>P-value</b>
<b>Child alive</b>			
No	716	14.9%	
Yes	4076	85.1%	
<b>Age of the mother</b>			<b>p&lt;0.0000</b>
15 – 19	476	9.9%	
20 – 24	1330	27.8%	
25 – 29	1366	28.5%	
30 – 34	831	17.3%	
35 – 39	562	11.7%	
40 – 44	193	4.0%	
45 – 49	34	0.7%	

Variable	Frequency	Percentage	P-value
<b>Region</b>			<b>p=0.0379</b>
Central	300	6.3%	
Coast	623	13.0%	
Eastern	700	14.6%	
Nairobi	125	2.6%	
North Eastern	352	7.3%	
Nyanza	661	13.8%	
Rift Valley	1597	33.3%	
Western	434	9.1%	
<b>Place of residence</b>			<b>p=0.3273</b>
Rural	3205	66.9%	
Urban	1587	33.1%	
<b>Education level</b>			<b>p=0.0034</b>
No education	1045	21.8%	
Primary	2491	52.0%	
Secondary	954	19.9%	
Higher	302	6.3%	
<b>Child sex</b>			<b>p=0.0188</b>
Female	2339	48.8%	
Male	2453	51.2%	
<b>Place of delivery</b>			<b>p=0.0645</b>
Government	2314	48.3%	
Private	148	3.1%	
Mission	315	6.6%	
Home	1964	41.0%	
Other	51	1.1%	

Variable	Frequency	Percentage	P-value
<b>Marital status</b>			<b>p=0.0002</b>
Divorced/Separated	254	5.3%	
Married	4041	84.3%	
Widowed	105	2.2%	
Never in union	392	8.2%	
<b>Source of drinking water</b>			<b>p=0.5266</b>
Borehole	412	8.6%	
Rainwater	129	2.7%	
Water bodies	1144	23.9%	
Public	701	14.6%	
Protected	756	15.8%	
Unprotected	618	12.9%	
Piped	856	17.8%	
Other	176	3.7%	
<b>Type of toilet facility</b>			<b>p=0.6172</b>
Flush toilet	417	8.7%	
Latrine	3196	66.7%	
No facility/bush/field	1166	24.3%	
Other	13	0.3%	
<b>Wealth Index</b>			<b>p=0.2129</b>
Poorer	970	20.2%	
Poorest	1697	35.4%	
Middle	808	16.9%	
Richer	701	14.6%	
Richest	616	12.9%	
<b>The total number of infants were 4,792 (n = 4,792)</b>			



Of the 4,792 infants in the sample, 14.9% did not survive but 85.1% survived. It was found that of the 4,792 infants, majority of them, 33.3% came from the Rift Valley region, 14.6% from Eastern region, 13.8% from Nyanza region and 13.0% from the Coast region. From Western there were 9.1%, North Eastern 7.3%, Central 6.3% and Nairobi there were 2.6%. On the age of the mother, infants whose mothers were of age group 25 – 29 were the majority with 28.5% of the total sample size while infants whose mothers were of age group 20 – 24 were 27.8% of the total number of infants.

From age group 30 – 34, infants were 17.3% of the total number, from age group 35 – 39, infants were 11.7% of the total number, from age group 15 – 19, infants were 9.9% of the total number of infants, from age group 40 – 45, infants were 4.0% of the total and from age group 45 – 49, infants were 0.7% of the total number. It was found out that rural dwellers were the majority in the study with 66.9% of the total number of infants. On the education level, infants whose mothers had achieved primary level were the majority with 52.0% of the total number while those that achieved no education were 21.8%. Infants whose mothers had achieved secondary were 19.9% of the total number and those had achieved higher education were 6.3%.

On the sex of the child, male infants were the majority with 51.2% of the total number of infants. On the place of delivery, majority of the infants were delivered in government hospitals with 48.8% of the total number, while infants delivered at home were 41.0% of the total sample size. Infants delivered in mission hospitals were 6.6% of the total, infants from private hospital were 3.1% of the total and infants delivered from other health institutions were 1.1%. On the marital status, majority of the infants were from married families with 84.3% of the total while infants whose parents were never in union were 8.2% of the total. Infants from divorced or separated families were 5.3% of the total and infants from widowed families were 2.2% of the total sample.

