Machine Learning Techniques to Identify Determinants of Infant and Child Mortality Based on Ethiopia Demographic and Health Surveys (EDHS)

By

Alemeshet Ketema Kabtiymer

A Thesis Submitted to

The Department of Computer Science and Engineering

School of Electrical Engineering and Computing

Presented in Partial Fulfillment of the Requirement for the Degree of Master's in Information Systems Engineering

Office of Graduate Studies

Adama Science and Technology University

Adama, Ethiopia

June 2019

Machine Learning Techniques to Identify Determinants of Infant and Child Mortality Based on Ethiopia Demographic and Health Surveys (EDHS)

Student Name: Alemeshet Ketema Kabtiymer

Advisor Name: Tilahun Melak(PhD.)

A Thesis Submitted to

The Department of Computer Science and Engineering

School of Electrical Engineering and Computing

Presented in Partial Fulfillment of the Requirement for the Degree of Master's in Information Systems Engineering

Office of Graduate Studies

Adama Science and Technology University

Adama, Ethiopia

June 2019

Approval of Board of Examiners

We, the undersigned, members of the Board of Examiners of the final open defense by Alemeshet Ketema Kabtiymer have read and evaluated his/her thesis entitled "Machine Learning Techniques to Identify Determinants of Infant and Child Mortality Based on Ethiopia Demographic and Health Surveys (EDHS)" and examined the candidate. This is, therefore, to certify that the thesis has been accepted in partial fulfillment of the requirement of the Degree of Master's in Information Systems Engineering

Approved By:

_____	_____	_____
Advisor	Signature	Date
_____	_____	_____
Chair Person	Signature	Date
_____	_____	_____
Internal Examiner	Signature	Date
_____	_____	_____
External Examiner	Signature	Date

DECLARATION

I hereby declare that this MSc Thesis is my original work, has not been presented for a degree in any other university, and all sources of materials used for this thesis have been properly acknowledged.

Alemeshet Ketema _____

Candidate Name signature

This MSc Thesis has been submitted for examination with my approval as thesis advisor:

Dr. Tilahun Melak _____

Advisor's Name signature

Date of submission: June 28, 2019

ACKNOWLEDGEMENTS

Firstly Thanks to almighty GOD and St. Virgin Mary, mother of Jesus Christ for giving me the ability to be where I am now and always with me where ever I go. You have done so much for me, I'm gland thanks, Amen.

Secondly, I extend my heartfelt grateful thank to my Advisor Tilahun Melak (Ph.D.) for the guidance to 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'm grateful thanks to my wife W/ro Selamawit Deres Hassen and my family as the whole, without whom this research wouldn't be succeeded. Brook Tesfaye, your helpful personality will always be a role in my heart.

Lastly not mean least I will thanks to the Samara University for sponsoring this thesis research.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

AOR: Adjacent Odd Ratio

AUC: Area under the Curve

CSA: Central Statistical Agency

CI: Confidence Interval

CV: Cross Validation

EMR: Electronic Medical Records

EDHS: Ethiopian Demographic and Health Surveys

FN: False Negative

FP: False Positive

FPR: False Positive Rate

IQR: inter-quartile range

LOO: Leave One Out

LR: Logistic Regression

MDG: Millennium development goal

MICE: Multivariate Imputation via Chained Equations

MoH: Ministry of Health

MR: Mortality Rate

NB: Naïve Bayes

NPV: Negative Predictive Value

PPV: Positive Predictive Value

RF: Random Forest

ROC: Receiver Operating Characteristic

ROSE: Random Over Sampling Example

SDG: sustainable development goal

SMOTE: Synthetic Minority Over-sampling Technique

SVM: Support Vector Machine

TN: True Negative

TP: True Positive

TPR: True Positive Rate

UN: United Nation

WEKA: Waikato Environment for Knowledge Analysis

WHO: World Health Organization

ABSTRACT

The Ethiopian government doing for the past two decades for attaining millennium development goals agenda for preventing childhood mortality by improving the child health's to change the country image to the rest of the world in reduction of childhood mortality. This study contributes some values in the improvement of childhood health by analyzing the determinants infant and child mortality by using machine learning techniques. Different reports indicate that the distribution of childhood mortality differs in the world. According to the United Nations Inter-Agency Group, 2017 report the global under-five mortality rate declined by 56 percent starting from 1990 to 2016. In 1990 the deaths per 1,000 live births were 93 percent and in 2016 counted 41 percent. Ethiopian demographic and health survey 2016 report of 2017 indicates that for the 5 years preceding the survey, the under-5 mortality rate was 67 deaths per 1,000 live births, and the infant mortality rate was 48 deaths per 1,000 live births. This study aimed to identify factors and developed predictive models using four supervised machine learning techniques namely C5.0 Decision tree, Random Forest, Support Vector Machine and Naïve Bayes algorithms using the 2016 EDHS dataset of 10,641 records. K-fold cross-validation, performance measures accuracy and Kappa and confusion matrix metrics used for performance measurements of models. SMOTE sampling method was used as the best sampling techniques. The child mortality rate was 60 deaths per 1000 live births. Number of under-five child in household (AOR = 16.08, (95% CI [13.3,19.44]), type of place of residence (AOR = 0.73, 95% CI [0.5,1.08]), source of drinking water (AOR = 0.97, [0.94,0.99]), wealth index difference (AOR = 0.97, [0.89,1.06]), family plan use (AOR = 1.03, 95% CI [0.99,1.07]), breastfeeding (AOR = 2.58, 95% CI [2.12,3.14]), place of delivery (AOR = 1.0062, 95% CI [0.9939,1.0187]), birth order number (AOR = 1.73, 95% CI [1.35,2.22]), and, preceding birth interval (AOR = 1.0075, [1.0017,1.0133]) were found to be determinants for child mortality. In this research work, Random Forest and C5.0 decision tree model had performance , accuracy (95.93%), (96.15%), sensitivity (65.26%), (65.78%), specificity (97.56%), (98.06%), Positive Predictive Value (PPV) (65.96%), (68.31%), Negative Predictive Value (NPV) 97.80%), (97.84%) respectively. The result obtained could support child health intervention programs in Ethiopia and used as a reference for the researcher.

Keywords: Infant and Child Mortality, Determinants, EDHS, Supervised machine learning, Sampling techniques, Metrics

CHAPTER 1: INTRODUCTION

1.1 Background

Ethiopia is one of the developing countries and working on different policies to reached and join a developed country. From this policy one of is the improvement of infant and child health's and reducing child mortality rate. The Ethiopian government doing for the past two decades for attaining millennium development goals agenda for preventing childhood mortality by improving the child health's. This improvement of the health of childhood by the government helps for the achievement of the government policy and the country has a good image to the rest of the world. This study contributes some values in the improvement of childhood health by analyzing the determinants infant and child mortality by using machine learning techniques. 2017 Ethiopian demographic health surveys reports indicate that the distribution of childhood mortality differs in the world. Infant and child mortality was defined according to Ethiopia Demographic and Health surveys, it was the death of babies between births to the fifth birthday and mortality rate (MR) was defined as the total number of infant and child deaths per 1,000 live births in a specified period of time. [1].

When we look at the distribution of childhood mortality, different report indicated that there was still a big issue in case of infant and child mortality. Globally, an estimated 6.6 million children under-five died in 2012, which 18,000 children under five dying every day. From this 3.2 million of the children did not reach their fifth birthday in 2012 which most died as a result of easily preventable infectious diseases in Africa continent's [2]. According to the United Nations Inter-Agency Group, 2017 report [3] the global under-five mortality rate declined by 56 percent starting from 1990 to 2016. In 1990 the deaths per 1,000 live births were 93 percent and in 2016 counted 41 percent. The burden of under-five deaths remains unevenly distributed. About 80 percent of under-five deaths occur in two regions, sub-Saharan Africa and Southern Asia. Six countries account for half of the worldwide under-five deaths, namely, India, Nigeria, Pakistan, Congo, Ethiopia, and China. The report indicates child mortality still remains one of the great wrongs of our modern world. According to 2017 Status report on Maternal Newborn Child and Adolescent Health [4] in 2015 79 countries globally have an under-five mortality rate above 25, and 47 of them not meet the proposed sustainable development goal (SDG) target of 25 deaths per 1000 live

births by 2030 if they continue their current trends in reducing under-five mortality. The report indicates that in 2015 under-five mortality rate for the continent was 76 deaths per 1,000 live births, which is 67% above the recommended SDG target of 25 deaths per 1,000.

Coming to Africa excluding North Africa 54 percent in under-five mortality from 180 deaths per 1,000 live births to 83 deaths per 1,000 live births from 1990 to 2015 was recorded. A great reduction in under-five mortality of 64 percent (73 per 1,000 in 1990 to 24 per 1,000 in 2015) was recorded in North Africa. The target goal of the SDG in 2030, in the annual rate of reduction of under-five mortality, should be 4.5% or more. According to this, only 6 countries in Africa have under-five mortality rates at or below the SDG target of 25 per 1,000 live births. These are Libya, Tunisia, Seychelles, Mauritania, Egypt and Cape Verde [4].

When we look Ethiopia in case of infant and child mortality based on Ethiopian demographic and health survey 2016 report of 2017 [1], the under-5 mortality rate was 67 deaths per 1,000 live births, and the infant mortality rate was 48 deaths per 1,000 live births. This means that 1 in 15 children in Ethiopia dies before reaching age 5, and 7 in 10 of the deaths occur during infancy. According to the report, there was a large variation in childhood mortality. Under-5 mortality ranges from a low of 39 deaths per 1,000 live births in Addis Ababa to a high of 125 deaths per 1,000 live births in Afar. The report also indicates the Under-5 mortality is higher in rural areas than in urban areas (83 versus 66 deaths per 1,000 live births). One of the factors for infant mortality rate declines with increases in the mother's education, falling from 64 deaths per 1,000 live births among children whose mothers have no education to 35 deaths per 1,000 live births among children whose mothers have more than secondary education. Boys are more likely to die in childhood than girls. In the report the gender gap is most pronounced in the neonatal period (within 1 month after birth), when male children are nearly twice as likely as female children to die (49 deaths compared with 26 deaths, per 1,000 live births, respectively.

In September 2016, the Human Rights Council resolution 33/11 indicate there is progress reduction of child mortality but not achieved by two thirds from 1990 to 2015 that the plan of Millennium Development Goal 4 planed by Ethiopia. The Council is made up of 47 United Nations Member States which are elected by the UN General Assembly [5].

The aim of this study was to identify factors for infant and child mortality and make a prediction model by applying machine learning algorithms and techniques based on Ethiopian Demographic and Health Survey (EDHS) 2016 dataset. The Ethiopian demographic and health survey (EDHS) 2016 was the new survey data collected in 2016 which the child record data contains 10,641 records and 1209 variable [1]. In this study all 1209 variables not used. The datasets were preprocessed and important features were selected from 1209 variable for model building. MICE (Multivariate Imputation via Chained Equations) [6] R package was used for handling missing data by creating multiple imputations as compared to a single imputation (such as the mean) takes care of uncertainty in missing values. For important features selection, Boruta algorithm [7] was used by using Boruta package in R. For prediction model supervised machine algorithms were applied on the selected features and best prediction model were selected by different model evaluation method.

For prediction and pattern finding in this study, machine learning techniques were used which was widely used in a variety of areas, including finance, driverless cars, image detection and speech recognition among others. In a world of high volume and varied datasets, machine learning techniques are an essential toolkit to provide actionable insights from the data [8]. For implementing the machine learning algorithms and techniques R Programming Language was used which an open source programming language developed by statisticians with the purpose of providing a diverse and high-quality open source statistical analysis language. The key strength of R is a wider scheme, with a massive form of newest packages, which build an implementation of complicated models quantity to one line of code.

1.2 Statement of the Problem

Infant and child mortality rates are basic indicators of a country's socioeconomic situation and quality of life (UNDP 2007) [1]. According to the National Planning Commission and united nation report [9] the childhood mortality not only concern Ethiopia but also too many Countries in the Africans. The report indicates that Ethiopia achieves Millennium Development Goals by two thirds by 2015. However, the mortality reduction was not uniform across the different childhood age groups (neonatal, infant, and under-five), geographic and socio-demographic population groups.

Ethiopia Demographic and Health Survey (2016 EDHS) [1] report indicate that for the 5-year period preceding the survey, the under-5 mortality rate is 67 deaths per 1,000 live births, and the infant mortality rate is 48 deaths per 1,000 live births. This means in another word 1 in 15 children in Ethiopia dies before reaching age 5, and 7 in 10 of the deaths occur during infancy. According to the report, 77% percent of currently married women have the potential for a high-risk birth and 62% percent of births have high mortality risks that are avoidable; 38% fall into a single high-risk category and 24% are in multiple high-risk categories. Regional differences also have large variations in childhood mortality which is Under-5 mortality ranges from a low of 39 deaths per 1,000 live births in Addis Ababa to a high of 125 deaths per 1,000 live births in Afar.

According to United Nations Children's Fund (UNICEF) in 2018 Bangladesh, Ethiopia, Guinea-Bissau, India, Indonesia, Malawi, Mali, Nigeria, Pakistan and the United Republic of Tanzania all together account for more than half of the world's newborn deaths [10].

According to the above-cited paper reports still, now infant and child mortality were high by different cases. Even if Ethiopia success the MDG in cases of infant and child mortality reduction there still varies in rural and urban. These indicate that it achieved the goal of the percentage in infant and child mortality not geographic and socio-demographic population groups.

Therefore, the goal of this study is identifying the factors that affect infant and child mortality by ranking the determinants and apply a machine learning algorithm to predict the infant and child mortality.

1.3 Objectives

1.3.1 General Objective

The general objective of this study was to identify determinants of infant and child mortality in Ethiopia and select the best predictive model by comparing a C5.0 Decision tree, Random Forest, Support Vector Machine and Naïve Bayes based on the performance of prediction on the new dataset.

1.3.2 Specific Objectives

In order to achieve the goal of the study, few experiments are performed as mentioned below:

- Perform data cleaning and preprocessing: includes basic operations, such as removing noise or outliers, mapping of missing and unknown values,
- Identify the major determinant attributes for the occurrences of infant and child mortality,
- Applied importance features selection techniques to select important features from the large dataset that related to "B5" which is dichotomous variable in the dataset that is a child "Alive" or "Died" group,
- Applied different sampling techniques for handling data imbalance problem,
- Perform or applied prediction models on the imbalanced and balanced data set,
- Building predictive models, implementing C5.0 Decision tree, Random Forest, Naïve Bayes and Support Vector Machine algorithms.
- Generate a rule-based model for analyzing the determinants of infant and child mortality,
- Compare models and select the best prediction model by Model evaluation technique.

1.4 Significance of the Study

The aim of this research was to identify the determinants of child mortality in Ethiopia based on 2016 EDHS dataset and build four prediction models. From the four models the best models were selected. The study contributed some information for professional's associations and other stakeholders such as NGOs in designing and implementing child health intervention programs and projects in the country. On the other hand, this study has been supported researchers in determining the appropriate model to use a given supervised machine learning models and different techniques for data preprocessing and importance feature selection methods.

1.5 Scope and Limitation of the Study

The study has focused on the prediction of the infant and child mortality by implementing machine learning algorithms based on the selected feature attributes from 2016 EDHS dataset. Some important variables that might affect infant and child mortality had many

missing values which were more than 50% missed. For building prediction model four supervised machine learning techniques namely, C5.0 Decision tree, Random Forest, Naïve Bayes and Support Vector Machine algorithms were used. The models were evaluated using a number of standard evaluation metrics K-fold cross-validation, performance measures accuracy and Kappa and confusion matrix.

1.6 Thesis Organization

This research was presented in six chapters. Chapter one highlights major issue relating to infant and child mortality at a global level and Ethiopia in particular and describe how this research was done. Chapter two discusses briefly literature review relating to factors associated with infant and child mortality. The way of different researcher done their study was discussed that related to this study. Chapter three discusses the experiment design and what method was used for data preprocessing and classification algorithms for prediction. The fourth Chapter describes the experiment of the research and prediction model was discussed. It comprises modeling, evaluation, presentation and interpretation of the rules discovered from the experiments. The fifth chapter describes result and discussion which discussed the result get from the experiment and the method apply on the data with its result outcome. Finally, in Chapter six conclusion and recommendations of the study are presented.