On the source of drinking water, majority of infants were from families that use water from lakes, ponds, streams, rivers with 23.9% of the total infants, piped water were 17.8%, protected water were 15.8%, public were 14.6%, unprotected water were 12.9%, borehole water were 8.6%, rain water were 2.7% and the rest were 3.7% of the total infants. On type of toilet facility, majority used latrines with 66.7% of the total infants, 24.3% of the total had no facility, 8.7% of the total infants used flush toilet and 0.3% of the total used other facilities. It was also found that infants from the poorest families were the majority with 35.4% of the total infants, followed by infants from poorer families with 20.2% of the total. Infants from middle class families were 16.9% of the total infants, from richer families were 14.6% of the total and from the richest families were 12.9% of the total infants. Among the 10 categorical variables given above, only 5 variables were significant in determining infant mortality. The variables were age of the mother, region, education level, sex of the child and marital status. The rest of the variables were not significant and were dropped.

#### 4.1.2.2 Continuous Variables

**Table 2. Description of continuous variables**

Variable	Min/Max	Mean (SD)	p-value
<i>Age of the mother at first birth</i>	6 and 39	19.4 (3.5)	p<0.0000
<i>Number of births in the last three years</i>	0 and 4	1.3 (0.5)	p<0.0000
<i>Family size</i>	0 and 12	3.2 (2.1)	p<0.0000
<i>Time to water source</i>	0 and 580	30.4 (48.5)	p<0.0000

The minimum age of the mother at her first birth was 6 while the maximum was 39. The average age of the mother at first birth was 19.4 with a standard deviation of 3.5. The minimum number of births in the last three years was 0 while the maximum was 4. The average number of births in the last three years was 1.3 with a standard deviation of 0.5. The minimum number of family size was 0 while the maximum was 12. The average number of family size was 3.2 with a standard deviation of 2.1. For the time to water source, the minimum time was 0 and the maximum time was 580 with an average time of 30.4 and standard deviation of 48.5. The study also found that all the four continuous variable were significant in determining infant mortality.

#### 4.1.3 Predicting Infant Mortality

**Table 3. Results of the three machine learning models**

Confusion Matrix		Random Forest		Logistic Regression		KNN Model	
		Predicted		Predicted		Predicted	
		<i>Alive</i>	<i>Dead</i>	<i>Alive</i>	<i>Dead</i>	<i>Alive</i>	<i>Dead</i>
<b>Observed</b>	<b>Alive</b>	808	21	806	124	814	137
	<b>Dead</b>	7	122	9	19	1	6
			%		%		%
<b>Accuracy</b>			97.1		86.1		85.6
<b>Sensitivity</b>			85.3		13.2		4.2
<b>Specificity</b>			99.1		98.9		99.9
<b>Positive predictive value</b>			94.6		67.9		85.7
<b>Negative predictive value</b>			97.5		86.7		85.6
<b>AUC</b>			92.2		56.1		52.0

The table above gives the results of the three machine learning models namely logistic regression, random Forests, and the K-nearest neighbor models. The infant mortality prediction accuracy was found to be high for all the three models, with random forest model having the highest overall

accuracy of 97.1% using the test data set, followed by logistic regression with 86.1% and K-nearest neighbors model with 85.6%. For sensitivity, the random forest had high sensitivity of 85.3% implying that among the three models, it was more accurate in identifying the dead cases.

All the three models had high specificity with random forest and K-NN model having the highest specificity meaning that the models were good in identifying the alive cases. However, the random forest model correctly identified 95% of the real dead cases ( $122 / (122 + 7)$ ) and 98% of real alive cases ( $808 / (808 + 21)$ ), meaning that the model is relatively better at predicting both real dead cases (positive) and alive cases (negative). The rest of the models (KNN and logistic regression) gave lower positive and negative predictive values compared to random forest model.