CHAPTER 2: LITERATURE REVIEW

This chapter provides a detailed review of the relevant literature to gain perception on this study. In the section some related works in the area of infant and child mortality prediction and determinants analysis presented. The distribution of childhood mortality from the global distribution to specific to Ethiopia has discussed. Data pre-processing, feature selection methods, handling of imbalance data, prediction model algorithms and model evaluation method were discussed with others researches on infant and child mortality prediction.

2.1 The trend in Infant and Child Mortality

As mentioned in chapter one in Background section different factors were accountable for the death of childhood which, the factors for mortality were changing in time. Depending on the factors that change in a time infrastructures facilities and awareness are changing from time to time. Therefore, it is necessary to give more emphasis on using the current data to identify the factors for infant and child mortality in order to achieve the goal and objectives of this study. To remind and shortly summarized Globally, an estimated 6.6 million children under_five died in 2012, from this 3.2 million of Africa continent's children did not reach their fifth birthday in 2012 which most died as a result of easily preventable infectious diseases [2]. According to United Nations Inter-Agency Group, 2017 report [3] the global under-five mortality rate declined by 56 percent through the year 1990 to 2016 that was 93 percent deaths per 1,000 live births in 1990 and in 41 percent counted in 2016. About 80 percent of under-five deaths occur in two regions, sub- Saharan Africa and Southern Asia. Six countries account for half of the worldwide under-five mortality, namely, India, Nigeria, Pakistan, the Democratic Republic of the Congo, Ethiopia, and China.

In Africa major reduction of 54 percent in under-five mortality from 180 deaths per 1,000 live births to 83 deaths per 1,000 live births from 1990 to 2015 occurred excluding North Africa. Only 6 countries in Africa have under-five mortality rates at or below the SDG target of 25 per 1,000 live births. These countries were Libya, Tunisia, Seychelles, Mauritania, Egypt and Cape Verde. From these Ethiopia has a rate of 59 per 1,000 live births under-five mortality [4]. This report indicates that around 60% of these children live in 10 countries: Angola, the Democratic Republic of the Congo, Ethiopia, India, Indonesia, Iraq, Nigeria, Pakistan, the Philippines, and Ukraine. Therefore this report indicates that Ethiopia was one

of the countries that not meet the proposed sustainable development goal (SDG) target of 25 deaths per 1000 live births by 2030 if they continue in this way in reducing under-five mortality. On the other hand United Nations Inter-Agency Group 2017 report [3] indicate that Ethiopia Under-five mortality rates were 58 per 1000 live birth in 2015 and the other report with Ethiopian demographic and health survey 2016 report of 2017 [1] indicates that for the 5-year period preceding the survey, the under-five mortality rate was 67 deaths per 1,000 live births, and the infant mortality rate was 48 deaths per 1,000 live births. This indicates the progress towards to infant and child mortality reduction is not reduced.

United Nations Children's Fund (UNICEF), 2018 report indicates that Bangladesh, Ethiopia, Guinea-Bissau, India, Indonesia, Malawi, Mali, Nigeria, Pakistan and the United Republic of Tanzania all together account for more than half of the world's newborn deaths [10]. The report indicates from 10 countries with the highest number of newborn deaths in 2016, and newborn mortality rates Ethiopia is one of them that ranked in the 5th place which understands still Ethiopia is not out from that the country highly contributes the infant and child mortality.

2.2 Related work

Many studies have been investigated in the determinants of childhood mortality in different socio-economic settings. Different factors may determine childhood mortality depending on the socio-economic and environmental situations. On this study, the researcher reviews different related work to this study and added value in the area.

In a study done by Shegaw investigated the potential applicability of data mining techniques to predict the risk of child mortality based upon community-based epidemiological datasets gathered by the BRHP (Butajira Rural Health Project) epidemiological study. The epidemiological database could be successfully mined to identify public health and socio-demographic determinants (risk factors) that are associated with infant and child mortality in rural communities. The researcher identify child mortality was associated with environment, household literacy, household health, age of the child, availability of windows in the house, household water, and even with household ethnicity. The researcher proved that these variations in child mortality among highland and lowland communities seems appropriate since there is an epidemic of malaria and meningitis in rural lowlands than in

highlands. The data was gained from the ten years' surveillance dataset of the BRHP epidemiological study. Neural network and decision tree data mining techniques were employed to build and test the models on sample dataset consisting 1,100 records taken randomly from the two classes of children (i.e. alive and died) of the ten years surveillance dataset of the BRHP epidemiological study which contains a total of 64,077 records was used to build and test both neural network and decision tree models. In features selection, the researcher has been used consultation with public health specialists, pediatricians, and researchers who have good knowledge and experience on the data set of BRHP database. [11].

The other study done by Desta has been investigated and identified the bio-demographic and socioeconomic determinants of infant and child mortality in Ethiopia in both the previous five years' period of the data on 2000 and 2005. In the study logistic regression was applied to determine the impact of bio-demographic and socioeconomic determinants on infant and child mortality. The researcher clearly show that among bio-demographic factors marital status, birth order, type of birth and preceding birth interval are the important proximate determinants for both infant and child mortality. However, breast feeding has a significant impact on infant mortality. Among Socioeconomic determinants education, household size and sex of household are the most important determinants for both infant and child mortality. This study also suggests that infant and child mortality widely varied between regions in Ethiopia due to unequal distribution of socioeconomic and health infrastructures [12].

The other study conducted by Be'emnetu in particular sites of Butajira rural health program has been explored the potential applicability of data mining to predict the determinants and pattern of under-five mortality for the Butajira rural health program sites. The researcher found under-five mortality was related with age of the child, days of exposure during episode, out-migration, in-migration, relationship, availability of windows in the house, type of roof during episode, number of oxen owned by family, radius of circular house in meters, number of land owned by family, distance to Butajira hospital, number of rooms in house, environment and source of water during episode. Two popular data mining algorithms C4.5 J48 Decision Trees and Naïve Bayes Classifier used to develop the predictive model with the dataset (11,600 cases). Using J48 decision tree rule based model was generated depend on the features. 10-fold cross-validation and 90% split test data used for performance

comparison, two predictive models. For handling missing values the researcher has been used field means [13].

The study has been done by Tesfahun also as Shegaw and Be'emnetu constructed predictive model using data mining techniques that identify and improve adult health status using BRHP (Butajira Rural Health Program) open cohort database. The researcher found six most important parameters/attributes that determine adult mortality are education status, outmigration, in migration, literacy, residence, and distance from the Butajira town. Decision tree and Naïve Bayes algorithms were used to build the predictive model on a sample dataset of 62,869 records and through three experiments and six scenarios. For feature selection survival Cox regression analysis was used in order to investigate significance selected features on adult mortality that cleaned data. The missed value field with mean. In this study, the researcher indicated that as compared to Bayes, the performance of the J48 pruned decision tree best performance [14].

The study has been conducted by [15] done based on the data from 2000, 2005 and 2011 of Ethiopia Demographic and Health Surveys (EDHS). In the study, Cox Regression was applied to assess the associations of child, mother and household characteristics with infant and under-five mortality. The researcher found that significant determinants of infant mortality are; sex of the child, birth size, birth interval, mother's education, mother's marital status and access to an improved toilet facility. Likewise, under-five mortality is influenced by the same characteristics. Region of residence was also a significant predictor of infant and under-five mortality. The researcher was not discussed in the paper how to handle the missing value and feature selection method and technique.

The study conducted by Jiregna identified factors that affect infant mortality based on 2014 Ethiopia mini Demographic and health survey dataset using multilevel count regression models and ZIP regression model. Logistic regression, survival analysis used for finding determinants of infant deaths and SAS version 9.2, STATA 13.0, R version 3.2.3 and ML WIN 2.36 statistical software packages used to analysis of the data. From Single level and multilevel count regression models with ZIP regression model region, household size, birth order, and birth interval were identified as significant factors for infant mortality. The study also showed that there is a significant regional variation of infant mortality [16].

The study conducted by Dechasa have been examined based on 2011 EDHS using logistic regression model were Breastfeeding, maternal education, family planning, preceding birth interval, presence of diarrhea, father's education, low birth weight and, age of the mother at first birth were found to be determinants for child mortality. From 11654 total number of children covered in the present study 846 (7.26%) were dead whereas 10808 (92.74%) were alive at the date of the survey and the child mortality rate was 72 deaths per 1000 live births. [17].

The study conducted by [18] has been identified Breastfeeding, maternal education, preceding birth interval, and low birth weight, family planning, and paternal education, age of the mother at first birth and presence of diarrhea for determinants of child mortality in this study. The researchers developed a web-based child mortality prediction model in Ethiopian by using local language data mining algorithm in limiting resource. The data source for this research was 2011 Ethiopian demographic and health survey data (EDHS 2011). Decision tree (using J48 algorithm) and rule induction (using PART algorithm) techniques were used on 11,654 records model prediction using Waikato Environment for Knowledge Analysis (WEKA) for Windows version 3.6.8. Statistical Package for Social Sciences (SPSS) version 20.0 used to identify determinants of child mortality by logistic regressions for Odds Ratio (OR) with 95% Confidence Interval (CI). For feature selection Information gain was applied and the imbalance, data was handled by Synthetic Minority Oversampling Technique (SMOTE).

The study conducted by Caluza has been designed a descriptive-correlation using data mining techniques specifically the decision tree algorithm called J48 algorithm for classifying child mortality rate. Results of the study indicate that annual population growth is highly correlated in predicting child mortality and generate three distinct rules respect to life expectancy at birth, annual population growth, and the gross domestic product [19].

The study conducted by [20] analyzed 35 Sub-Saharan-African countries, and 13 Asian countries, contains information on 824,017 children from 48 countries which were sampled in 131 individual surveys over the period from 1992 to 2015. In the study, Multilevel STAR model was used for the hierarchical structure of the data, and to allow for heterogeneity between countries. The study investigated strong non-linear effects for the baseline hazard, the household size, the age of the mother, the BMI of the mother, and the birth order of the

child. Additionally, they found considerable differences in determinants between Asian and Sub-Saharan Asian countries.

The study conducted by Sonkar worked on supervised machine learning to predict the mortality risk in the elderly using biomarkers. This study used Logistic Regression (LR) and Support Vector Machine (SVM) which were used in the prediction of the mortality in elderly people. Each technique can predict the mortality risk using biomarker features in their own particular way. The main goal of the study was to measure the classifier accuracy and compare the performance of both the models [21].

The study has been conducted by Caluza was Predicting Children Mortality using decision tree called J48 algorithm in classifying child mortality rate, life expectancy at birth, annual population growth, and the gross domestic product is. Annual population growth is highly correlated in predicting child mortality and by using the model three rules were generated [22].

The study conducted by [23] investigated and addressed basic medical care, during pregnancy and after delivery which was a major caused for infant mortality in India. Kendall rank correlation coefficient was used for feature selection and models were builds based on machine learning. This model trained using the past data records of both successful and unsuccessful births and the model used to predict the key factors like a high-risk mother and thus they can be given correct medical attention to mitigate the risk, hence reduce the infant mortality rate.

The study has been conducted by Munyamahoro applied the quantitative method in which was undertaken among 3015 mothers who have children less than five years in a rural area and urban area using Cox-regression model. The study was indicated that the major factors to death were inadequate access to basic medical care, home environmental problems, malnutrition, and another infectious disease. The study shows the wealth index difference, age, and education level of the mother affect the life of a child. This study has been conducted on a small dataset that was collected a sampled data by [24].

As we look in the above different study have been done by different researchers and they have been used different techniques to solve the problem in case of identifying the determinants of child mortality. Most of the determinants in the paper reviewed were

Mother's Characteristics: like mother's attainment of education, health service utilization, age at birth, marital status, and contraceptive use, literacy, education status; Child's Characteristics: Birth interval, sex of the child, and preceding birth interval; Household Characteristics: Household wealth, region, residence, availability of safe drinking water, and improved toilet facility.

The study here now conducted by the researcher differ and have some similarity from the above-reviewed papers in feature selection, in handling missing values, the technique used for predicted model and tools used. In this study, the researcher investigated the determinants of infant and child mortality based on the EDHS 2016 dataset. The study was determined the factors for childhood mortality of the whole region in Ethiopia and the determinants were analyzed by machine learning techniques using R tools and for some statistical summary and to convert the dataset format to comma separated value we used Statistical Package for Social Sciences (SPSS) version 20.0. The missing values were handled by using MICE (Multivariate Imputation via Chained Equations) [6] which was commonly used package by R users. The feature selection technique used by the researcher was Boruta algorithm which was based on the random forest algorithm. Random forest algorithm was used for importance feature ranking. In this study, the supervised machine learning algorithms; random forest, c5.0 decision tree, support vector machine, naïve Bayes and the neural network were applied to build a prediction. From the whole predictive models, best predictive models were selected by model evaluation techniques. The performance of the classifiers [25] were measured in terms of different standard metrics like confusion matrix, accuracy, and Kappa, 10-fold validation.

CHAPTER 3: DATA AND METHODOLOGIES

Here, in this study the overall research design was to build a model that identify and predicts the probability of infant and child mortality which was child alive or die. This study was conducted in Ethiopia based on 2016 EDHS dataset. The survey target groups were women age 15-49 and men age 15-59 in randomly selected households across Ethiopia [1]. RStudio tools are utilized as means to address the research problem.

Supervised machine learning algorithms were used for the prediction of infant and child mortality and analyzed the main factors of infant and child mortality in 2016 EDHS dataset. By generating rule based model using C5.0 decision tree and bivariate logistic regression with adjacent odd ratio the determinants were listed. The determinants class distribution in child was alive or not listed by cross tabulation with dependent variable "B5" which was binary outcome variable. From supervised machine learning algorithms, four algorithms were selected for prediction and the models were trained on selected on the feature from the original dataset. Later the performance of the models were evaluated using K-fold cross-validation, performance measures accuracy and Kappa, and confusion matrix.

3.1 Methods of Data Understanding

3.1.1 Source of Data

The data was obtained from the Ethiopian Demographic and Health Survey (EDHS) 2016, which implemented by the Central Statistical Agency (CSA) at the request of the Ministry of Health (MoH) at the national level [1]. The data can be also downloaded from the DHS database [26]. The EDHS collects nationally-representative data on women of child-bearing age (15-49 years) and their children. The data used for under-five mortality estimation were collected in the birth history section of the Woman's Questionnaire from 10,641 women aged 15-49.

3.1.2 Data Analysis

The data analysis in this study have tow stage. In the first stage to develop first insight into the data, relevance analysis like descriptive data visualization was done using statistical tool (SPSS20) that also converted the data comma delimited (csv) format. The second stage was done

in RStudio for data preprocessing like data cleaning, missing value handling, handling imbalance data, feature selection, data split and building model.

3.1.3 Variables Included in the Study

The 2016 Ethiopia Demographic and Health Survey (EDHS) dataset contains 10,641 and 1209 variables which contain information on children's dataset and throughout Ethiopia with all births to interviewed women in the 5 years preceding the survey. The data contain a dependent variable B5 indicating whether the child was alive at the time of the survey. The "Dead" for mothers who had one or more dead child [1]. The dependent variable used in this study was binary, which was coded as 1 if death has not occurred and coded as 0 if death has occurred. On the other hand based on the literature reviewed and with feature selection method, the major proximate determinants of infant and child mortality for this study listed in Table 3.1. The major factors selected from the original dataset that contains 10,641 and 1209 variables were 34 variables.

Table 3.1. Variables description

	Variable code	Variable Name			Variable Name
1	V024	Region	18	V704	Husband/partner's occupation
2	V025	Type of place of residence	19	V716	Respondent's occupation
3	V106	Highest educational level	20	BORD	Birth order number
4	V113	Source of drinking water	21	B0	Child is twin
5	V116	Type of toilet facility	22	B4	Sex of child
6	V137	Number of children in under 5 in household	23	B5	Child is alive
7	V158	Frequency of listening to radio/television	24	B9	The child lives with whom

8	V169A	Owns a mobile telephone	25	B11	Preceding birth interval (months)
9	V190	Wealth index combined	26	B15	Live birth between births
10	V201	Total children ever born	27	B19	The current age of the child in months
11	V208	Births in the last five years	28	M4	Duration of breastfeeding
12	V212	Age of respondent at 1st birth	29	M15	Place of delivery
13	V312	Current contraceptive method	30	M18	Size of the child at birth
14	V404	Breastfeeding	31	H11	Had diarrhea recently
15	V463A	Smokes cigarettes	32	H31	Had to cough in last two weeks
16	V501	Current marital status	33	S1107A	Have you ever chewed Chat?
17	V701	Husband/partner's education level	34	S1107C	Have you ever taken a drink that contains alcohol

3.2 Methods of Data Pre-processing

Data preprocessing is one of the most important phases of the data analysis activity which involves the construction of the final data set which the data was fed into the modeling tool from the preliminary raw data. The Real-world knowledge incline to be extremely disposed to noisy, missing, and inconsistent due to their typically huge size and their likely origin from multiple, heterogeneous sources. Therefore Preparing data is required to get the best results what done within machine learning algorithms on data [27]. To clean the disposed or the original dataset the dataset must be passing through the following data pre-processing methods and techniques.

3.2.1 Data Cleaning

During collecting or entering data, transforming or extracting data, exploring or analyzing data and submitting the draft report for peer review the data may have some errors, outliers and some missing. Therefore; data cleaning was needed for fixing an error occurred [28]. In our data we found 22% of features were completed from 1209 features and the rest of the features contain missing values. The variables decoded to easily manageable which the variables that have many label's options. The variables that contains high missing values were deleted.

3.2.2 Missing Value Handling

Missing values refer to the values for one or more attributes in a data was absent or jumped during data collection. This presence of missing values can affect the performance of a classifier constructed using that dataset as a training sample. Scientifically rates of less than 1% missing data are generally considered minor and 1-5% manageable. However, 5-15% needs to refined strategies and more than 15% may cruelly impact any kind of interpretation. [29].

Here in this study, the researcher tries to discuss different methods and techniques to handle missing data and the best methods were selected. The following methods were some of them in case of handling missing values:

Ignore the tuple: When the class label is missing this method is advisable and the method is not very effective unless the tuple contains several attributes with missing values. It is especially poor when the percentage of missing values per attribute varies considerably [27].

Fill in the missing value manually: This approach was time-consuming and may not be feasible given a large data set with many missing values [27].

Replace with a summary: The most commonly used imputation technique. Summarization here is the mean, mode, or median for a respective column [30].

Use a universal constant to fill in the missing value: Replace all missing attribute values by the same constant such as a label "Unknown [27].

Random replace: replace the missing values with a randomly picked value from the respective column. This technique would be appropriate where the missing values row count is insignificant [30].

Imputation: This technique preserved all cases by replacing missing data with a probable value based on other available information. A simple procedure for imputation was to replace the missing value with the mean or median.

For this study, the researcher chooses imputation approaches that have been used the mice package [31] in RStudio for imputation (Multivariate Imputation by Chained Equations) for handling the missing values. The reason why the researcher selects this methods was which were a mice package gives functions that can impute continuous, binary, and ordered/unordered categorical data, and imputing each incomplete variable with a separate model that, means in one different values were filled in feature column missed space rather than fill the same values in all missed space. Creating multiple imputations as compared to a single imputation (such as mean) takes care of uncertainty in missing values. Therefore the MICE imputed data on a variable by variable based on specifying an imputation model per variable. [6].

3.2.3 Feature Selection Methods

Feature selection [32] [33] was a technique for identifying a subset of features by removing irrelevant or redundant features. The importance of feature selection was reducing the cost of learning by reducing the number of features. Several feature selection methods [7] have been introduced in the machine learning domain like Information Gain and Chi-squared, Stepwise Logistic Regression and R Package "Boruta" were some of them. The main aim of these techniques is to remove irrelevant or redundant features from the dataset and select the important features to feed to machine learning predictive modeling techniques.

The researcher used Boruta algorithms in this study for feature selection. The reason was Boruta algorithm infers features' relevance using the estimate of their importance from Random Forest and all feature is selected. The other reason why we select Boruta is an all-relevant feature selection algorithm that aims at finding both strongly and weakly relevant features from the dataset [34].

Boruta algorithm was one of a feature selection algorithm that was based on Random Forest which removes variables that do not perform well recursively at each iteration. The algorithms started by creating an initial ranking of features in the random forest. Then the algorithm performed an iterative procedure during which the smallest amount important features were sequentially turned into contrast features by permuting their values between objects. Then the threshold level was increased by a predefined step and procedure was repeated until the self-consistence was achieved. The mechanism of its work was first, it copies the dataset, and shuffle the values in each column. These values are called shadow features. Then, it trains a classifier, such as a Random Forest Classifier, on the dataset. By doing this, it ensures that an idea of the importance -via the Mean Decrease Accuracy or Mean Decrease Impurity- for each of the features of data set. The higher the score, the better or more important and the algorithm assigns variables to three classes: (Confirmed, Tentative, and Rejected). [35].

3.2.4 Variables Importance Rank

In building the prediction model, the aim was not only to make the most accurate predictions of the response but also to identify which predictor variables were the most important to make these predictions. In this study, the variables ranked the result get from Boruta feature selection algorithm and it ranked by random forest algorithm [36].

3.2.5 Data Split

Data splitting includes dividing the data into an explicit training dataset used to prepare the model and an unseen test dataset used to evaluate the model's performance on unseen data [37]. The data was split to train data set and test data set in RStudio by using caret package [38] which has a function createDataPartition that conducts data splits within groups of the data. The dataset was randomly categorized into two groups, training, and test groups for model validation as recommended [39].

3.2.6 Imbalanced Data Handling

Data imbalanced [40] is the data proportion of one class out numbers and the other class has a large proportion. This type of proportion more frequently in binary classification problems

than multi-level classification problems. From the term imbalanced we understand that the disproportion encountered in the dependent (response) variable.

The EDHS 2016 dataset the proportion of class for variable "B5" "Alive" and "Dead" group from 10641 records "Dead" group has 6% and "Alive" group is 94%. From here we understand that the dependent variable has imbalance proportion of "Alive" and "Dead" groups. Therefore, we must handle this data to get a good predictive model result. In this study, the researcher looks at different sampling methods and select the best one depending on its best performance of the area under curve (AUC) [37] on the trained model. In this study, we used Random over Sampling Examples (ROSE) and DMwR packages for quickly perform sampling strategies [41].

3.2.6.1 Undersampling the majority class

One of the most common and simplest strategies to handle imbalanced data is to undersample the majority class which majority class is under-sampled by randomly removing samples from the majority class population till the minority class becomes some specified percentage of the majority class. This forces the learner to experience varying degrees of under-sampling and at higher degrees of under-sampling, the minority class has a larger presence in the training set [40], [41]. The techniques of undersampling methods were to stochastically delete some majority class instances to adjust the data distribution.

3.2.6.2 Oversampling the minority class

Oversampling [40] increases the number of minority class instances to balance the distribution of classes which is the reverse of undersampling. The technique here oversampling simply duplicate minority instances and it adds no new information to the dataset and could cause overfitting of classifiers. Although the training accuracy of such data set is high, the accuracy on unseen data is worse [40].

3.2.6.3 Synthetic Minority Over-Sampling Technique (SMOTE)

Instead of replicating and adding observations from the minority class, it overcomes imbalances by generates artificial data. In regards to synthetic data generation, SMOTE

(Synthetic Minority Oversampling Technique) [40], [41] consists of synthesizing elements for the minority class, based on those that already exist

3.2.6.4 Under sampling and oversampling

This technique uses both under and oversampling on the imbalanced data. It used method = "both" which, is the minority class is oversampled with replacement and majority class is under-sampled without replacement. The data generated from oversampling have an expected amount of repeated observations and data generated from undersampling is run-down of necessary data from the first data. This leads to inaccuracies in the resulting performance [40], [41].

3.2.6.5 Random Over Sampling Example (ROSE)

Random over Sampling Example (ROSE) helps to generate data synthetically and the data generated using ROSE is considered to provide a better estimate of original data. For sampling the imbalance dataset in R tools there is ROSE package provides functions to deal with binary classification problems in the presence of imbalanced classes [42].

3.3 Methods of Building a Predictive Modelling

Predictive modeling is used to create a statistical model of unseen upcoming behavior depends on the trained dataset. The predictive modeling in trading is a modeling process wherein predict the probability of an outcome using a set of predictor variables [43]. If the dependent variable was a binary response (yes/no) [44] there are a number of different methods algorithms were considered for use. These are; Logistic Regression, Random forest, Survival Analysis, Neural Network, Naïve Bayes, Decision Tree, Support Vector Machine, K-Nearest Neighbors. The EDHS 2016 child record data set have a binary response and the above-listed algorithm is compatible with this study. However, for this particular research problem, the researcher used classification algorithms such as decision tree (C5.0), random forest, Naïve Bayes classifier and support vector machine for make prediction model and analyzed the data. The predictive algorithms were executed on the top twenty features that were selected from 34 features that were the feature selected by Boruta feature selection method. The classifier was practice on a balanced and unbalanced dataset for each of the classifier algorithms.

In this study to train and evaluate the model created we use a package called caret [38] which stands for classification and regression training that available in R tools. The prediction model was built relying on the training group which has 7450 cases (70% of the whole dataset) and top twenty (20) features that selected and ranked by feature selection methods. In the end, a good model was selected by algorithms evaluation metrics.

3.3.1 Random Forest

Random forest algorithm [31] is a supervised classification and regression algorithm which, randomly creates a forest with several trees. The algorithm was an ensemble learning approach based on classification trees that can solve both types of problems. In simple words, Random forest builds multiple decision trees (called the forest) and glues them together to get a more accurate and stable prediction.

3.3.2 Decision trees

Decision tree [45] learners enable classification via tree structures modeling the relationships among all features and potential outcomes in the data. All decision trees begin with a trunk representing all points are part of the same cohort. Iteratively, the trunk is split into narrower and narrower branches by forking-decisions based on the intrinsic data structure. At every step, splitting the data into branches may include binary or multinomial classification. The final decision is obtained when the tree branching process terminates. The terminal (leaf) nodes represent the action to be taken as the result of the series of branching decisions. For predictive models, the leaf nodes provide the expected forecasting results given the series of events in the decision tree.

In decision tree algorithms rule assimilation easily by end users, classification steps of decision tree induction are simple and fast, and also tree construction does not require any domain knowledge [46]. A “divide-and-conquer” approach to the problem of learning from a set of independent instances leads naturally to a style of representation called a decision tree [47]. Decision trees can easily handle a mix of numeric and categorical attributes and can even classify data for which attributes are missing [48].

There are many implementations of decision trees, but one of the most well-known is the C5.0 algorithm. This algorithm was developed by computer scientist J. Ross Quinlan as an

improved version of his prior algorithm, C4.5, which itself is an improvement over his ID3 (Iterative Dichotomiser 3) algorithm [49].

In R, decision tree algorithm can be implemented using rpart and C50 package. In addition, the caret package was used for doing cross-validation [50]. Size of a decision tree often changes the accuracy of the prediction. In general, bigger tree means higher accuracy, but if the tree is too big, it overfits the data and results in decreased accuracy and robustness. Pruning is a technique used in determining the size of the tree. Often, you would grow a very large tree and starts to trim in down while measuring the change in accuracy and robustness [49].

Important Terminology related to Decision Trees

Root Node: It represents the entire population or sample and this further gets divided into two or more homogeneous sets.

Splitting: it's a method of dividing a node into two or a lot of sub-nodes.

Decision Node: A sub-node splits into further sub-nodes, then it was called a decision node.

Leaf/ Terminal Node: Nodes don't split is called Leaf or Terminal node.

Pruning: It is the process of taking away a sub-nodes of a decision node is called pruning. It is the opposite process of splitting.

Branch / Sub-Tree: A sub section of the entire tree is called branch or sub-tree.

Parent and Child Node: A node, which is divided into sub-nodessubsection is called a parent node of sub-nodes whereas sub-nodes are the child of parent node [30].

A rule-based classification technique used a collection of if ... then ... rules for a rule-based classifier that extract a set of rules show the relationships between the attributes of a dataset and the class label. The Accuracy of a rule is the fragment of instances that satisfy both the previous and consistent rule, normalized by those satisfying the antecedent [49]. In this study, the researcher was used classification rules using C5.0 decision tree by using C5.0 library install in R language in RStudio platform.

3.3.3 Naive Bayes

Bayesian classifier [51] is one of the classifier methods use training data to calculate an observed probability of each outcome based on the provided feature values. Bayesian classifier [52] is a statistical classifier and a practical learning algorithm that can predict class membership probabilities.

Naive Bayes classifier is one of supervised machine learning algorithm that used to build a Naive Bayesian model which make and particularly useful prediction for very large data sets. Apart from being simple, Naive Bayes is known to outperform even highly advanced classification methods. An important feature of naive Bayes classifier is that it only requires a small amount of training data to estimate the parameters necessary for classification. [44].

Bayes theorem provides a way of computing posterior probability P(c|x) from P(c), P(x) and P(x|c). Observe the equation provided here:

P(c/x) = P(x/c)P(c)/P(x)

Where,

P(c|x) is the posterior probability of class (target) given predictor (attribute).

P(c) is the prior probability of class.

P(x|c) is the likelihood which is the probability of predictor given class.

P(x) is the prior probability of predictor [44].

The Naive Bayes algorithm [29] is named as such because it makes some "naive" assumptions about the data. In particular, Naive Bayes assumes that all of the features in the dataset are equally important and independent. The Naive Bayes algorithm describes a simple method to apply Bayes' theorem to classification problems. The Naive Bayes algorithm was: Simple, fast, and very effective and well with noisy and missing data. It requires relatively few examples for training, but also works well with very large numbers of examples and easy to obtain the estimated probability for a prediction.

3.3.4 Support Vector Machine

One of the supervised machine algorithms that we used in this study is Support vector machines (SVM) [53] that exist in different forms, linear and non-linear that separate different categories of data. SVM algorithms have been implemented in R language within five packages namely, e1071, kernlab, klaR, svmpath , shogun. The model in this study built as part of the experiment which consists of kernel used in training and prediction. SVM has different kernel type for a model built and here in this study linear, polynomial, radial basis and sigmoid kernel types used and the best one is selected.

3.4 Methods of Performance Evaluation for Predictive Models

In this section, we evaluate the performance of the predictive models that were used in building predict model by using a number of standard evaluation metrics like K-fold cross-validation, performance measures accuracy and Kappa and confusion matrix. By using caret there are different popular evaluation metrics with RStudio [37]. Those are:

3.4.1 Accuracy and Kappa

These metrics are the default metrics used to evaluate algorithms on binary and multiclass classification datasets in caret. Accuracy is the percentage of correctly classified instances out of all instances. Kappa or Cohen's Kappa is like classification accuracy, except that it is normalized at the baseline of random chance on your dataset. It is a more useful measure to use on problems that have an imbalance in the classes [37].

3.4.2 K-fold cross-validation

K-fold cross-validation [37] is a method used to evaluate the models by splitting the dataset into k-subsets without replacement. This process is repeated K time and we obtain K models and determined for each instance in the dataset, and an overall accuracy estimate is provided. In this study, we have a tendency to used 10-fold cross-validation evaluation method where ever the data are randomly partitioned into 10 mutually exclusive subsets with approximately equal size. The testing operation is then repeated 10 times where at the ith evaluation iteration. The dependent variable used in this study was binary which experiment comes under dichotomous classification, this means that each fold contains roughly the same proportions of the two types of class labels.

The performance of the models [54] were compared based on the accuracy obtained in the prediction of unseen test data. Therefore each iteration of the evaluation process, the following metrics are calculated:

Sensitivity: is the true positive rate also called the recall. It is the number of instances from the positive (first) class that actually predicted correctly.

Sensitivity = TP/TP + FN

Specificity is also called the true negative rate. It is the number of instances from the negative class (second class) that were actually predicted correctly.

Specificity = TN/TN + FP

Precision: It represents the percentage of tuples that the classifier has labeled as positive are actually positive

Precision = TP/TP + FP

3.4.3 Confusion matrix

A confusion matrix [55] is a tool used to analyze the performance of a classification model is recognized on a set of test data for which the true values were known in table format. The confusion matrix provides a quick understanding of model accuracy and the types of errors in the building of models. Performance of such systems is commonly evaluated using the data in the matrix known as confusion matrix which is a binary classification of two by two table counting of the number of the four outcomes of a binary classifier.