#### 4.1.4 Receiver Operating Characteristics (ROC) Curve

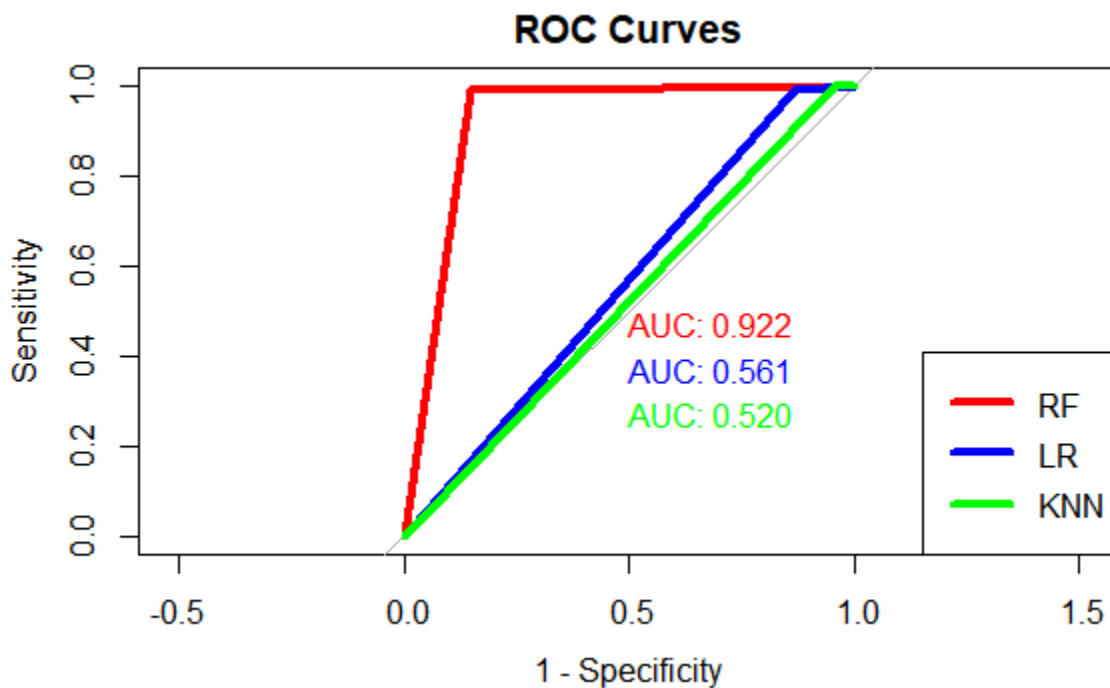


Figure 1. ROC Curves for the three models

The figure above shows the ROC curves for the three models (Random forest, logistic regression and K-nearest models). From the graph, the curve of the random forest model shows the highest AUC value (Area Under the Curve) which is approximately 92%, implying that it is the best model at classifying dead and alive cases when compared with the rest of the models (KNN and logistic regression).