Table 3.2. Confusion Matrix

N=Number of instances	Predicted: YES	Predicted: NO
Actual: YES	TP	FN
Actual: NO	FP	TN

In the above Table 3.2 there are two possible predicted classes: "yes" and "no". If we were predicting the presence of an instance in that category, for example, "yes" would mean the presence of the instance of that category, and "no" would mean the absence of the instance of that category. A binary classifier predicts all data instances of a test dataset as either positive or negative. Possibly the classification normally produces only four outcomes as true positive, true negative, false positive and false negative [55].

CHAPTER 4: IMPLEMENTATION AND RESULTS

This chapter presents the results of what we discuss and design in chapter three and the results were obtained by applying the four models and the performance of the models. In the section descriptive statistical summary and machine learning techniques used by the researcher presented.

4.1 Descriptive statistical summary

In this study there are a total of 10,641 children were included in the study and out of 10,641 children, 5987 (51.4%) were male. The majority of the children, 9668 (83.0%), were from rural areas of the country. The wealth index difference among the family who participates in the survey were 5775(54.3%) poor, 1466(13.8%) middle, and 3400(32.0%) richer. 7888 (74.1%) mothers not using any conservative method and 7271(68.3%) mothers were delivered in the home. 7618 (71.6%) mothers ages were less than twenty years at the age of the first childbearing. The mortality percentage of the infant was 50.9% and the child mortality was 49.1 %. (Table 4.1)

Table 4.1. Determinants class distribution

Determinants * Child is alive Cross tabulation					
			Child is alive		
			No	Yes	Total
Region	Tigray	Count	41	992	1033
		% of Total	.4%	9.3%	9.7%
	Afar	Count	90	972	1062
		% of Total	.8%	9.1%	10.0%
	Amhara	Count	49	928	977
		% of Total	.5%	8.7%	9.2%
	Oromia	Count	87	1494	1581
		% of Total	.8%	14.0%	14.9%
	Somali	Count	103	1402	1505
		% of Total	1.0%	13.2%	14.1%
	Benishangul	Count	64	815	879
		% of Total	.6%	7.7%	8.3%

	SNNPR	Count	71	1206	1277
		% of Total	.7%	11.3%	12.0%
	Gambela	Count	44	670	714
		% of Total	.4%	6.3%	6.7%
	Harari	Count	41	564	605
		% of Total	.4%	5.3%	5.7%
	Addis Adaba	Count	14	447	461
		% of Total	.1%	4.2%	4.3%
	Dire Dawa	Count	31	516	547
		% of Total	.3%	4.8%	5.1%
Total		Count	635	10006	10641
		% of Total	6.0%	94.0%	100.0%
Type of place of residence	Urban	Count	67	1907	1974
		% of Total	.6%	17.9%	18.6%
	Rural	Count	568	8099	8667
		% of Total	5.3%	76.1%	81.4%
Total		Count	635	10006	10641
		% of Total	6.0%	94.0%	100.0%
		% of Total	6.0%	94.0%	100.0%
Source of drinking water	PROTECTED WATER	Count	338	6011	6349
		% of Total	3.2%	56.5%	59.7%
	UNPROTECTED WATER	Count	297	3995	4292
		% of Total	2.8%	37.5%	40.3%
Total		Count	635	10006	10641
		% of Total	6.0%	94.0%	100.0%
Type of toilet facility	IMPROVED TOILET	Count	29	586	615
		% of Total	.3%	5.5%	5.8%
	NON IMPROVEDTOILET	Count	606	9420	10026
		% of Total	5.7%	88.5%	94.2%
Total		Count	635	10006	10641
		% of Total	6.0%	94.0%	100.0%
Number of children aged under 5 in the household not include neighboring	0	Count	156	194	350
		% of Total	1.5%	1.8%	3.3%
	1	Count	276	3483	3759
		% of Total	2.6%	32.7%	35.3%
	2	Count	143	4375	4518
		% of Total	1.3%	41.1%	42.5%
	3	Count	45	1677	1722
		% of Total	.4%	15.8%	16.2%
	4	Count	13	222	235

		% of Total	.1%	2.1%	2.2%
	5	Count	0	37	37
		% of Total	0.0%	.3%	.3%
	6	Count	2	18	20
		% of Total	.0%	.2%	.2%
Total		Count	635	10006	10641
		% of Total	6.0%	94.0%	100.0%
Wealth index combined	Poor	Count	399	5376	5775
		% of Total	3.7%	50.5%	54.3%
	Middle	Count	80	1386	1466
		% of Total	.8%	13.0%	13.8%
	Richer	Count	156	3244	3400
		% of Total	1.5%	30.5%	32.0%
Total		Count	635	10006	10641
		% of Total	6.0%	94.0%	100.0%
Total children ever born	<2	Count	61	1409	1470
		% of Total	.6%	13.2%	13.8%
	2-5	Count	345	5700	6045
		% of Total	3.2%	53.6%	56.8%
	6-9	Count	194	2567	2761
		% of Total	1.8%	24.1%	25.9%
	>9	Count	35	330	365
		% of Total	.3%	3.1%	3.4%
Total		Count	635	10006	10641
		% of Total	6.0%	94.0%	100.0%
Births in the last five years	1	Count	157	4121	4278
		% of Total	1.5%	38.7%	40.2%
	2	Count	276	4576	4852
		% of Total	2.6%	43.0%	45.6%
	3	Count	166	1175	1341
		% of Total	1.6%	11.0%	12.6%
	4	Count	32	128	160
		% of Total	.3%	1.2%	1.5%
	5	Count	4	6	10
		% of Total	.0%	.1%	.1%
Total		Count	635	10006	10641
		% of Total	6.0%	94.0%	100.0%
Age of respondent at 1st birth	<20	Count	470	7148	7618
		% of Total	4.4%	67.2%	71.6%
	20-29	Count	154	2692	2846
		% of Total	1.4%	25.3%	26.7%

	30-39	Count	11	161	172
		% of Total	.1%	1.5%	1.6%
	40-49	Count	0	1	1
		% of Total	0.0%	.0%	.0%
Total		Count	635	10006	10641
		% of Total	6.0%	94.0%	100.0%
Current contraceptive method	Not using	Count	515	7373	7888
		% of Total	4.8%	69.3%	74.1%
	Using	Count	120	2633	2753
		% of Total	1.1%	24.7%	25.9%
Total		Count	635	10006	10641
		% of Total	6.0%	94.0%	100.0%
Breastfeeding	No	Count	393	3428	3821
		% of Total	3.7%	32.2%	35.9%
	Yes	Count	242	6578	6820
		% of Total	2.3%	61.8%	64.1%
Total		Count	635	10006	10641
		% of Total	6.0%	94.0%	100.0%
Husband/partner's education level	No education	Count	324	4604	4928
		% of Total	3.2%	46.0%	49.2%
	Primary	Count	186	3034	3220
		% of Total	1.9%	30.3%	32.2%
	Secondary	Count	43	972	1015
		% of Total	.4%	9.7%	10.1%
	Higher	Count	36	734	770
		% of Total	.4%	7.3%	7.7%
	Don't know	Count	2	73	75
		% of Total	.0%	.7%	.7%
Total		Count	591	9417	10008
		% of Total	5.9%	94.1%	100.0%
Husband/partner's occupation	Not working and didn't work	Count	75	1103	1178
		% of Total	.7%	11.0%	11.8%
	Working	Count	516	8314	8830
		% of Total	5.2%	83.1%	88.2%
Total		Count	591	9417	10008
		% of Total	5.9%	94.1%	100.0%
Birth order number	1	Count	136	2031	2167
		% of Total	1.3%	19.1%	20.4%
	2-3	Count	176	3162	3338
		% of Total	1.7%	29.7%	31.4%

	4-6	Count	203	3182	3385
		% of Total	1.9%	29.9%	31.8%
	7+	Count	120	1631	1751
		% of Total	1.1%	15.3%	16.5%
Total		Count	635	10006	10641
		% of Total	6.0%	94.0%	100.0%
Child is twin	Single birth	Count	577	9786	10363
		% of Total	5.4%	92.0%	97.4%
	1st of multiple	Count	23	116	139
		% of Total	.2%	1.1%	1.3%
	2nd of multiple	Count	35	104	139
		% of Total	.3%	1.0%	1.3%
Total		Count	635	10006	10641
		% of Total	6.0%	94.0%	100.0%
Sex of child	Male	Count	376	5107	5483
		% of Total	3.5%	48.0%	51.5%
	Female	Count	259	4899	5158
		% of Total	2.4%	46.0%	48.5%
Total		Count	635	10006	10641
		% of Total	6.0%	94.0%	100.0%
Preceding birth interval	<2 year	Count	204	1910	2114
		% of Total	2.4%	22.6%	25.0%
	2-5 year	Count	239	5001	5240
		% of Total	2.8%	59.1%	61.9%
	>5 year	Count	51	1048	1099
		% of Total	.6%	12.4%	13.0%
Total		Count	494	7966	8460
		% of Total	5.8%	94.2%	100.0%
Live birth between births	No	Count	492	7906	8398
		% of Total	5.8%	93.3%	99.1%
	Yes	Count	7	69	76
		% of Total	.1%	.8%	.9%
Total		Count	499	7975	8474
		% of Total	5.9%	94.1%	100.0%
Place of delivery	HOME	Count	490	6781	7271
		% of Total	4.6%	63.7%	68.3%
	HEALTH CARE	Count	145	3225	3370
		% of Total	1.4%	30.3%	31.7%
Total		Count	635	10006	10641
		% of Total	6.0%	94.0%	100.0%
	Very large	Count	193	3021	3214

Size of the child at birth		% of Total	1.8%	28.4%	30.2%
	Average	Count	223	4196	4419
		% of Total	2.1%	39.4%	41.5%
	Very small	Count	219	2789	3008
		% of Total	2.1%	26.2%	28.3%
Total		Count	635	10006	10641
		% of Total	6.0%	94.0%	100.0%
Age at death	Infant Death	Count	323		323
		% of Total	50.9%	0%	50.9%
	Child Death	Count	312		312
		% of Total	49.1%	0%	49.1%
Total		Count	635		635
		% of Total	100.0%	0%	100.0%

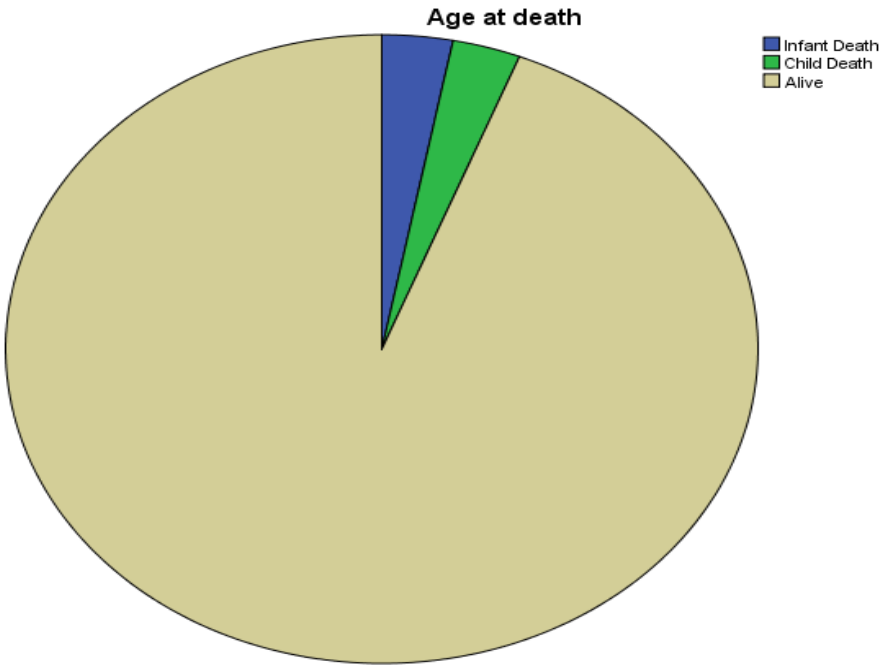

Figure 4.1. Infant and child death distribution

In the study in feature selection process from 1902 original features 34 variables are selected and the rest of were ignored based on the literature reviewed, by ignoring the feature that have high missing values and by using Boruta and random forest algorithms for selecting the top 20 (twenty) variables in feature importance related to variable B5 which is child is "Died" or "Alive". The top 20 (twenty) variables were feed to the models. By using Bivariate

logistic regression with an adjacent odd ratio the determinant correlation with target B5 class variable. The p_value < 0.001 has the most correlated with the target class variable B5 (

Table **4.2**). The table indicate predictor with the confidence interval range in adjacent odd ratio for child was alive or not.

Table 4.2. Bivariate logistic regression with AOR to distribution of Child alive or not

Logistic regression predicting B5: 1(alive) vs 0(died)

	Crude OR(95%CI)	adj. OR(95%CI)	P(Wald's test)	P(LR-test)
V024	1.0051 (0.9768,1.0342)	0.9913 (0.9535,1.0305)	0.657	0.657
V025	0.5 (0.39,0.65)	0.73 (0.5,1.08)	0.118	0.115
V113	0.98 (0.96,1.01)	0.97 (0.94,0.99)	0.017	0.019
V137	2.9 (2.58,3.26)	16.08 (13.3,19.44)	< 0.001	< 0.001
V190	1.16 (1.1,1.22)	0.97 (0.89,1.06)	0.487	0.488
V201	0.92 (0.89,0.95)	0.79 (0.72,0.87)	< 0.001	< 0.001
V208	0.52 (0.47,0.57)	0.1 (0.09,0.13)	< 0.001	< 0.001
V312	1.04 (1.01,1.08)	1.03 (0.99,1.07)	0.179	0.171
V404	3.12 (2.64,3.68)	1.23 (0.94,1.62)	0.137	0.138
V501	0.92 (0.83,1.02)	0.95 (0.83,1.08)	0.415	0.42
V701	1.16 (1.06,1.26)	1.08 (0.96,1.21)	0.196	0.188
V716	1.0016 (0.999,1.0041)	0.9994 (0.9962,1.0026)	0.724	0.724
BORD	0.96 (0.88,1.04)	1.73 (1.35,2.22)	< 0.001	< 0.001
B0	0.4 (0.33,0.48)	0.54 (0.4,0.72)	< 0.001	< 0.001
B9	1.7 (1.2,2.41)	4.72 (3.25,6.84)	< 0.001	< 0.001
B11	1.0132 (1.0087,1.0177)	1.0075 (1.0017,1.0133)	0.012	0.01
B19	0.99 (0.98,0.99)	1.01 (1.01,1.02)	< 0.001	< 0.001
M4	2.24 (1.99,2.51)	2.58 (2.12,3.14)	< 0.001	< 0.001
M15	1.0176 (1.0062,1.0292)	1.0062 (0.9939,1.0187)	0.324	0.308
M18	0.92 (0.87,0.98)	0.96 (0.89,1.02)	0.204	0.205

Log-likelihood = -1461.7195

No. of observations = 10641

AIC value = 2965.4391

4.2 Data Preprocessing

4.2.1 Missing Value Handling

The 2016 Ethiopia Demographic and Health Survey (EDHS) dataset contains 10,641 and 1209 variables which contain information on children's dataset. As we discussed in chapter three in section 3.5.2 the column which the missing values greater than 15% were removed and the variables that not have any relation to the outcome variable were removed. Finally, 34 variables were selected from 1209 variables and the result look like in Figure 4.2.

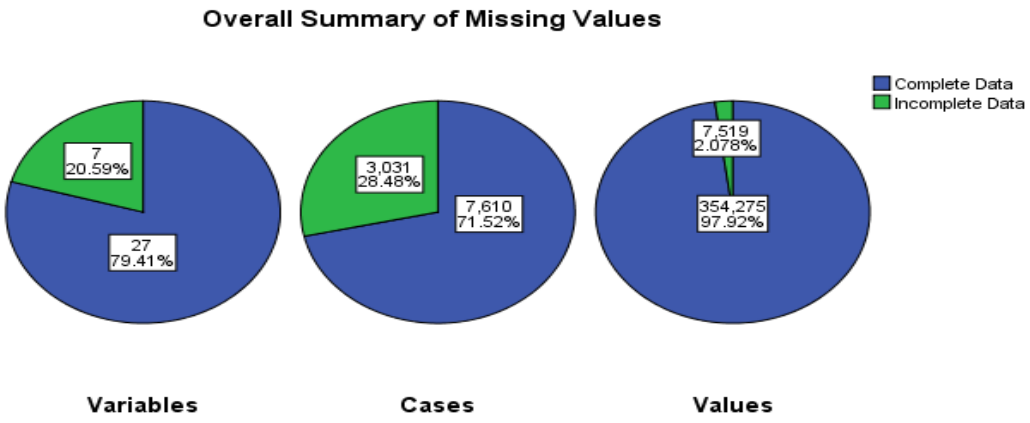

Figure 4.2. Overall of missing values

From the figure, we understand that 7 variables have missing values with 20.59% and in case of values, 7,519 values were missed with 2.078% from the total of values.