#### 4.1.5 Variable importance measures for the random forest model

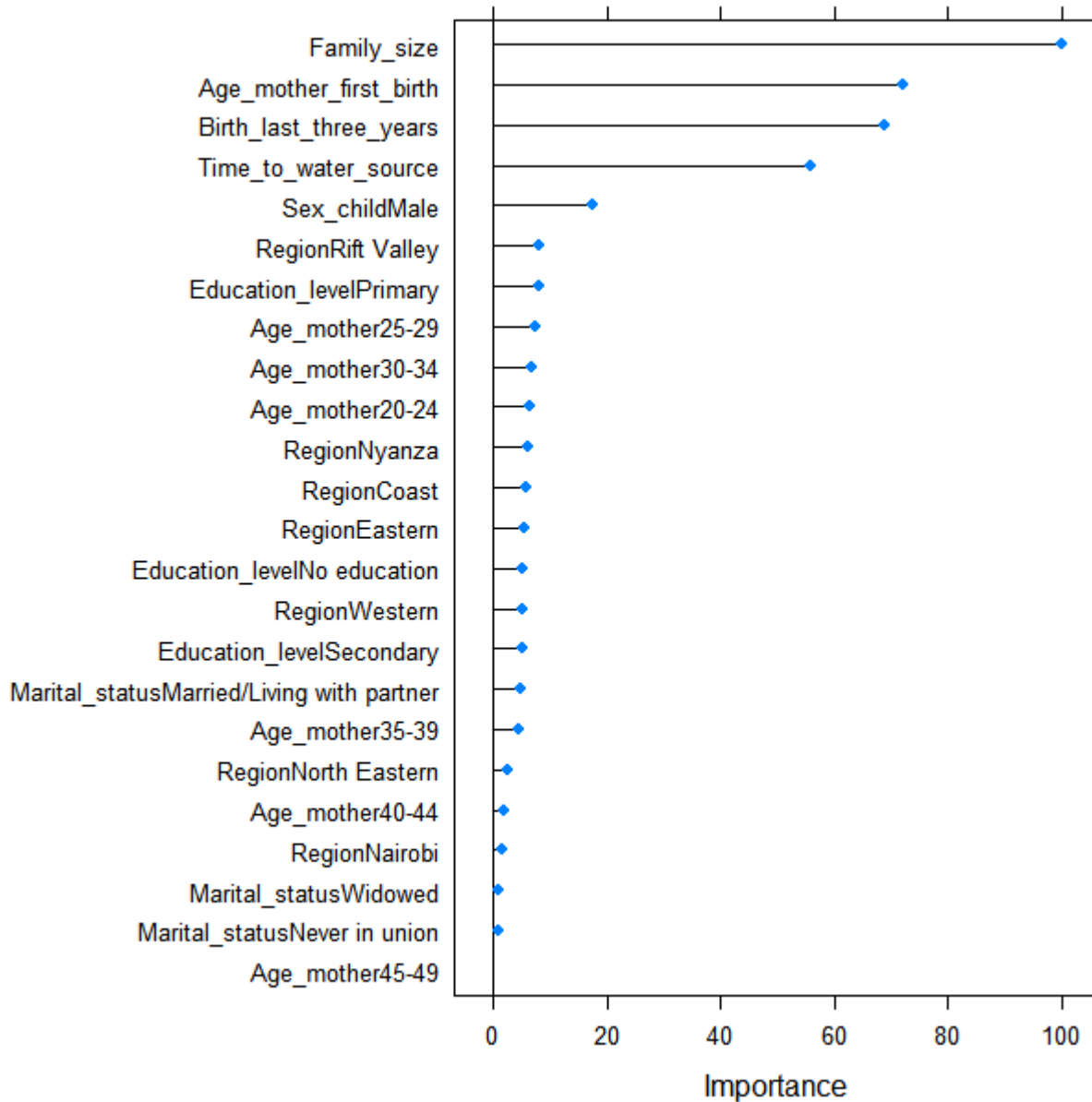


Figure 2. Variable Importance for the random forest model

Variable importance measure refers to how important a variable is for predicting infant mortality across all the cross-validation estimates [7]. The figure below shows the variable importance measures for the random forest model (categories of variables based on their Mean Decrease in Gini coefficient) with family size, age of the mother during her first birth and number of births in the last three years as the top 3 variables in the model. The other important factors that appeared in the top five variables were time to water source and sex of child (male).

## 5 CHAPTER 5: DISCUSSION AND CONCLUSION

### 5.1 Discussion

The results for this study found out that the random forest model had higher prediction accuracy and AUC compared to logistic regression and K-nearest neighbors, hence it's the best performing model. Another study done in Ethiopia [7] compared three machine learning models (Random forest, K-Nearest Neighbors and Logistic Regression) and the traditional logistic regression model. The results indicated that Random Forest model was the best performing model with 97.1% accuracy, followed by logistic regression with 86.1% accuracy and lastly K-NN with 85.6% accuracy. The reason was that the random forest model considers the outcomes from many different decision trees, thus more accurate compared to others [35].

The results of the model showed that family size, age of the mother at her first birth, the number of births for the last three years, time to water source and sex of the child were among the top 5 important predictors of infant mortality in Kenya and without the variables the model accuracy decreases.

From the findings of the best performing machine learning model, male children showed importance in predicting infant mortality compared with female children. Another study in Ethiopia [1] showed that male children were at higher risk of dying before celebrating their first birthday. This was also supported by other studies in which infant mortality rate was higher for males than females [36] and [22]. It has been studied that male children are at higher risk of dying in the first month of life because of high vulnerability to infectious diseases [7]. The reason may be that female infants have a biological advantage against many causes of death than boys hence less vulnerable to infectious diseases during their first months of life [1].

### 5.2 Conclusion

The study used three supervised machine learning algorithms to predict infant mortality in Kenya and identify important risk factors that will help in policy making. Random forest models provided a better predictive power than logistic regression and K-nearest neighbors in predicting infant mortality in Kenya. The model also revealed some important predictors of infant mortality, therefore the model can be used for policy making decisions regarding the survival of infants in Kenya. Factors such as family size, age of the mother at her first birth, the number of births for the last three years, time to water source and sex of the child play a major role in childhood survival chances in Kenya especially infants.

### **5.3 Future Research**

Future work should be done using regression methods to investigate how the factors affect the infant mortality quantitatively.

### **5.4 Limitation of the Study**

The KDHS data had several missing values for important predictors, hence it was difficult to include them in this study.