4.2.2 Feature Selection Methods

The important features were selected by using Boruta algorithms which respect to an outcome variable "B5" with assigning the variables to three classes: (Confirmed, Tentative, and Rejected) which looks like Figure 4.3.

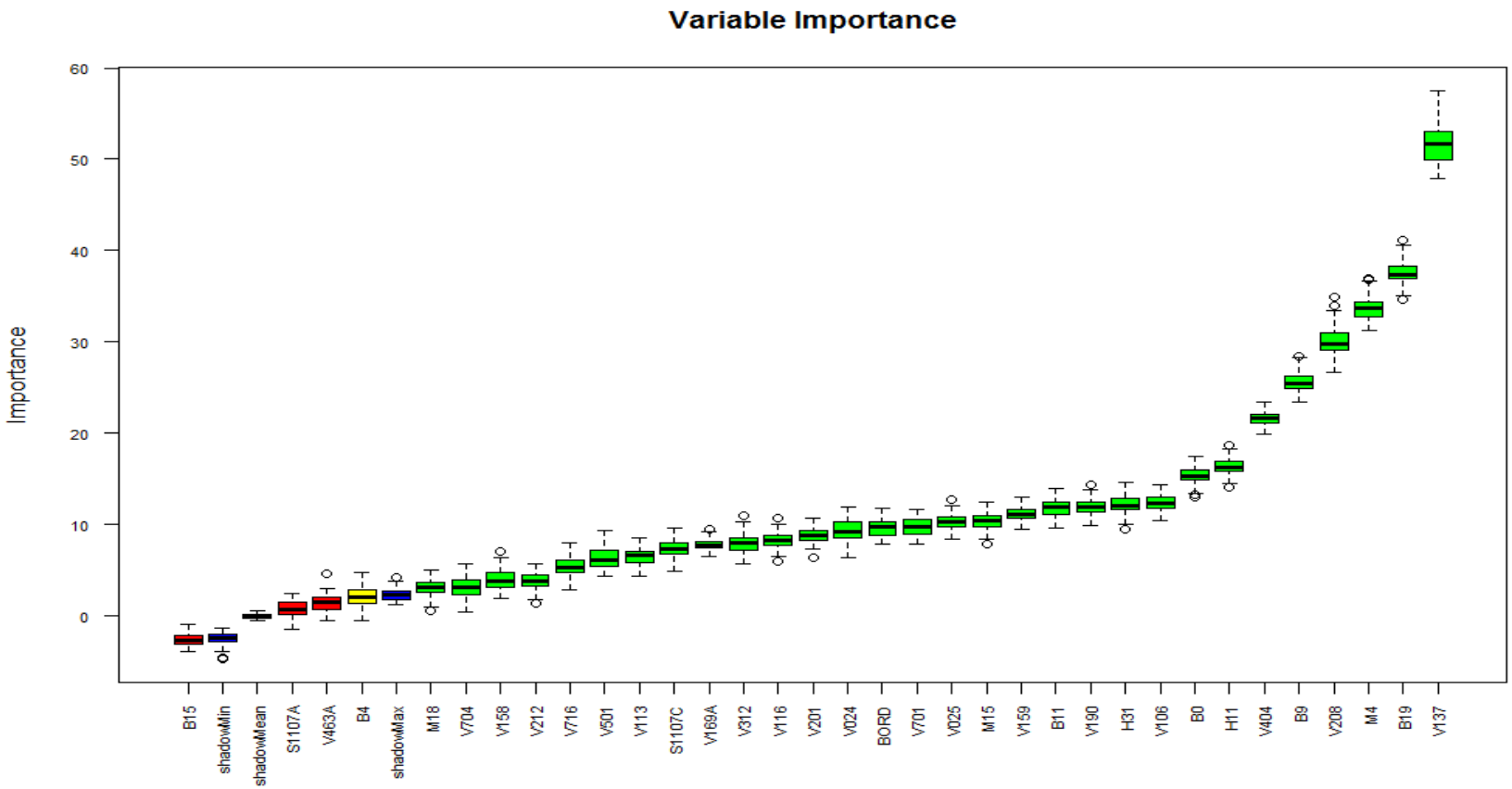

Figure 4.3. Boruta features importance result.

In Figure 4.3 the variables in boxplot sorted by increasing importance and colored in green are those which were classified as relevant. Variables colored in red are those which are irrelevant and yellow colored indicate tentative variable. The blue color boxplot indicates that shadow attributes that created by Boruta algorithm for benchmark or reference in variable importance comparing. The decision to confirm, reject and tentative are depended on mean imputation, median imputation, maximum imputation, and normal hits. The ranked variables by mean imputation look like in **Error! Not a valid bookmark self-reference.**.

Table 4.3. Ranked Boruta features importance result.

Variable	meanImp	Decision		Variable	meanImp	Decision
V137	51.668838	Confirmed		BORD	9.671205	Confirmed
B19	37.560339	Confirmed		V024	9.359591	Confirmed
B4	33.634897	Confirmed		V201	8.876538	Confirmed
V208	30.000387	Confirmed		V116	8.335360	Confirmed
B9	25.670768	Confirmed		V312	7.962722	Confirmed
V404	21.614949	Confirmed		V169A	7.867425	Confirmed
H11	16.336983	Confirmed		S1107C	7.292216	Confirmed
B0	15.406808	Confirmed		V113	6.544518	Confirmed
V106	12.365924	Confirmed		V501	6.391137	Confirmed
H31	12.266567	Confirmed		V716	5.443755	Confirmed
V190	11.919100	Confirmed		V158	3.980583	Confirmed
B11	11.840412	Confirmed		V212	3.827046	Confirmed
V159	11.161854	Confirmed		M18	3.160596	Confirmed
M15	10.370131	Confirmed		V704	3.138065	Confirmed
V025	10.352005	Confirmed		B4	2.171734	Tentative
V701	9.801510	Confirmed				

4.2.3 Features Importance Rank

Using Boruta feature selection algorithm and random forest twenty most important variables were selected from 34 variables which indicated in Figure 4.4.

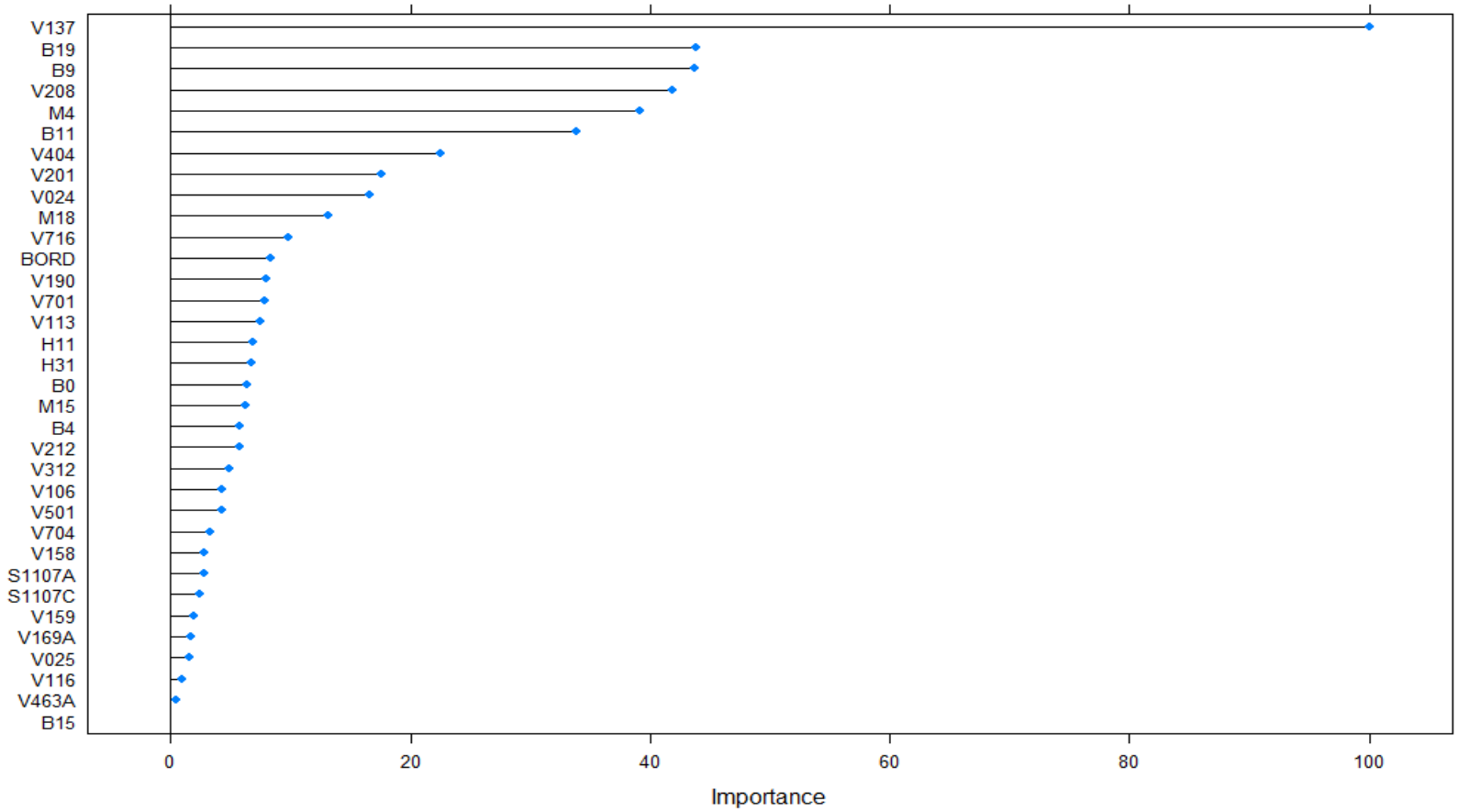

Figure 4.4. Top twenty features ranked by random forest

4.2.4 Data Split

The dataset was randomly categorized into two groups, training, and test groups. The training group comprised of 7450 cases (70% of the whole dataset). The prediction model was built relying on the training group. The remaining 3191 cases (30% of the whole dataset) were allocated as the test group for model validation recommended [39].

4.2.5 Imbalanced Data Handling

To solve the problem of data imbalance we used the sampling method what discuss in chapter three under imbalance data handling. Each sampling technique was experimented here and use the best one. Using five sampling methods the imbalance data was handled as the following.

A number of train data class distribution before sampling from total 7450 observations 445(6%) were under 0 class which was child died and 7005(94%) observations were under 1 class which was child alive.

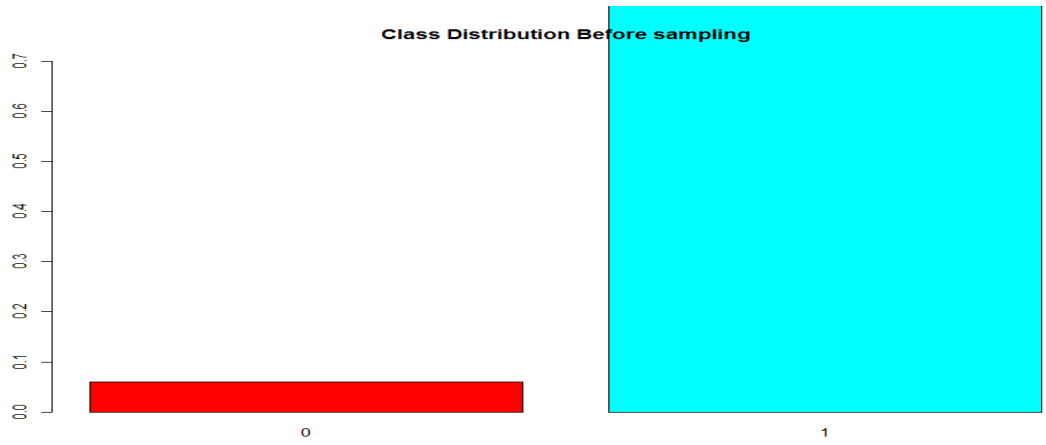

Figure 4.5. Class distribution before applying sampling techniques

The number of train data class distribution after undersampling method from total 7450 observations minimized to 878 observations which randomly removing samples from the

majority class population until the minority class becomes some specified proportion of the majority class. In this sampling technique, 433(49.3%) observations were under 1 class which remove observation and 445(50.7%) observations were under 0 class. A number of train data class distribution after oversampling technique increases the number of minority class instances to balance the distribution of classes. In this case, oversample minority class until it reaches the highest class which increase the minority class from 445 instances to 6934 instances. A number of train data class distribution after both under and oversampling the minority class is oversampled with replacement and majority class is under-sampled without replacement. A number of train data class distribution after ROSE sampling technique create "synthetic" data for classifiers from imbalanced datasets. A number of train data class distribution after SMOTE sampling instead of replicating and adding the observations from the minority class, it overcomes imbalances by generates artificial data. Here the technique doubles the minority class from 445 to 890. Table 4.4 indicates that the dependent variable class distribution after and before sampling technique what have discussed.

Table 4.4. Data imbalance sampling technique result

Sampling Methods	Class 0: child "Died"	Class 1: child "Alive"	Total
Before Sampling	445	7005	7450
	6%	94%	100%
under sampling	445	433	878
	50.7%	49.3%	100%
Over Sampling	6934	7005	13939
	49.7%	50.3%	100%
Both Under and Over Sampling	3704	3746	7450
	49.7%	50.3%	100%
ROSE	3704	3746	7450
	49.7%	50.3%	100%
SMOTE	890	890	1780
	50%	50%	100%

For comparison, the Random Forest classifier model was used and computed. Using each trained balanced data the test data was predicted on the model. The best sampling methods were selected based on the prediction performance of the Random Forest classifier model which was computed on each trained balanced data and the test data based on resamples method with metric Accuracy and Kappa which indicated in Table 4.5.

Table 4.5. Comparing Sampling Method

Accuracy						
	Min	1st Qu.	Median	Mean	3rd Qu.	Max.
ROSE	0.8077	0.8268	0.8375	0.8375	0.8492	0.8669
OVER	0.9286	0.9316	0.9382	0.9377	0.9428	0.9500
UNDER	0.7816	0.8433	0.8628	0.8614	0.8873	0.9101
BOTH	0.9195	0.9264	0.9342	0.9341	0.9409	0.9516
SMOTE	0.9101	0.9347	0.9373	0.9373	0.9421	0.9528
Kappa						
	Min	1st Qu.	Median	Mean	3rd Qu.	Max.
ROSE	0.6155	0.6537	0.6751	0.6751	0.6985	0.7338
OVER	0. 8572	0. 8633	0. 8765	0. 8754	0. 8857	0. 9000
UNDER	0. 5630	0. 6867	0. 7260	0. 7230	0. 7749	0. 8204
BOTH	0. 8388	0. 8529	0. 8684	0. 8682	0. 8818	0. 9032
SMOTE	0.7788	0.8375	0.8426	0.8443	0.8574	0.8833

4.3 Make Predictions on Unseen Test Data

Here, classifier models were built to predict the infant and child mortality using Random forest, C5.0 decision tree, Support Vector Machine, Naïve Bayes algorithms. The models prediction performance on the unseen dataset were measured and compared on the models trained on the balanced and unbalanced train dataset.