## References

- [1] Masrie Getnet Abate, Dessie Abebaw Angaw, and Tamrat Shaweno. Proximate determinants of infant mortality in ethiopia, 2016 ethiopian demographic and health surveys: results from a survival analysis. *Archives of Public Health*, 78(1):1–10, 2020.
- [2] Joseph Misati Akuma. Regional variations of infant mortality in kenya: Evidence from 2009 kdhs data. *Mediterranean Journal of social sciences*, 4(9):425–425, 2013.
- [3] Najat Ali, Daniel Neagu, and Paul Trundle. Evaluation of k-nearest neighbour classifier performance for heterogeneous data sets. *SN Applied Sciences*, 1(12):1–15, 2019.
- [4] Ethem Alpaydin. *Introduction to machine learning*. MIT press, 2020.
- [5] Shweta Bajpai, Monika Semwal, Ram Bajpai, Josip Car, Andy Hau Yan Ho, et al. Health professions’ digital education: Review of learning theories in randomized controlled trials by the digital health education collaboration. *J Med Internet Res*, 21(3):e12912, 2019.
- [6] Samir Kumar Bandyopadhyay and Shawni Dutta. Machine learning approach for confirmation of covid-19 cases: Positive, negative, death and release. *MedRxiv*, 2020.
- [7] 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):1–16, 2020.
- [8] Rung-Ching Chen, Christine Dewi, Su-Wen Huang, and Rezzy Eko Caraka. Selecting critical features for data classification based on machine learning methods. *Journal of Big Data*, 7:1–26, 2020.
- [9] Ashley N Dalrymple, Simon A Sharples, Nathan Osachoff, Adam Parker Lognon, and Patrick John Whelan. A supervised machine learning approach to characterize spinal network function. *Journal of neurophysiology*, 121(6):2001–2012, 2019.
- [10] Christopher M Florkowski. Sensitivity, specificity, receiver-operating characteristic (roc) curves and likelihood ratios: communicating the performance of diagnostic tests. *The Clinical Biochemist Reviews*, 29(Suppl 1):S83, 2008.
- [11] Paul Gatabazi, Sileshi Fanta Melesse, and Shaun Ramroop. Comparison of three classes of marginal risk set model in predicting infant mortality among newborn babies at kigali university teaching hospital, rwanda, 2016. *BMC pediatrics*, 20(1):1–11, 2020.
- [12] Guang-Bin Huang, Qin-Yu Zhu, and Chee-Kheong Siew. Extreme learning machine: theory and applications. *Neurocomputing*, 70(1-3):489–501, 2006.

- 
- [13] Erhabor Idemudia and Klaus Boehnke. *Psychosocial Experiences of African Migrants in Six European Countries: A Mixed Method Study*. Springer Nature, 2020.
- [14] Tony Jebara. *Machine learning: discriminative and generative*, volume 755. Springer Science & Business Media, 2012.
- [15] Max Kuhn, Jed Wing, Stew Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, et al. caret: Classification and regression training. r package version 6.0-86. Available at: <https://cran.r-project.org/web/packages/caret/caret.pdf> (accessed March 20, 2020), 2020.
- [16] Danning Liu. Determinants of infant mortality in kenya-analysis of kenya dhs 2003 and 2008/9. 2014.
- [17] Dianbo Liu, Leonardo Clemente, Canelle Poirier, Xiyu Ding, Matteo Chinazzi, Jessica T Davis, Alessandro Vespignani, and Mauricio Santillana. A machine learning methodology for real-time forecasting of the 2019-2020 covid-19 outbreak using internet searches, news alerts, and estimates from mechanistic models. *arXiv preprint arXiv:2004.04019*, 2020.
- [18] Viengsakhone Louangpradith, Eiko Yamamoto, Souphalak Inthaphatha, Bounfeng Phoumalaysith, Tetsuyoshi Kariya, Yu Mon Saw, and Nobuyuki Hamajima. Trends and risk factors for infant mortality in the lao people’s democratic republic. *Scientific Reports*, 10(1):1–11, 2020.
- [19] Peter M Macharia, Emanuele Giorgi, Pamela N Thurair, Noel K Joseph, Benn Sartorius, Robert W Snow, and Emelda A Okiro. Sub national variation and inequalities in under-five mortality in kenya since 1965. *BMC public health*, 19(1):1–12, 2019.
- [20] Nivedita Manohar Mathkunti and Shanta Rangaswamy. Machine learning techniques to identify dementia. *SN Computer Science*, 1(3):1–6, 2020.
- [21] Kathy McKay, Allison Milner, and Myfanwy Maple. Women and suicide: Beyond the gender paradox. *International Journal of Culture and Mental Health*, 7(2):168–178, 2014.
- [22] Hayelom Gebrekirstos Mengesha and Berhe W Sahle. Cause of neonatal deaths in northern ethiopia: a prospective cohort study. *BMC public health*, 17(1):1–8, 2017.
- [23] W Henry Mosley and Lincoln C Chen. An analytical framework for the study of child survival in developing countries. *Bulletin of the world Health Organization*, 81:140–145, 2003.
- [24] Iqbal Muhammad and Zhu Yan. Supervised machine learning approaches: A survey. *ICTACT Journal on Soft Computing*, 5(3), 2015.
- [25] LJ Muhammad, Md Milon Islam, Sani Sharif Usman, and Safial Islam Ayon. Predictive data mining models for novel coronavirus (covid-19) infected patients’ recovery. *SN Computer Science*, 1(4):1–7, 2020.



- 
- [26] Furqan Rustam, Aijaz Ahmad Reshi, Arif Mehmood, Saleem Ullah, Byung-Won On, Waqar Aslam, and Gyu Sang Choi. Covid-19 future forecasting using supervised machine learning models. *IEEE access*, 8:101489–101499, 2020.
- [27] Paul Sajda. Machine learning for detection and diagnosis of disease. *Annu. Rev. Biomed. Eng.*, 8:537–565, 2006.
- [28] R Saravanan and Pothula Sujatha. A state of art techniques on machine learning algorithms: a perspective of supervised learning approaches in data classification. In *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 945–949. IEEE, 2018.
- [29] Daniel R Schrider and Andrew D Kern. Supervised machine learning for population genetics: a new paradigm. *Trends in Genetics*, 34(4):301–312, 2018.
- [30] Pramod Singh. Supervised machine learning. In *Learn PySpark*, pages 117–159. Springer, 2019.
- [31] Aayush Kumar Singha, Devaraj Phukan, Sneha Bhasin, and Ramraj Santhanam. Application of machine learning in analysis of infant mortality and its factors. *Work Pap*, pages 1–5, 2016.
- [32] Rubita Sudirman, Narges Tabatabaey-Mashadi, and Ismail Ariffin. Aspects of a standardized automated system for screening children’s handwriting. In *2011 First International Conference on Informatics and Computational Intelligence*, pages 49–54. IEEE, 2011.
- [33] Nur Amalina Diyana Suhaimi and Hafiza Abas. A systematic literature review on supervised machine learning algorithms. *PERINTIS eJournal*, 10(1):1–24, 2020.
- [34] Jamal Uddin and Zakir Hossain. Predictors of infant mortality in a developing country. *Asian Journal of Epidemiology*, 1(1):1–16, 2008.
- [35] Shahadat Uddin, Arif Khan, Md Ekramul Hossain, and Mohammad Ali Moni. Comparing different supervised machine learning algorithms for disease prediction. *BMC medical informatics and decision making*, 19(1):1–16, 2019.
- [36] Berhe Weldearegawi, Yohannes Adama Melaku, Semaw Ferede Abera, Yemane Ashebir, Fisaha Haile, Afework Mulugeta, Frehiwot Eshetu, and Mark Spigt. Infant mortality and causes of infant deaths in rural ethiopia: a population-based cohort of 3684 births. *BMC public health*, 15(1):1–7, 2015.
- [37] Li Yan, H Zhang, Jorge Goncalves, Yang Xiao, Maolin Wang, Yuqi Guo, Chuan Sun, Xiuchuan Tang, Liang Jin, Mingyang Zhang, et al. A machine learning-based model for survival prediction in patients with severe covid-19 infection. 2020.
- [38] Danzhen You, Lucia Hug, Simon Ejemyr, Priscila Idele, Daniel Hogan, Colin Mathers, Patrick Gerland, Jin Rou New, Leontine Alkema, et al. Global, regional, and national levels and trends in under-5 mortality between 1990 and 2015, with scenario-based projections

to 2030: a systematic analysis by the un inter-agency group for child mortality estimation.  
*The Lancet*, 386(10010):2275–2286, 2015.