4.3.1 Naïve Bayes classifier: Prediction on Test Data

The Naïve Bayes model was trained on a dataset of 70% and make a prediction on 30% of the unseen test dataset. After train the Naïve Bayes model on train data set by caret package we can make a prediction on unseen test data set. As we can see in the confusion matrix, the model correctly predict 63 children as died and 2820 children as alive. It has wrongly predicted 81 children as died and 127 children as alive.

```
          (a)        (b)    <-classified as
       Reference
Prediction   0        1
     0      63       81    (a): class 0 child died
     1     127     2820    (b): class 1 child was alive
```

The accuracy of prediction of the model was 90.35% with 95% CI (confidence interval): which the accuracy was lie in the range of (0.8927, 0.9135). The model indicates that the Sensitivity: 33.15% and the Specificity: 93.99%. The Balanced Accuracy was 63.56%

On the other hand when we train the model on balanced train dataset by SMOTE sampling and we get the result with Confusion Matrix as the following:

```
          (a)        (b)    <-classified as
       Reference
Prediction   0        1
     0      64       80    (a): class 0 child was died
     1     126     2921    (b): class 1 child was alive
```

The accuracy of the prediction of the model was increased from 90.35% to 93.54% with 95% CI (confidence interval) with an accuracy range (0.9264, 0.9437). The Sensitivity increase from 33.15% to 33.68% % and the Specificity increased from 93.99% to 97.33. The Balanced Accuracy was increased from 63.56% to 65.50% (Table 4.6).

Table 4.6. Performance of Naïve Bayes classifier

	Naïve Bayes prediction model	
Metrix	Unbalanced Data	Balanced Data
Accuracy	90.35%	93.54%
95% CI	(0.8927, 0.9135)	(0.9264, 0.9437)
Sensitivity	33.15%	33.68%
Specificity	93.99%	97.33%
Balanced Accuracy	63.56%	65.50%
Positive Predicted Value	25.82%	44.44%
Negative Predicted Value	95.69%	95.87%

4.3.2 C5.0 classifier: on Train Dataset

By using caret package C5.0 classifier was trained on unbalanced and balanced train dataset. After the train, the model prediction was made on the unseen test dataset and we can get the result of Confusion Matrix and Statistics. The confusion Matrix for unbalanced train dataset on a new unseen test dataset looks like the following matrix

```
          (a)        (b)    <-classified as
          Reference
Prediction   0        1          (a): class 0 child was died
         0   108      22
         1   82       2979       (b): class 1 child was alive
```

Table 4.7. Performance of C5.0 classifier

	C5.0 prediction model	
Metrix	Unbalanced Data	Balanced Data
Accuracy	96.74%	96.15%
95% CI	(0.9606, 0.9733)	(0.9542, 0.9679)
Sensitivity	56.84%	65.78%
Specificity	99.26%	98.06%
Balanced Accuracy	78.05%	81.92%
Positive Predicted Value	83.10%	68.31%
Negative Predicted Value	97.32%	97.84%

From the Confusion Matrix understand the model predict correctly 108 children as classified under 0 indicate children were died and 2979 classified under 1 which indicate the child as alive. And it indicate that 82 children as died but not correctly predicted and 22 children as alive but wrongly predicted that was falsely predicted. The accuracy of prediction of the model was 96.74% with 95% CI (confidence interval): which the accuracy lay in range (0.9606, 0.9733). The model indicates that the Sensitivity: 56.84% and the Specificity: 99.26%. The Balanced Accuracy was 78.05% (Table 4.7)

On the other hand when we train the model on balanced train dataset which balanced SMOTE sampling and we get the result with Confusion Matrix as the following:

```
          (a)       (b)       <-classified as
       Reference
Prediction  0        1
       0   125       58       (a): class 0 child was died

       1   65        2943     (b): class 1 child was alive
```

The accuracy of the prediction of the model was a little decreased from 96.74% to 96.15% with 95% CI (confidence interval) with an accuracy range (0.9542, 0.9679). The Sensitivity increase from 56.84% to 65.78% and the Specificity decrease from 99.26% to 98.06%. The Balanced Accuracy was increased from 78.05% to 81.92% (Table 4.7).

4.3.2.1 Rules from decision trees

Using C5.0 Decision tree 8 rules are generated. Those are the following

```
Rules:

Rule 1: (126/12, lift 15.1)
        V137 > 0
        V137 <= 1
        V208 > 1
        V404 > 0
        B9 <= 0
        M4 <= 1
        ->  class 0  [0.898]

Rule 2: (81/16, lift 13.3)
        V137 <= 1
        V208 > 2
        B9 <= 0
        M4 <= 1
        ->  class 0  [0.795]

Rule 3: (24/5, lift 12.9)
        V137 <= 1
        B0 > 1
        M4 <= 1
        ->  class 0  [0.769]

Rule 4: (233/77, lift 11.2)
        V137 <= 0
        B9 <= 0
        ->  class 0  [0.668]

Rule 5: (453/172, lift 10.4)
        V137 <= 1
        V208 > 1
        B9 <= 0
        M4 <= 1
        ->  class 0  [0.620]

Rule 6: (4245, lift 1.1)
        M4 > 1
        ->  class 1  [1.000]

Rule 7: (117, lift 1.1)
        V137 <= 0
        B9 > 0
        ->  class 1  [0.992]

Rule 8: (10291/479, lift 1.0)
        V137 > 0
        ->  class 1  [0.953]
```

Rule 1: If Number of children resident in the household and aged 5 and under is one and Bi rths in last five years greater than 0ne and child is lives with respondent and child is not bre astfeed the child died. The probability of child die was 89.9%

Rule 2: If Number of children resident in the household and aged 5 and under is one or is n ot there and Births in last five years greater than two and child is lives with respondent and child is not breastfeed the child died with a probability of 79.5%

Rule 3: If Number of children resident in the household and aged 5 and under is one or not t here and Child is twin is greater than first of multiple and child is not breastfeed the child d ied with a probability of 76.9%

Rule 4: If Number of children resident in the household and aged 5 and under is not there a nd Births in the last five years with no birth the child died with a probability of 66.9%

Rule 5: If Number of children resident in the household and aged 5 and under is one or not t here and Child was lives with respondent and child is not breastfeed the child died with a pr obability of 62.0%

Rule 6: If a child is not breastfed the child died with a probability of 100.0%

Rule 7: If Number of children resident in the household aged 5 and under is not there and th e child was lives with other place child was alive with a probability of 99.2%

Rule 8: If Number of children resident in the household aged 5 and under is more than one th e probability of child was alive was 95.3%. For more detail the tree generated by C5.0 classifier looks like (Figure 4.6))

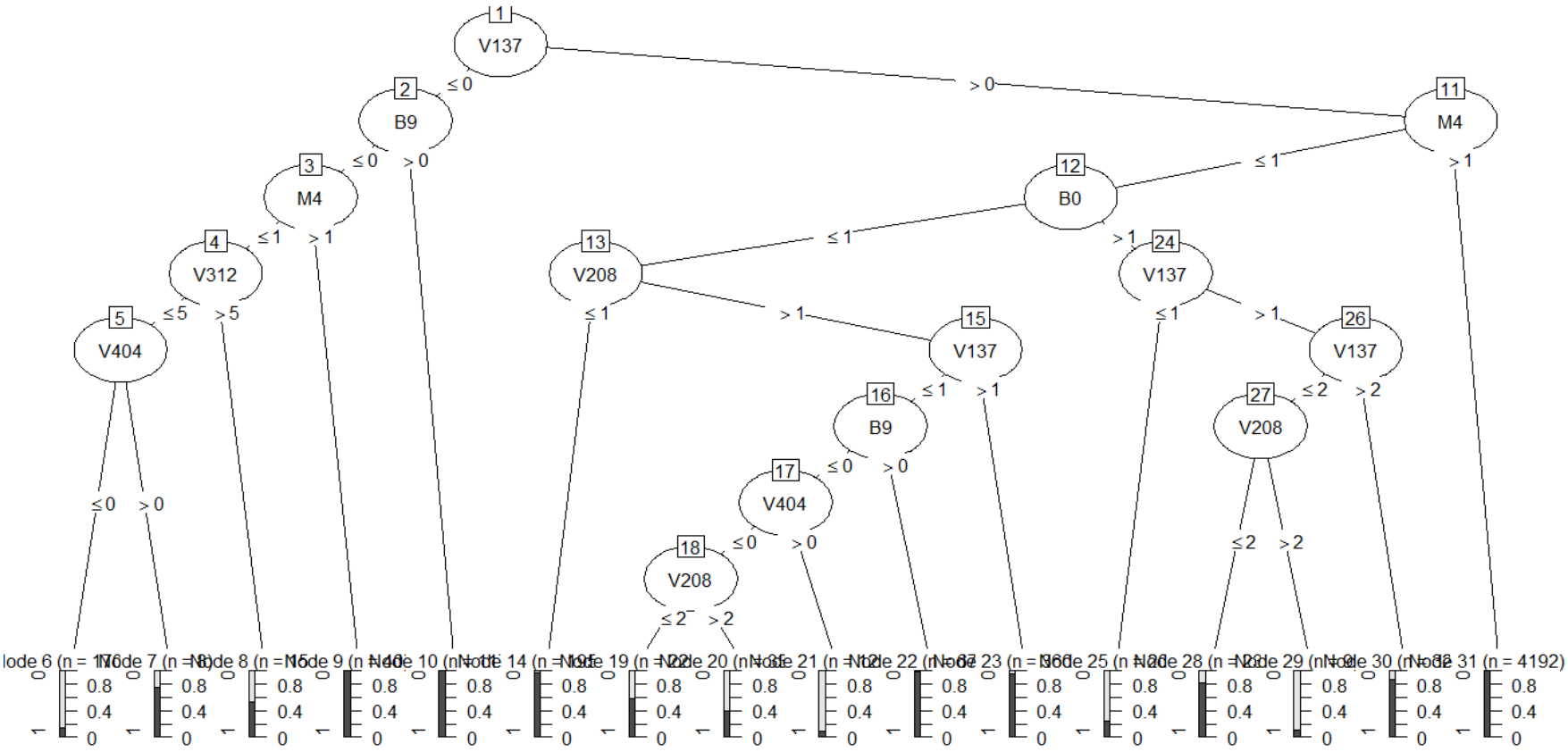

Figure 4.6. C5.0 Rule-Based Decision Tree

4.3.3 SVM classifier: unbalanced and balanced train dataset

The SVM classifier was trained on unbalanced and balanced train dataset within SVM-Kernel: Polynomial, sigmoid, and radial. After the train, the model prediction was made on the unseen test dataset and we can get the result of Confusion Matrix and Statistics for all of three SVM-Kernel: Polynomial, sigmoid, and radial. The prediction and confusion Matrix for unbalanced train dataset on a new unseen test dataset looks like the following matrix(Table 4.8).

Table 4.8. Confusion Matrix of SVM Imbalance data

SIGMOID				POLYNOMIAL				RADIAL		
	0	1			0	1			0	1
0	61	91		0	54	27		0	82	10
1	129	2909		1	136	2974		1	108	2991

As we can see in the confusion matrix, the model correctly predict 61 children as died and 2909 children as alive. It has wrongly predicted 91 children as died and 129 children as alive in sigmoid. In polonomial model correctly predict 54 children as died and 2974 children as alive. It has wrongly predicted 27 children as died and 136 children as alive. On the othe radial model indicated correctly predict 82 children as died and 2991 children as alive. It has wrongly predicted 10 children as died and 108 children as alive(Table 4.8). From the three model Radial SVM was good with 96.30% accuracy performance.

Table 4.9. Performance of SVM on prediction on imbalance data

	Type of SVM with its Kernel		
Metrix	SVM Kernel polynomial	SVM Kernel sigmoid	SVM Kernel radial
Accuracy	94.89%	93.07%	96.30%
95% CI	(0.9407, 0.9563)	(0.9214, 0.9393)	(0.9559, 0.9693)
Sensitivity	28.42%	32..10%	43.15%
Specificity	99.10%	96.93%	99.66%
Balanced Accuracy	63.76%	64.52%	71.14%
Pos Pred Value	66.67%	39.87%	89.13%
Neg Pred Value	95.63%	9575	96.52%

On the other hand, the prediction and confusion Matrix for balanced train dataset on a new unseen test dataset looks like the following matrix.

Table 4.10. Confusion Matrix for SVM balanced data

SIGMOID				POLYNOMIAL				RADIAL		
	0	1			0	1			0	1
0	121	383		0	132	129		0	138	97
1	69	2618		1	58	2872		1	52	2904

Table 4.11. Performance of SVM on prediction on balanced data

Metrix	Type of SVM with its Kernel		
	SVM Kernel polynomial	SVM Kernel sigmoid	SVM Kernel radial
Accuracy	94.14%	85.84%	95.33%
95% CI	(0.9323, 09493)	(0.8458, 0.8703)	(0.9454, 0.9604)
Sensitivity	69.47%	63.68%	72.63%
Specificity	95.70%	87.23%	96.76%
Balanced Accuracy	82.58%	75.46%	84.70%
Pos Pred Value	50.58%	24.01%	58.72%
Neg Pred Value	98.02%	97.43%	98.24%

4.3.4 Random Forest model: on train dataset

Random Forest classifier was trained on unbalanced and balanced train dataset. After the train, the model prediction was made on the unseen test dataset and we can get the result of Confusion Matrix and Statistics. The confusion Matrix for unbalanced train dataset on a new unseen test dataset looks like the following matrix

```
          (a)          (b)     <-classified as
          Reference
Prediction  0          1
         0  109        22      (a): class 0 child was died
         1  81         2979    (b): class 1 child was alive
```

As we can see in the confusion matrix, the model correctly predict 109 children as died and 2979 children as alive. It has wrongly predicted 81 children as died and 22 children as alive. When we train the model on balanced trainining dataset that has been balanced usning SMOTE sampling, we get the following result described with confusion matrix:

```
          a)           (b)    <-classified as
          Reference
Prediction  0        1
         0  124      64      (a): class 0 child was died
         1  66       2937    (b): class 1 child was alive
```

Table 4.12. Performance of Random Forest classifier

Metrix	Random Forest prediction model	
	Unbalanced Data	Balanced Data
Accuracy	96.77%	95.93%
95% CI	(0.961, 0.9736)	(0.9518, 0.9659)
Sensitivity	57.36%	65.26%
Specificity	99.26%	97.56%
Balanced Accuracy	78.31%	81.56%
Positive Predicted Value	83.21%	65.957%
Negative Predicted Value	97.35%	97.80%

The performance of the prediction model indicated in Table 4.12 compares with balanced and unbalanced train data. From the Table 4.12, we understand that the accuracy of the model with balanced train dataset was less than the accuracy of imbalanced train dataset, but the average accuracy of balanced train dataset was greater than from imbalanced train dataset because of the sensitivity was increased during balance train dataset. The average accuracy was increased from 78.31% to 81.56% because of balanced train dataset.

4.4 Evaluation

In this section, the performance analysis of classification algorithms of Random forest, C5.0 decision tree, Support Vector Machine, Naïve Bayes algorithms were carried out with a common dataset using R tool. In order to evaluate the performance of trained models, we use 10-fold cross-validation to estimate accuracy using caret packages. The models were estimated their accuracy of the trained model with 10-fold cross-validation by metric Accuracy and Kappa. Finally, the best results were selected depending on the result of the metric in Table 4.13 and Table 4.14 for unbalanced and balanced train dataset.

Table 4.13. 10-fold cross-validation on imbalance target data

Accuracy						
	Min	1st Qu.	Median	Mean	3rd Qu.	Max.
RF	0.9516129	0.9583473	0.9603765	0.9605370	0.9624161	0.9717742
C5.0	0.9543624	0.9610869	0.9644304	0.9639814	0.9664430	0.9705094
NB	0.9395973	0.9396176	0.9408602	0.9404030	0.9409396	0.9422819
SVM	0.9449664	0.9516453	0.9536606	0.9540489	0.9557492	0.9624161
Kappa						
	Min	1st Qu.	Median	Mean	3rd Qu.	Max.
RF	0.4758923	0.5759818	0.5955283	0.59996876	0.6217060	0.7095620
C5.0	0.5361197	0.5926548	0.6312709	0.62242921	0.6484728	0.7027781
NB	0.0000000	0.0000000	0.0000000	0.00416074	0.0000000	0.0419296
SVM	0.2607875	0.3833237	0.4363440	0.42853281	0.4770035	0.5329154

Table 4.14. 10-fold cross-validation on balance target data

Accuracy						
	Min	1st Qu.	Median	Mean	3rd Qu.	Max.
RF	0.9635974	0.9735011	0.9754011	0.9763511	0.9804813	0.9850107
C5.0	0.9657388	0.9724599	0.9775161	0.9764576	0.9804658	0.9850107
NB	0.8823529	0.8983957	0.9063169	0.9054034	0.9122056	0.9197861
SVM	0.9218415	0.9331376	0.9373999	0.9369720	0.9411607	0.9496788
Kappa						
	Min	1st Qu.	Median	Mean	3rd Qu.	Max.
RF	0.9092110	0.9345761	0.9393038	0.9416507	0.9516911	0.9632880
C5.0	0.9149406	0.9321970	0.9444443	0.9420862	0.9518417	0.9631224
NB	0.6900708	0.7362753	0.7584696	0.7534749	0.7726160	0.7953612
SVM	0.8052933	0.8342286	0.8432437	0.8433385	0.8540138	0.8746409

4.5 Summary of implementation

The dataset used in this study has many missing values and the variables in the dataset were very large which 1209 attributes were counted. From 1209 attributes selecting the important features were very difficult that feed to the prediction model. The dataset was pre-processed by using data pre-processing techniques such as missing values were imputed, balancing of the value of target variable using different sampling algorithms and SMOTE algorithm was selected at the end, the data transformed, normalization and scaled. For variable importance and ranking, Random Forest and Boruta algorithms were used. SPSS 20 software used to convert the machine understandable format and different statistical description was performed on it. By caret package the six models were built and trained. For support vector machine SVM algorithm we build three models depend on its kernel. Those were sigmoid kernel, SVM radial basis kernel and SVM polynomial kernel on training-test data set. Generally, six models were built. For models evaluation, 10-fold cross-validation was used with metrics of accuracy, Kappa and ROC. Finally, in order to evaluate the performance of all the models, specificity, sensitivity and average classification accuracy of each model was compared.

CHAPTER 5: DISCUSSION

5.1 Discussion

The goal of the research has been to identify the factors for childhood mortality and perform an experiment to evaluate the performance of four supervised machine learning algorithms. Initially, the classifiers have been trained on a set of training sample with the imbalanced dataset. The result showed the prediction accuracy is slightly low. Therefore, we understand that data balancing has a major impact on the prediction accuracy of the classifier and perform different resampling methods for data balancing.

In case of filling or handling missing values, most researchers used mean values of the feature itself. That means for the missed value in the features the same value was filled and this may be not all records in the features not the same. Here in this study, we used mice package in RStudio for multiple imputations that was given functions that can impute continuous, binary, and ordered/unordered categorical data, and imputing each incomplete variable with a separate model.

The dataset used in this study has many missing values and the features in the dataset were very large which 1209 features were counted. From 1209 features selecting the importance features were very difficult that feed to prediction models. Therefore the dataset was pre-processed by using data pre-processing techniques and depending on literature review, expert domain and by using feature selection technique 34 variables were selected. Some of the features were decoded to avoid overfitting and outliers. After 34 variables were selected using Boruta and random forest algorithms twenty attributes were feed to supervised machine learning algorithms. The 34 features were listed with its description in **Error! Reference source not found.**Table 3.1 and the top 20 features selected from it by using random forest algorithm. From the features listed in the Table 3.1the determinants that cause infant and child mortality was clearly listed in Table 4.1 with the percentage contribute to child mortality. And also the rule based model was build using C5.0 decision tree with 8 rules discussed in section **Error! Reference source not found.**. As we understand from Table 4.1 region to region there is a mortality difference that we understand from table Somali and Afar region have the highest mortality rate compared to the rest of the region. Also, the type of place residence has a great impact on child mortality which was the

childbirth in a rural area have a high probability for death rather than child birth and live in urban. In this study, the determinants like the presence of diarrhea founded by [18] and [17] were not now the determinants or factors of child mortality in this study.

The researcher found in this study infant and child mortality significantly influenced by Number of under-five child in house hold (AOR = 16.08, (95% CI [13.3,19.44]), type of place of residence (AOR = 0.73, 95% CI [0.5,1.08]), source of drinking water (AOR = 0.97, [0.94,0.99]), wealth index difference (AOR = 0.97, [0.89,1.06]), family plan use (AOR = 1.03, 95% CI [0.99,1.07]), breast feeding (AOR = 2.58, 95% CI [2.12,3.14]), place of delivery (AOR = 1.0062, 95% CI [0.9939,1.0187]), birth order number (AOR = 1.73, 95% CI [1.35,2.22]), and, preceding birth interval (AOR = 1.0075, [1.0017,1.0133]) were found to be determinants for child mortality using multivariate logistic analysis with adjacent odd ratio that indicated in

Table **4.2**.

For handling imbalance data we used different data imbalance handling methods as illustrated in chapter 4 in section 4.2.5. From that, the best sampling methods were selected by applying the sampling method on the trained model and tested on test data based on the Area under the ROC curve. The sampling was evaluated by the accuracy of respective predictions using inbuilt function roc.curve allows us to capture roc metric. From the roc.curve result SMOTE 93.74% highest mean accuracy and we used the balanced data sampled by SMOTE. The reason we use SMOTE not only its highest accuracy as we discuss in chapter 3 in section imbalance data handling method SMOTE generate synthetic data rather than replicating and adding the observations from the minority class under-sampling the majority class like over and under sampling techniques. The overall sampling techniques performance using Accuracy indicated in Table 4.5.

The estimation of accuracy and Kappa using 10-fold cross-validation illustrated in section 4.4 in Table 4.13 and Table 4.14 which trained models to have the highest accuracy regarding imbalanced and balanced data. From Table 4.13 on the imbalance data, C5.0 algorithm produced a more accurate model with a mean accuracy of 96.39% and Kappa 62.24%. Table 4.14 indicates also the C5.0 algorithm produced a more accurate model with a mean accuracy of 97.64% and Kappa 94.24%. The accuracy is an estimate that not yet applied on a new

dataset that used to prediction. Depending on unseen data the accuracy of the model's result were listed in Table 5.1.

Table 5.1. Performance the models in prediction on unseen test data.

Prediction unseen dataset on imbalance train data				
	RF	C5.0	SVM	NB
Accuracy	0.9677	0.9674	0.963	0.7728
Kappa	0.6627	0.6585	0.5646	0.0099
95% CI	(0.961, 0.9736)	(0.9606, 0.9733)	(0.9559, 0.9693)	(0.7579, 0.7872)
Sensitivity	0.57368	0.56842	0.43158	0.0052632
Specificity	0.99267	0.99267	0.99667	1.0000000
Balanced Accuracy	0.78318	0.78055	0.71412	0.5026316
Prediction unseen dataset on balance train data				
Accuracy	0.9593	0.9615	0.9533	0.9354
Kappa	0.6344	0.6498	0.6247	0.3499
95% CI	(0.9518, 0.9659)	(0.9542, 0.9679)	(0.9454, 0.9604)	(0.9264, 0.9437)
Sensitivity	0.65263	0.65789	0.72632	0.33684
Specificity	0.97867	0.98067	0.96768	0.97334
Balanced Accuracy	0.81565	0.81928	0.84700	0.65509

From Table 5.1 we understand that data imbalance has an impact on the performance of prediction on the unseen new dataset. When we look random forest model the accuracy of prediction increased from 96.74% to 98.43% when the data was balanced train data. In the same sensitivity and specificity increased when the data was balanced. The most noticeable change here is the significant increase in the average or balanced accuracy of the classifiers which is the average of sensitivity and specificity. From the Table 5.1, we understand the C5.0 model have higher prediction accuracy as compared to the other models in data imbalance. But when we look at the average accuracy random forest model is greater than the other and it is a good classifier. In the same hand, C5.0 model has higher prediction accuracy as compared to the other models in data imbalance. But also now the average accuracy of the random forest model is greater than the other models. Therefore, the random forest is a good classifier with 96.74% accuracy, 79.53% average accuracy in imbalanced train data and with 98.43% accuracy, 95.22% average accuracy in balanced train data.

CHAPTER 6: CONCLUSIONS AND RECOMENDATIONS

6.1 Conclusion

Identify the factors for infant and child mortality by apply supervised machine algorithm for building predictive models and make a prediction on the unseen dataset was the main target of this study. The models we used in the study have a statistical difference in its performance.

For experimentation, four supervised machine learning algorithms were used and six models were built on the features selected dataset. From the six built models, three of the models was trained using Support Vector Machine (SVM) algorithm on its SVM kernel. These SVM models were Sigmoid, polynomial and radial SVM, and from three of them, radial SVM kernel was selected as the best performance model with 96.30% performance accuracy and 71.14% balanced or average accuracy. Another three models were built using Random Forest (RF), Naïve Bayes (NB), and C5.0 algorithms. From the whole models, the random forest and C5.0 decision tree classifier performed lower error measures than the rest of the models.

The dataset missing values were handled by multiple imputations by MICE (Multivariate Imputation via Chained Equations) R packages and feature selection method was applied by Boruta algorithm. The importance of the top tweeny features were ranked with Random Forest and Boruta algorithms. For the problem of imbalanced data different sampling algorithms were applied and at the final SMOTE algorithm was selected based on its performance on trained models. The cleaned and selected dataset was split into two part which is train and test dataset by caret packages in R. Initially, the models were built over imbalanced values of mortality factors attributes and later same models are trained over balanced dataset to study the impact and importance of the balanced dataset in modeling. Finally, we get that Random Forest and C5.0 models were a good classifier on the unseen dataset.

A limitation of this study is the data is large have much missing values and the missing values were important but not filled by the respondent. Respondents were likely to forget events

that occurred in the past. This introduces recall bias in the reporting of events over the preceding 10 year period. In addition, the place of residence of some respondents at the time of the survey may have changed over the period from the date of birth of the child.

6.2 Future Work and Recommendations

This research only focused on four algorithms: Random Forest, C5.0, Support Vector Machine and Naïve Bayes but there were other supervised machine learning algorithms may have a good performance on a result of experiments. Also on model evaluation method, we used k-fold cross-validation resampling method but there are other resampling methods like leave one out cross validation. The metric we used in this study were Accuracy, Kappa and for future work also there are other metrics like Rsquare and ROC. In this study, the dataset which we used was 2016 EDHS dataset and we analyzed the determinants in the dataset. In the future by comparing the dataset starting from 2000 EDHS to there and by applying time series machine learning algorithms to compare the determinants that change through the time. In this, for handling missing values we have used MICE (Multivariate Imputation via Chained Equations) by using MICE package in RStudio but there were other techniques and packages in RStudio there for handling missing value like Amelia, VIM, MissForest, FactorMineR.

References

[1] C. S. Agency and D. Program, "Ethiopia Demographic and Health Survey 2016," Central Statistical Agency (CSA); DHS Program, Addis Ababa, Ethiopia, 2017.

[2] C. Nkosazana, L. Carlos, A. Akinwumi and C. Helen, "MDG Report: “Assessing Progress in Africa toward the Millennium Development Goals”," United Nations Economic Commission for Africa, Addis Ababa, Ethiopia, 2015.

[3] H. Lucia, S. David, Y. S., A. M. and Y. &Danzhen, "‘Levels & Trends in Child Mortality: , Estimates Developed by the UN Inter-Agency Group for Child Mortality Estimation’, United Nations Children’s Fund," United Nations Inter-Agency Group for Child Mortality Estimation (UN IGME), New York, 2017.

[4] A. UNION, "2017 Status report on Maternal Newborn Child and Adolescent Health: Focusing on Unfinished Business in Africa," Addis Ababa, Ethiopia, 2017.

[5] UNICEF and WHO, "Reaching the Every Newborn National 2020 Milestones Country progress, plans andmoving forward Geneva:," World Health Organization, Geneva, 2017.

[6] A. Tsai, "Methods for Multiple Imputing Analysis with R," 1 12 2019. [Online]. Available: https://www.analyticsvidhya.com/blog/2016/03/tutorial-powerful-packages-imputing-missing-values/.

[7] B. Miron and Kursa, Boruta for those in a hurry, July 17, 2018.

[8] P. Alex, C. Ben, P. Steven, d. P. Valerie and L. Zhixin, "Practical Application of Machine Learning Within Actuarial Work by Modelling, Analytics and Insights in Data working party," 2018.

[9] P. C. National and N. United, "Assessment of Ethiopia's Progress towards the Millennium Development Goals(Millennium Development Goals Report 2014)," Addis Ababa, Ethiopia, 2015.

[10] T. D., M. F. and R. (. Katherine, "Every child alive the urgent need to end newborn deaths," United Nations Children's Fund (UNICEF), Swezerland, 2018.

[11] S. Anagaw, "Application of Data Mining Technology to Predit Child mortality Patterns: The Case of Butarija Rural Health Project," *Journal of Healthcare Information Management,* (June 2002).

[12] M. Desta, "" Infant and Child Mortality in Ethiopia: The role of Socioeconomic, Demographic and Biological factors In the previous five years period of 2000 and 2005"," 2011.

[13] T. Be'emnetu, ""predicting the pattern of under-five mortality in Ethiopia using data mining technology: the case of Butajira rural health program," Addis Ababa, Ethiopia, 2012.

[14] H. Tesfahun, "application of data mining for predicting adult mortality," Addis Ababa, Ethiopia, 2012.

[15] N. Assefa, A. Gebeyehu, B. Terefe, G. Tesfayi and P. Roger, ""An Analysis of the Trends, Differentials and Key Proximate Determinants of Infant and Under-Five Mortality in Ethiopia": Further Analysis of the 2000, 2005, and 2011 Dememgraphic and health survey," *MoFED and UNICEF,* pp. 1 - 36, 2013.

[16] O. Jiregna, ""determinants of infant mortality in Ethiopia: a multilevel count regression models"," Addis Ababa, Ethiopia, 2016.

[17] D. Bedada, ""Determinant of Under-Five Child Mortality in Ethiopia.," *American Journal of Statistics and Probability,* vol. 6, no. 4, p. 198, 2017.

[18] T. Brook, A. Suleman, E. Noah, D. Legesse, S. Syed-Abdul and K. Mihiretu, ""“Determinants and development of a web-based child mortality prediction model in resource-limited settings: A data mining approach”: Ethiopia.," *Computer Methods and Programs in Biomedicine,* vol. 140, p. 45–51, 2017.

[19] Caluza and L. J. B., ""“Machine Learning Algorithm Application in Predicting Children Mortality”: A Model Development," *International Journal of Information Sciences and Application,* 2018.

[20] H. Kenneth, L. Stefan, S. Judith and S. Johannes, "Modeling Under-5 Mortality through Multilevel Structured Additive Regression with Varying Coefficients for Asia and Sub-Saharan Africa," 2017.

[21] P. Sonkar, "Application of supervised machine learning to predict the mortality risk in elderly using biomarkers. Masters dissertation," 2017..

[22] L. J. B. Caluza., "Machine Learning Algorithm Application in Predicting Children Mortality: A Model Development,," *International Journal of Information Sciences and Application,* pp. 1-6, 2018.

[23] K. S. Aayush, P. Devaraj, B. Sneha and S. Ramraj, "Application of Machine Learning in Analysis of Infant Mortality and its Factors," *Researchgate,* 2016.

[24] F. Munyamahoro, "An Empirical Analysis of Death of Children Under Five Years in Rwanda," *iMedPub Journals,* 2017.

[25] O. Gordon, "A comparative study of decision tree and naïve bayesian classifiers on verbal autopsy datasets," October 2015.

[26] [Online]. Available: http://www.measuredhs.com/.

[27] L. Brett, Machine Learning with R: Learn how to use R to apply powerful machine learning, Birmingham: Packt Publishing Ltd, 2013.

[28] ACAPS., ACAPS: Data cleaning- dealing with messy data., 2016.

[29] B. Lantz, Machine Learning with R Second Edition; Discover how to build machine learning algorithms, prepare data, and dig deep into data prediction techniques with R;, BIRMINGHAM - MUMBAI: Packet Publishing, 2015.

[30] S. Manohar, Mastering Machine Learning with Python in Six Steps: A Practical Implementation Guide to Predictive Data Analytics Using Python, Bangalore, Karnataka, India: Apress, 2017.

[31] S. Buuren and K. Groothuis-Oudshoorn, " mice: Multivariate imputation by chained equations in R.," *Journal of statistical software,* 2011.

[32] R. Revathy and R. Lawrance., "“Comparative Analysis of C4.5 and C5.0 Algorithms on Crop Pest Data”," *International Journal of Innovative Research in Computer and Communication Engineering,* 1, March 2017.

[33] Y. Pinar, "Filter Based Feature Selection Methods for Prediction of Risks in Hepatitis Disease," *International Journal of Machine Learning and Computing* , Vols. Vol. 5, , pp. No. 4, , August 2015.

[34] P. Wiesław, W. Mariusz, N. Rafał and W. R. R., "Application of all-relevant feature selection for the failure analysis of parameter-induced simulation crashes in climate models," *GeoScientific,* 2016.

[35] R. Witold, W. Mariusz and P. Wiesław, "Feature Selection for Data and Pattern Recognition," *Springer,* 2015.

[36] L. Gilles, "Understanding Random Forests from theory to practice: from theory to practice," 2015.

[37] B. Jason, Machine Learning Mastery With R, Melbourne, Australia, 2017.

[38] P. Max Kuhn, Predictive Modeling with R and the caret Package, 2013.

[39] E. Duchesnay and L. Tommy, Statistics and Machine Learning in Python, Release 0.1, 2018.

[40] Z. Zhuoyuan, C. Yunpeng and L. Ye, "Oversampling method for imbalanced Classification," *Computing and Informatics*, pp. 1017- 1037, 2015.

[41] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research,* p. 321–357, 2002.

[42] L. Nicola, M. Giovanna and T. Nicola, "ROSE: A Package for Binary Imbalanced Learning," *The R Journal,* June 2014.

[43] P. Milind, Predictive Modeling for Algorithmic Trading, 2016.

[44] "Machine learning with python tutorial;," 2015. [Online].

[45] S. MIDAS and I. Dinov, Data Science and Predictive Analytics; Decision Tree Divide and Conquer Classification, (October 2018).

[46] H. J. and K. M., " Data Mining Concepts and Techniques.," New York. USA, (2001).

[47] Ian H.Witten, Eibe(2nd ed.), Data mining: practical machine learning tools and techniques, (2005).

[48] J. Grus, Data Science from Scratch, United States of America: (O'Reilly), (2015).

[49] P. Rutvija and P. Jayati, "C5.0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning," *International Journal of Computer Applications,* 2015.

[50] K. Max, Classification Using C5.0 UseR!, Pfizer Global R&D, 2013.

[51] B. Lantz, Machine Learning with R, Discover how to build machine learning algorithms,First published, BIRMINGHAM - MUMBAI: Packet publishing, 2013.

[52] G. Anshul and M. Rajni, "Performance Comparison of Naïve Bayes and J48 Classification Algorithms," *International Journal of Applied Engineering Research,* 2012.

[53] Sanjay and S. Kumar, "Predicting and Diagnosing of Heart Disease Using Machine Learning Algorithms," *International Journal Of Engineering And Computer Science,* pp. 21623-21631, 2017.

[54] S. Sherif, E. Radwa, A. M. Ahmed and T. Q. Waqas, "Comparison of machine learning techniques to predict all-cause mortality using fitness data: the Henry ford exercIse testing (FIT) project," *BMC Medical Informatics,* 2017.

[55] A. Arivarasan and D. M. karthikeyan, "Classification based Performance analysis using Naïve-Bayes J48 and Random forest algorithms," *International Journal of Applied Research,* pp. 174-178, 2017.

[56] M. F. K. R. Tara Dooley, "United Nations Children's Fund, Every child alive: The urgent need to end newborn deaths," UNICEF, Switzerland, (2018).

[57] N. P. Commission, "Ethiopia 2017 Voluntary National Review on SDGs Government Commitments, National Ownership and Performance Trends," Addis Abeba, June 2017.

[58] R. R. Witold, W. ´. Mariusz and P. Wiesław, " Feature Selection for Data and Pattern Recognition, Studies in Computational Intelligence," *Springer-Verlag Berlin Heidelberg,* 2015.

[59] P. Rutvija and J. P., "C5.0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning," *International Journal of Computer Applications,* 16, May 2015.

[60] C. S. A. (. [Ethiopia] and ICF, "Ethiopia Demographic and Health Survey 2016," (Central Statistical Agency (CSA), Addis Ababa, Ethiopia, and Rockville, Maryland, USA: CSA and ICF, 2017.

[61] S. Raschka, "Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning," November 2018.

[62] J. Gareth, W. Daniela, H. Trevor and T. Robert, An Introduction to Statistical Learning with Applications in R, New York: Springer, 2017.

Appendix A

The source code used in the study using RStudio

```
infant <- read.csv("infant.csv", header =TRUE) # Load the data set
table(infant$B5) # Display the class distribution of alive and died
#confert B5 to factor variable
infant$B5 <- as.factor(infant$B5)
# Handling missing Values
#Produce NAs in 1% of the data
infant.mis <- prodNA(infant, noNA = 0.001)
head(infant.mis)
##Visualizing “Missing Data”
md.pattern(infant.mis)
#########################################################
#Multivariate Imputation by Chained Equations (MICE) for Handling Missing Values
imputed_infant <- mice(infant.mis, m=5, maxit = 10, method = 'pmm', seed = 500)
#m: the number of imputations made per missing observation (5 is normal–generates 5 data sets with imputed/original values)
#maxit: the number of iterations
#method: We use ’probable means. Seed: Values to randomly generate from
summary(imputed_infant)   # display the summary of imputed value.
```

```
########################Split the dataset into Train and Test dataset################
```

```
############################### load the libraries ###################################
library(caret)
library(klaR)
# define an 70%/30% train/test split of the dataset
set.seed(333)
split=0.70
indinfant <- createDataPartition(infant$B5, p=split, list=FALSE)
train_infant <- dataset[indinfant, ]
test_infant <- dataset[-indinfant, ]
```

```
##############################################################################
########## Feature or variable selection using Boruta Algorithm ################
##############################################################################
#Load the packages
library(Boruta)
library(mlbench)
library(randomForest)
library(caret)
library(car)
library(rFerns) # for shorter time importance selection
set.seed(123)
Boruta_infant<- Boruta(B5 ~ ., data = infant, doTrace = 2)
print(Boruta_infant)
plot(Boruta_infant)
plot(Boruta_infant, las = 2, cex.axis = 0.7) # plot the feature with rejected, tentative and confirmed features
plotImpHistory(Boruta_infant)
getNonRejectedFormula(Boruta_infant)  # Display tentative and confirmed features
getConfirmedFormula(Boruta_infant) # Display confirmed features only
imporinfant= imps[imps$decision != 'Rejected', c('meanImp', 'decision')]
imporinfant [order(-imporinfant $meanImp), ]  # descending sort
# Plot variable importance
plot(Boruta_infant, cex.axis=.7, las=2, xlab="", main="Variable Importance")
##############################################################################
# Rank features  importance using random forest.
#Load the packages
library(caret)
library(randomForest)
# Set the random number seed generator
set.seed(333)
# Fit a random forest model for B5 against everything in the model (~.)
Rfmod_infant <- train(B5~., data=infant, method="rf")
# Take a look at the output
```

```
Rfmod_infant
#The varImp function knows what kind of model was fitted and knows how to estimate variable importance.
varImp(Rfmod_infant, scale=TRUE)
#Plot Var importance
varImp(Rfmod_infant, scale=TRUE) %>% plot()
```

#After appling the feature selection methods top twenty features were selected and the file name was “selected” and it loaded ##

```
selected <- read.csv("selected.csv", header =TRUE)
###Then split the selected dataset into Train and Test dataset
#########################################################################
# load the libraries
library(caret)
library(klaR)
selected$B5 <- as.factor(selected$B5) # Convert the outcome to factor
# define an 70%/30% train/test split of the dataset split=0.70
trainIndexselected <- createDataPartition(selected$B5, p=split, list=FALSE)
trainselected <- selected[ trainIndexselected,]
testselected  <- selected[-trainIndexselected,]
#########################################################################
####################Handling Class Imbalance Problem ###################
#########################################################################
########################################OVERSAMPLING###################
#over sampling
data_balanced_over <- ovun.sample(B5 ~ ., data = trainselected, method = "over")$data
table(data_balanced_over$B5)
CrossTable(data_balanced_over$B5, prop.chisq = FALSE)
##Undersampling
data_balanced_under <- ovun.sample(B5 ~ ., data = trainselected, method = "under", seed = 1)$data
table(data_balanced_under$B5)
```

```
#########################SMOTE
selected_balanced_smote <- SMOTE(B5~., trainselected)

#Both over and under
data_balanced_both <- ovun.sample(B5 ~ ., data = trainselected, method = "both", p=0.5, N=7450, seed = 1)$data
table(data_balanced_both$B5)
#########################ROSE
data_balanced_rose <- ROSE(B5 ~ ., data = trainselected, seed = 1)$data
table(data_balanced_rose$B5)
#Evaluete the sampling techniques with the model
# load libraries
library(mlbench)
library(caret)
# 10-fold cross-validation with 3 repeats
trcnt <- trainControl(method="repeatedcv", number=10, repeats=3)
metric <- "Accuracy"
##Now Let’s compute the model using each data and evaluate its accuracy.build model
#build decision tree models
data_balanced_both$B5 <- as.factor(data_balanced_both$B5)
tree.rose <- train(B5 ~ ., data = data_balanced_rose, method="rf", metric=metric, preProc=c("BoxCox"), trControl=trcnt, na.action=na.omit)
tree.over <- train(B5 ~ ., data = data_balanced_over, method=" rf ", metric=metric, preProc=c("BoxCox"), trControl=trcnt, na.action=na.omit)
tree.under <- train(B5 ~ ., data = data_balanced_under, method=" rf ", metric=metric, preProc=c("BoxCox"), trControl=trcnt, na.action=na.omit)
tree.both <- train(B5 ~ ., data = data_balanced_both, method=" rf ", metric=metric, preProc=c("BoxCox"), trControl=trcnt, na.action=na.omit)
tree.smote <- train(B5 ~ ., data = selected_balanced_smote, method=" rf ", metric=metric, preProc=c("BoxCox"), trControl=trcnt, na.action=na.omit)
```

```
#make predictions on unseen dataset
pred.tree.rose <- predict(tree.rose, newdata = testselected)
pred.tree.over <- predict(tree.over, newdata = testselected)
pred.tree.under <- predict(tree.under, newdata = testselected)
pred.tree.both <- predict(tree.both, newdata = testselected)
pred.tree.smote <- predict(tree.smote, newdata = testselected)

#It’s time to evaluate the accuracy of respective predictions. Using inbuilt function roc.curve allows us to capture roc metric.
roc.curve(testselected$B5, pred.tree.rose, main ='ROSE')
roc.curve(testselected$B5, pred.tree.over, main ='OVER')
roc.curve(testselected$B5, pred.tree.under, main ='UNDER')
roc.curve(testselected$B5, pred.tree.both, main ='BOTH UNDER and OVER')
roc.curve(testselected$B5, pred.tree.smote, main ='SMOTE')
################################
results <- resamples(list(ROSE=tree.rose, OVER=tree.over, UNDER=tree.under, BOTH=tree.both, SMOTE=tree.smote))
parallelplot(results, metric = "Accuracy")
splom(results, metric = "Accuracy")
summary(results)
dotplot(results)
```

Model Train and Model Evaluation

```
################################Unbalance data ##############################
# load libraries
library(mlbench)
library(caret)
library(gmodels)
# 10-fold cross-validation with 3 repeats
trcnt <- trainControl(method="repeatedcv", number=10, repeats=3) metric <- "Accuracy"
set.seed(333)     # Set the seed for train the model and evaluate
#Random Forest model for imbalance dataset
```

```
sel.rfun <-train(B5~.,data=trainselected,method="rf",metric=metric, preProc=c("BoxCox"),
        trControl=trcnt, na.action=na.omit)
print(sel.rfun)  # print the result of the model
confusionMatrix(predict(sel.rfun, testselected), testselected$B5) # display result in confition
matrix with different model evalution techniques
# C5.0 Decision Tree for imbalance Dataset
set.seed(333) # Set the seed for train the model and evaluate
sel.c5.0un<-train(B5~.,data=trainselected,method="C5.0",metric=metric,
preProc=c("BoxCox"),  trControl=trcnt, na.action=na.omit)
summary(sel.c5.0un) # Display the summary of the model
confusionMatrix(predict(sel.c5.0un, testselected), testselected$B5) # ) # display result in
confition matrix with different model evalution techniques
# Naïve Bayes model for imbalance dataset
set.seed(333) # Set the seed
fit.nb <- train(B5~., data=trainhiwi, method="nb", metric=metric, trControl=trcnt)
confusionMatrix(predict(fit.nb, testhiwi), testhiwi$B5)
sel.nbun<-train(B5~.,data=trainselected,method="nb",metric=metric,
preProc=c("BoxCox"), trControl=trcnt, na.action=na.omit)
confusionMatrix(predict(sel.nbun, testselected), testselected$B5) ) # display result in
confition matrix with different model evalution techniques
#Support Vector Machine for imbalanced dataset
set.seed(333) # Set seed
sel.svmun <- train(B5~., data=trainselected, method="svmRadial", metric=metric,
preProc=c("BoxCox"),  trControl=trcnt, na.action=na.omit)
confusionMatrix(predict(sel.svmun, testselected), testselected$B5) ) # display result in
confition matrix with different model evalution techniques
# Compare results the trained models with the accuracy metric and display the result
ensembleResults <-resamples(list(RF=sel.rfun, C5.0=sel.c5.0un, NB=sel.nbun,
SVM=sel.svmun))
summary(ensembleResults) # Display the  compared result of the models
dotplot(ensembleResults)  # Display in figures
```

##Balanced Data ##########################

```
# load libraries
library(mlbench)
library(caret)
# 10-fold cross-validation with 3 repeats
trcnt <- trainControl(method="repeatedcv", number=10, repeats=3)
metric <- "Accuracy"
#Random Forest for Balanced train dataset by SMOTE sampling
set.seed(333)
sel.rf <- train(B5~., data=selected_balanced_smote, method="rf", metric=metric, preProc=c("BoxCox"), trControl=trcnt, na.action=na.omit) # Train the model
confusionMatrix(predict(sel.rf, testselected), testselected$B5) ) # display result in confition matrix with different model evalution techniques

#C5.0 Model for balanced dataset by SMOTE sampling
set.seed(333)
sel.c5.0 <- train(B5~., data=selected_balanced_smote, method="C5.0", metric=metric, preProc=c("BoxCox"), trControl=trcnt, na.action=na.omit)
confusionMatrix(predict(sel.c5.0, testselected), testselected$B5) ) # display result in confition matrix with different model evalution techniques
# Naïve Bayes for balanced dataset by SMOTE sampling
set.seed(333)
sel.nb <- train(B5~., data=selected_balanced_smote, method="nb", metric=metric, preProc=c("BoxCox"), trControl=trcnt, na.action=na.omit)
confusionMatrix(predict(sel.nb, testselected), testselected$B5) ) # display result in confition matrix with different model evalution techniques
#Support Vector Machine for balanced dataset by SMOTE sampling
set.seed(333)
sel.svm<- train(B5~., data=selected_balanced_smote, method="svmRadial", metric=metric, preProc=c("BoxCox"), trControl=trcnt, na.action=na.omit) # train by Radial kernel
confusionMatrix(predict(sel.svm, testselected), testselected$B5) ) # display result in confition matrix with different model evalution techniques
# Support Vector Machines with Polynomial Kernel
```

```
set.seed(333)
sel.svmPoly <- train(B5~., data=trainselected, method="svmPoly", metric=metric, preProc=c("BoxCox"), trControl=trcnt, na.action=na.omit)
confusionMatrix(predict(sel.svmPoly, testselected), testselected$B5) ) # display result in confition matrix with different model evalution techniques
# Support Vector Machines with Radial Basis Function Kernel
set.seed(333)
sel.svmRadialCost <- train(B5~., data=trainselected, method="svmRadialCost", metric=metric, preProc=c("BoxCox"), trControl=trcnt, na.action=na.omit)
confusionMatrix(predict(sel.svmRadialCost, testselected), testselected$B5)
# Support Vector Machines with Radial Basis Function Kernel
# Compare results
ensembleResults <- resamples(list(RF=sel.rf, C5.0=sel.c5.0, NB=sel.nb, SVM=sel.svm))
# Compare results
summary(ensembleResults)
dotplot(ensembleResults)
```